**IPL SCORE PREDICTOR**

**Submitted for**

**CSET211 - Statistical Machine Learning**

**Submitted by:**

**E23CSEU1243 Ritesh Pandey**

**E23CSEU1256 Mohd Kayes Ansari**

**E23CSEU1260 Pritam Kumar Mishra**

**Submitted to: Dr. Prashant Kapil**

**July-Dec 2024**

**SCHOOL OF COMPUTER SCIENCE AND ENGINEERING**

A close-up of a logo

Description automatically generated

**ABSTRACT**

The Indian Premier League (IPL) is renowned for its unpredictability, making accurate score prediction a challenging task. This project aims to develop a robust machine learning model to predict the scores of IPL matches.

By leveraging historical match data, player statistics, team performance metrics, and external factors like pitch conditions and weather, we aim to create a predictive model that can provide valuable insights to fans, analysts, and betting enthusiasts. The model will employ a combination of feature engineering, data preprocessing, and advanced machine learning techniques to capture the intricate patterns and dependencies within the data.

Through rigorous evaluation and validation, we strive to enhance the accuracy and reliability of our predictions, contributing to a deeper understanding of the factors influencing IPL match outcomes.

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No** | **Content** | **Page No** |
| **1.** | **Introduction** | 3 |
| **2.** | **Related Surveys** | 4 |
| **3.** | **Datasets** | 5 |
| **3.1** | **Data Preprocessing** | 5 |
| **4.** | **Methodology** | 6 |
| **4.1** | **Hardware and Software Requirements** | 6 |
| **4.2** | **Performance Metrics** | 6 |
| **5.** | **Result and Analysis** | 7 |
| **6.** | **Conclusions and Future Works** | 8 |

1. **Introduction**

**The Unpredictable Game: An AI-Driven Approach to IPL Score Prediction**

The Indian Premier League (IPL) has captivated cricket fans worldwide with its high-octane matches and unpredictable outcomes. Predicting the outcome of an IPL match, especially the first innings score, has been a perennial challenge for fans, analysts, and betting enthusiasts alike. This project aims to leverage the power of artificial intelligence to develop a robust model capable of accurately predicting the scores of IPL matches.

**Motivation**

The primary motivation behind this project is to delve into the intricate dynamics of cricket matches and harness the capabilities of AI to extract meaningful insights from historical data. By developing an accurate prediction model, we can:

1. **Enhance Fan Experience:** Provide fans with data-driven predictions and analysis, elevating their viewing experience.
2. **Assist Cricket Analysts:** Offer valuable insights to analysts and commentators, aiding in their pre-match and in-match analysis.
3. **Inform Betting Strategies:** Provide a statistical edge to betting enthusiasts, helping them make informed decisions.

**Objective**

The primary objective of this project is to build a machine learning model that can accurately predict the scores of an IPL match based on various factors such as:

* **Team Performance:** Historical performance of the batting and bowling teams.
* **Player Statistics:** Individual player performance metrics, including batting and bowling averages, strike rates, and economy rates.
* **Pitch Conditions:** Factors like pitch type, weather conditions, and venue.
* **Match Context:** Information about the match, such as the stage of the tournament and the importance of the match.

By leveraging advanced machine learning techniques and a comprehensive dataset, we aim to develop a model that can outperform traditional statistical methods and provide reliable predictions.

The most crucial **contribution** that the project aims to do is to **provide valuable insights.** The insights gained from the model can be used to inform decision-making by teams, coaches, and players, as well as providing valuable information to fans and analysts.

1. **Related Surveys**

**We studied a total of 9 research papers and** **some key findings of different research papers and surveys include:**

1. The research paper proposes a machine learning model, primarily using Random Forest, to predict IPL match outcomes. The model leverages historical data from 2008-2019, incorporating player and match-specific features, and achieves an accuracy of 88.10%.
2. The research analyzes IPL matches, rating players based on performance and considering factors like team form and strength. It introduces Batting and Bowling Indices for novel analysis. Decision Tree and Logistic Regression achieve accuracies over 87% and 95%, respectively.
3. The research aims to predict IPL match winners using machine learning algorithms like Decision Tree, Naive Bayes, KNN, and Random Forest. After data preprocessing and feature selection, Random Forest is identified as the most accurate model, outperforming others in terms of accuracy and error rate.
4. The research utilizes Linear and Ridge Regression models to predict first-inning scores in IPL matches. After data preprocessing and feature selection, Linear Regression emerges as the more accurate model. The system is implemented in Python with a web interface for user-friendly predictions.
5. The research predicts IPL match outcomes using a Multilayer Perceptron (MLP) model. The MLP, trained on 634 matches, achieves 71.66% accuracy on 2018 matches, outperforming other models and demonstrating its potential for predicting IPL matches.
6. The study develops an AI-powered system to detect and manage crop diseases. By training machine learning models on annotated crop images, the system accurately identifies diseases, leading to improved crop yield and management practices for farmers.
7. The research predicts IPL match winners using machine learning models, including Random Forest, SVM, Logistic Regression, and KNN. Random Forest achieved the highest accuracy (88.1%) in predicting match outcomes based on factors like strike rate. Future research could focus on predicting individual player performance and identifying the man of the match.
8. The research predicts IPL match scores using machine learning models, including Extra Tree Regressor. The model, trained on historical and real-time data, achieves 90% accuracy. Future research aims to improve accuracy by incorporating additional factors like team dynamics and social media sentiment analysis.
9. The research predicts IPL match scores using machine learning models, including Extra Tree Regressor. The model, trained on historical and real-time data, achieves 90% accuracy. Future research aims to improve accuracy by incorporating additional factors like team dynamics and social media sentiment analysis.

While significant progress has been made in IPL score prediction using machine learning, there is still room for improvement. Future research can focus on refining **existing techniques**, **exploring novel approaches**, and **incorporating real-time data** to enhance prediction accuracy. By leveraging the power of AI, we can continue to unlock the secrets of the unpredictable world of cricket

1. **Datasets**

The project consists of 3 datasets which are as follows:

1. **Deliveries**: The data set aims to provide ball-by-ball data which includesover, ball number, batsman, bowler, runs scored, wickets taken, extras, and other relevant details.
2. **Matches:** The data set consist of 757 IPL fixtures with team details, venue, toss winner, toss decision, match winner, win margins, timings, player of the match and umpire’s names.
   1. **Data preprocessing**

The project employs several data preprocessing techniques to prepare the IPL data for building a machine learning model. The techniques are:

1. **Handling Missing Values (Implicit):** The code doesn't explicitly address missing values. It's likely assuming there are no missing values or that they are negligible.
2. **Removing Irrelevant Columns:** The code identifies and removes columns considered irrelevant for predicting the scores. These include mid, date, venue, batsman, bowler, striker, and non-striker.
3. **Filtering Teams:** The code focuses on a consistent set of teams throughout the IPL seasons. It creates a list of these teams (const\_teams) and then filters the data to only include matches where both the batting team and bowling team are from this list. This ensures the model is trained on data with consistent team representations.
4. **Filtering Overs:** The code removes matches where the number of overs played in the first inning is less than 5. This is likely done to focus on complete innings or innings where a significant portion has been played, leading to a more reliable score prediction.

**Data Acquisition**

1. **Data Import:** The project starts by importing the IPL data from a CSV file using pandas.
2. **Exploratory Data Analysis (EDA):**
   * The code checks the data shape and uses describe () and info () functions to understand data types and identify potential issues.

**Data Cleaning:** Columns considered irrelevant for predicting the first inning score are identified and removed (e.g., mid, date, venue).

**Handling Categorical Features:** Label encoding is applied to convert categorical features like bat\_team and bowl\_team into numerical values using LabelEncoder from scikit-learn. This allows the model to work with these features.

1. **Methodology**

The code implements several regression algorithms for comparison:

1. Decision Tree Regressor
2. Linear Regression
3. Random Forest Regressor
4. Support Vector Machine (SVM) Regression
5. XGBoost Regressor
6. K-Nearest Neighbours Regressor (KNN)

**Model Evaluation:**

* Evaluates each model on the testing set using metrics like R-squared (score) and Mean Absolute Error (MAE) to understand how well the model predicts the actual scores.
* Prints the training and testing scores for each model.

**Model Selection:**

* Compares the performance of all models based on their R-squared scores.
* Chooses the model with the highest R-squared score on the testing set (presumably the Random Forest Regressor in this case based on the comments) as the best model for prediction.

**4.1. Hardware and Software Requirements**

The hardware requirements for running this project are fairly minimal. Some of the requirements are mentioned below:

1. **CPU:** Any modern processor (Intel i3 or AMD Ryzen 3 or better) should suffice.
2. **Storage:** Enough storage space to hold the dataset and the train the model. The size of the dataset is not massive so the model will be relatively small.

The Software Requirements include:

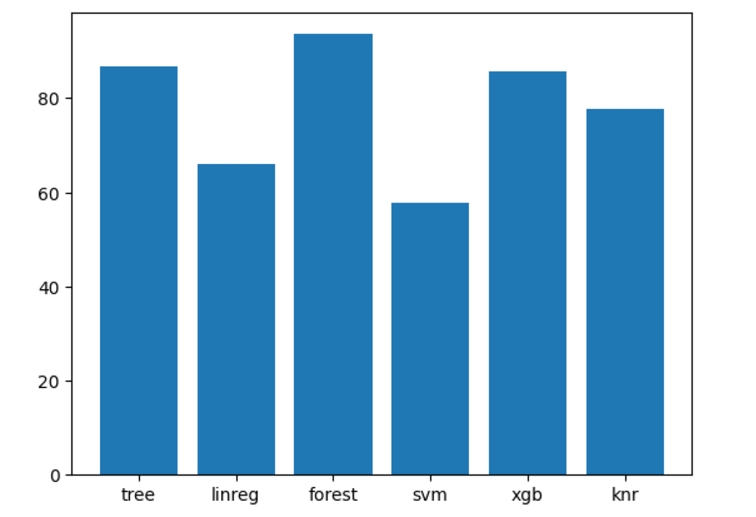
* **Python 3.x:** This is the programming language your code is written in.
* **Libraries:** Several Python libraries are used in your code. These can be installed using the pip command in your terminal. Here are the required ones:
  + 1. pandas
    2. numpy
    3. seaborn
    4. matplotlib
    5. scikit-learn (includes most of the machine learning models used)
    6. xgboost (for XGBoost model)
    7. pickle (for saving the model)

**4.2. Performance Metrics**

The project utilizes several performance metrics to evaluate the different machine learning models for predicting the first inning score in IPL matches. Here's a breakdown:

1. **R-squared Score (Train & Test):** This metric is reported for each model after training on the training data and evaluating on the testing data. It measures the proportion of variance in the target variable (first inning score) explained by the independent variables (features) in the model. Higher R-squared values indicate a better fit between the model and the data.
2. **Mean Absolute Error (MAE):** This metric is calculated for each model on the testing data. It measures the average absolute difference between the predicted scores and the actual scores. Lower MAE values indicate better prediction accuracy.
3. **Mean Squared Error (MSE):** This metric is calculated for each model on the testing data. It measures the average squared difference between the predicted scores and the actual scores. Lower MSE values indicate better prediction accuracy.
4. **Root Mean Squared Error (RMSE):** This metric is calculated for each model on the testing data. It's the square root of MSE and provides an error measure in the same units as the target variable (runs). Lower RMSE values indicate better prediction accuracy.

* The project displays a bar chart comparing the test R-squared scores of all models, visually indicating the best performing model in terms of R-squared.
* The code showcases example predictions for two test cases, highlighting the predicted score vs. the actual score. This provides a glimpse into the model's prediction capabilities.

**5. Results and Analysis**

**Model Comparison**

The code cleaned the data, removed irrelevant columns and encoded categorical features. We included Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for evaluation.

Out of the 6 models deployed, Random Forest gives the best performance with test case score of 93.62%. The code showcases evaluation metrics like MAE, MSE, and RMSE for each model on the test set. These metrics help analyze how well the model generalizes to unseen data.

Some of the limitations of the project includes:

* The code doesn't explicitly mention feature engineering, which could potentially improve model performance.
* The model's accuracy might be limited by the data quality and the chosen set of features.

**6. Conclusion**

The project successfully built a machine learning model to predict scores of an IPL match. It implemented a regression approach using six different models: **Decision Tree, Linear Regression, Random Forest, Support Vector Regression, XGBoost, and K-Nearest Neighbours**.

The data underwent cleaning, feature engineering (including one-hot encoding for categorical features), and splitting into training and testing sets. Evaluation metrics like R-squared and Mean Absolute Error helped identify the **Random Forest model as the best performer with a test score of 93.62%**

The project demonstrates the ability to predict scores based on factors like teams, runs, wickets, and overs bowled. While Test 1 and Test 2 showcase sample predictions, including more test cases would strengthen the analysis.

**Future works:**

The long-term aim is to:

* Explore feature engineering techniques and create new features.
* Adding additional factors like toss result and weather.
* Train multiple models and compare their performance using techniques like cross-validation.

By addressing the limitations and exploring advanced techniques, the model can be further optimized to provide more accurate predictions.

**References**

**GitHub Link:** https://github.com/pandit-001/IPL\_Score.git

**Research papers and Surveys: (Page 4)**

1. [**https://www.bing.com/ck/a?!&&p=f7da7909416ddc253f0fe84b9435ba8e1a67bc2a4a60bbb3e4ee797bc003a3a2JmltdHM9MTczMDkzNzYwMA&ptn=3&ver=2&hsh=4&fclid=22efaef5-d3da-64a6-1ec4-bdb4d22865bc&psq=ipl+score+predictor+research+paper+summaries&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zNTUwNjExMzlfUHJlZGljdGlvbl9vZl9JUExfTWF0Y2hfT3V0Y29tZV9Vc2luZ19NYWNoaW5lX0xlYXJuaW5nX1RlY2huaXF1ZXM&ntb=1**](https://www.bing.com/ck/a?!&&p=f7da7909416ddc253f0fe84b9435ba8e1a67bc2a4a60bbb3e4ee797bc003a3a2JmltdHM9MTczMDkzNzYwMA&ptn=3&ver=2&hsh=4&fclid=22efaef5-d3da-64a6-1ec4-bdb4d22865bc&psq=ipl+score+predictor+research+paper+summaries&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zNTUwNjExMzlfUHJlZGljdGlvbl9vZl9JUExfTWF0Y2hfT3V0Y29tZV9Vc2luZ19NYWNoaW5lX0xlYXJuaW5nX1RlY2huaXF1ZXM&ntb=1)
2. [**https://www.bing.com/ck/a?!&&p=23971813305a2b3ecc522bfa51e8c3cbc9ec7c93a3a80bd04207a9a565b1506eJmltdHM9MTczMDkzNzYwMA&ptn=3&ver=2&hsh=4&fclid=22efaef5-d3da-64a6-1ec4-bdb4d22865bc&psq=ipl+score+predictor+research+paper+summaries&u=a1aHR0cHM6Ly9pZWVleHBsb3JlLmllZWUub3JnL2RvY3VtZW50LzkxNTM5NzI&ntb=1**](https://www.bing.com/ck/a?!&&p=23971813305a2b3ecc522bfa51e8c3cbc9ec7c93a3a80bd04207a9a565b1506eJmltdHM9MTczMDkzNzYwMA&ptn=3&ver=2&hsh=4&fclid=22efaef5-d3da-64a6-1ec4-bdb4d22865bc&psq=ipl+score+predictor+research+paper+summaries&u=a1aHR0cHM6Ly9pZWVleHBsb3JlLmllZWUub3JnL2RvY3VtZW50LzkxNTM5NzI&ntb=1)
3. [**https://www.bing.com/ck/a?!&&p=e91872b7e0f9f55c74480f24a02df3253a0fd81626829f22a5920e197c0dbcd4JmltdHM9MTczMDkzNzYwMA&ptn=3&ver=2&hsh=4&fclid=22efaef5-d3da-64a6-1ec4-bdb4d22865bc&psq=ipl+score+predictor+research+paper+summaries&u=a1aHR0cHM6Ly9pY3RhY3Rqb3VybmFscy5pbi9wYXBlci9JSlNDX1ZvbF8xMV9Jc3NfMV9QYXBlcl8yXzIxOTlfMjIwNC5wZGY&ntb=1**](https://www.bing.com/ck/a?!&&p=e91872b7e0f9f55c74480f24a02df3253a0fd81626829f22a5920e197c0dbcd4JmltdHM9MTczMDkzNzYwMA&ptn=3&ver=2&hsh=4&fclid=22efaef5-d3da-64a6-1ec4-bdb4d22865bc&psq=ipl+score+predictor+research+paper+summaries&u=a1aHR0cHM6Ly9pY3RhY3Rqb3VybmFscy5pbi9wYXBlci9JSlNDX1ZvbF8xMV9Jc3NfMV9QYXBlcl8yXzIxOTlfMjIwNC5wZGY&ntb=1)
4. [**https://www.bing.com/ck/a?!&&p=460dd274e4c50ed771c84752951617b4ad159eba0e9aa55c32a2a62c3c0eb10cJmltdHM9MTczMDkzNzYwMA&ptn=3&ver=2&hsh=4&fclid=22efaef5-d3da-64a6-1ec4-bdb4d22865bc&psq=ipl+score+predictor+research+paper+summaries&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zNjQ2MzE5NTNfRmlyc3RfaW5uaW5nX3Njb3JlX3ByZWRpY3Rpb25fb2ZfYW5fSVBMX21hdGNoX3VzaW5nX21hY2hpbmVfbGVhcm5pbmc&ntb=1**](https://www.bing.com/ck/a?!&&p=460dd274e4c50ed771c84752951617b4ad159eba0e9aa55c32a2a62c3c0eb10cJmltdHM9MTczMDkzNzYwMA&ptn=3&ver=2&hsh=4&fclid=22efaef5-d3da-64a6-1ec4-bdb4d22865bc&psq=ipl+score+predictor+research+paper+summaries&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zNjQ2MzE5NTNfRmlyc3RfaW5uaW5nX3Njb3JlX3ByZWRpY3Rpb25fb2ZfYW5fSVBMX21hdGNoX3VzaW5nX21hY2hpbmVfbGVhcm5pbmc&ntb=1)
5. [**https://www.bing.com/ck/a?!&&p=6468441384a0cce5d95d75cbee6716a84d9df3f1658e7f67c683331d7fa68b67JmltdHM9MTczMDkzNzYwMA&ptn=3&ver=2&hsh=4&fclid=22efaef5-d3da-64a6-1ec4-bdb4d22865bc&psq=ipl+score+predictor+research+paper+summaries&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zMjc5MDQwMDlfUHJlZGljdGluZ19PdXRjb21lX29mX0luZGlhbl9QcmVtaWVyX0xlYWd1ZV9JUExfTWF0Y2hlc19Vc2luZ19NYWNoaW5lX0xlYXJuaW5n&ntb=1**](https://www.bing.com/ck/a?!&&p=6468441384a0cce5d95d75cbee6716a84d9df3f1658e7f67c683331d7fa68b67JmltdHM9MTczMDkzNzYwMA&ptn=3&ver=2&hsh=4&fclid=22efaef5-d3da-64a6-1ec4-bdb4d22865bc&psq=ipl+score+predictor+research+paper+summaries&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi8zMjc5MDQwMDlfUHJlZGljdGluZ19PdXRjb21lX29mX0luZGlhbl9QcmVtaWVyX0xlYWd1ZV9JUExfTWF0Y2hlc19Vc2luZ19NYWNoaW5lX0xlYXJuaW5n&ntb=1)
6. [**https://www.irjmets.com/uploadedfiles/paper/issue\_4\_april\_2022/20791/final/fin\_irjmets1651051572.pdf**](https://www.irjmets.com/uploadedfiles/paper/issue_4_april_2022/20791/final/fin_irjmets1651051572.pdf)
7. **https://ijarcce.com/wpcontent/uploads/2024/04/IJARCCE.2024.134144.pdf**
8. **https://www.irjweb.com/IPL%20SCORE%20PREDICTION.pdf**
9. **https://www.irjmets.com/uploadedfiles/paper/volume3/issue\_5\_may\_2021/10362/1628083416.pdf**