|  |
| --- |
| Electrical & Computer Engineering & Computer Science  University of New Haven [Acceptance Rate + Statistics] |

|  |  |
| --- | --- |
| SPRING 22 |  |



Contents

[Project Name 2](#_Toc96341549)

[Executive Summary 2](#_Toc96341550)

[Technical Report 3](#_Toc96341551)

[Highlights of Project 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341552)

[Submitted on: 3](file:///C:\Users\ardia\Desktop\TEACHING\SPRING%2022\DSCI6002%20Data%20Science\Projects\Technical%20report%20Template.docx#_Toc96341553)

[Abstract 4](#_Toc96341554)

[Methodology 5](#_Toc96341555)

[Results Section 5](#_Toc96341556)

[Discussion 6](#_Toc96341557)

[Conclusion 6](#_Toc96341558)

[Contributions/References 7](#_Toc96341559)

|  |
| --- |
| IPL Data Analytics for New IPL Franchise |

|  |  |  |
| --- | --- | --- |
| Executive Summary In India, one of the most played sports is cricket. Indian cricket standards reached their highest point once the IPL began, and many gifted players had the opportunity to highlight their abilities on a stage like the IPL where many foreign cricketers compete. So, we decided to predict player performances, team performance and chance of winning the cup using machine learning algorithm by which a new IPL franchise can easily decide what to do in their beginning course of actions.  GitHub Project Link: <https://github.com/DSCI6007-Heavenhikers/Final_Project> | | |
| person at a table writing in a notebook with people around | | |
|  | **Questions?**  Contact : |  |

|  |
| --- |
| Technical Report |

|  |  |
| --- | --- |
| IPL DATA ANALYTICS FOR  NEW IPL FRANCHISE |  |
| **Highlights**   * Predicting player performances so that a new IPL franchise can decide which player to pick and which player they can avoid. * This also helps teams to analyze their performance by comparing it with other franchises based on their scores. * We also included a lot of visualizations describing various features of dataset through Exploratory data analysis. * We used flask app for deploying the model for better interaction and usage. |

## Abstract

Finding useful insights from the provided dataset is the crux of data science and analytics. In other words, seeking solutions that might benefit business. In this article, we will look at how to begin using Python for data analysis. NumPy, Pandas, Matplotlib, and Seaborn are the Python packages that we use in this notebook.

We used the IPL (Indian Premier League) Dataset provided on Kaggle Datasets, which was sourced from cricsheet. Since the IPL is one of the most well-known cricket competitions in the world, anyone who is familiar with cricket should be able to relate to the issues we aim to address and the inquiries we make.

[Here you can find a few useful tips on](https://www.youtube.com/watch?v=bZTWx2bftaw) producing a great pitch

Executive Summary

1. First we preprocessed the data by cleaning the data such as adding the missing values and dropping unnecessary Nan values.

2. Later we performed several plotting for better understanding of dataset.

3. Next step is model training and testing

4. We used 8 ML algorithms for predicting

5. We used 5-fold cross validation techniques.

6. Used super-learning to combine the best model into single layer.

7. used pickle library to store the output locally.

8. Finally deployed the model using flask application

Introductory Section

Indian Premier League (IPL) is the most-attended cricket league in the world and all franchises in the IPL needs help for finding best players. By using the IPL dataset, we can decide whether to buy a player based on their statistics.

Review of available research

The entire IPL dataset is thoroughly analyzed, and numerous features required for IPL evaluation are visualized. The winner of any two teams has been compared using a variety of machine learning algorithms. There are few models that attempt to rank players using either straightforward formulas or a small number of mathematical models. In the absence of substantial amounts of valuable data sets, efficiency was quite low. This is because when these models were proposed, there simply was not enough data. The T20 game has specific needs that the present models cannot meet.

## Methodology

## We used 8 machine learning algorithms Logistic Regression, Random Forest, KNeighbors, Decision Tree, Ada Boost, Bagging, Gradient Boosting, XGBoost. 5-fold cross validation is being used. PCA does not help much but worse the condition in this dataset. Used Super leaner to combine our best model into a single layer. The highest accuracy we can achieve is 94 percent with an Super Leaner which is pretty Good!

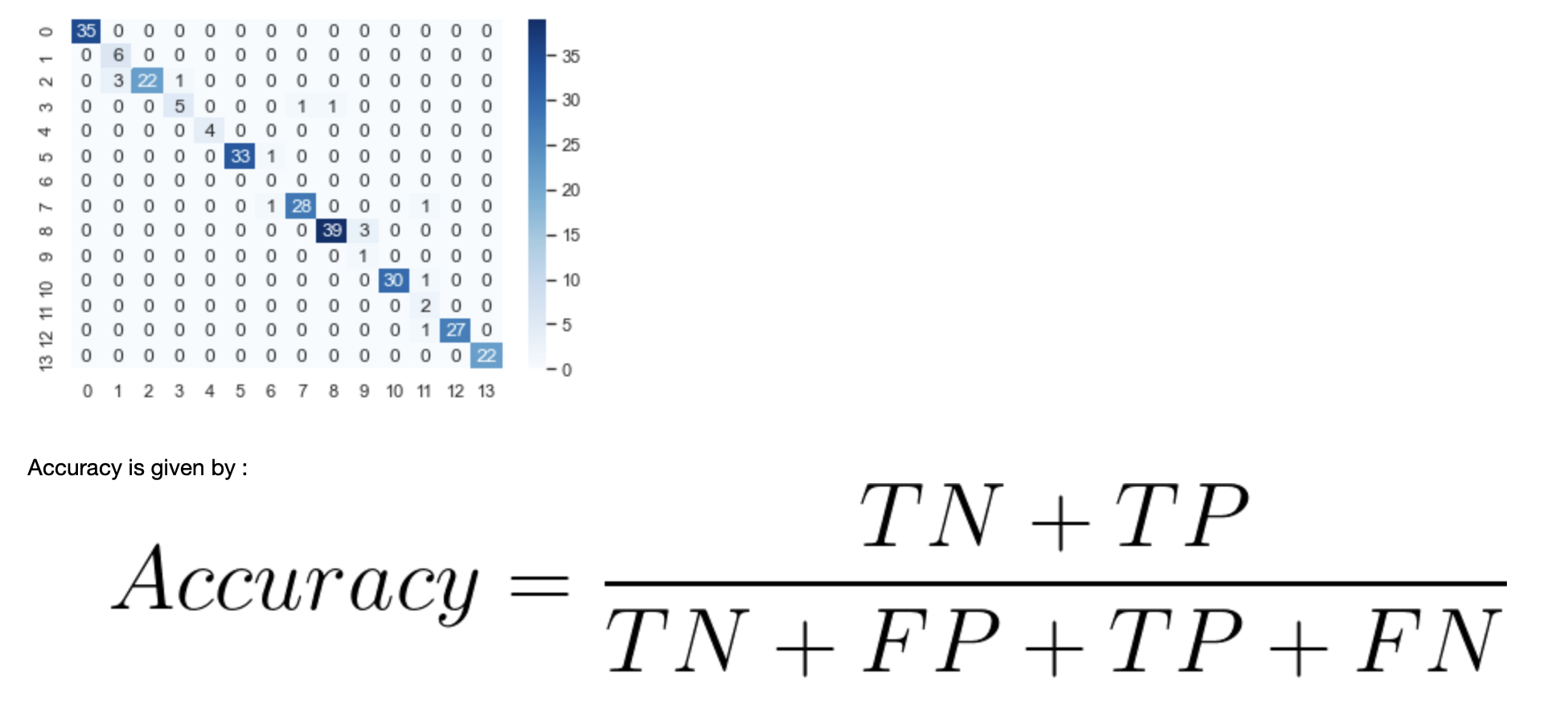
## Results Section

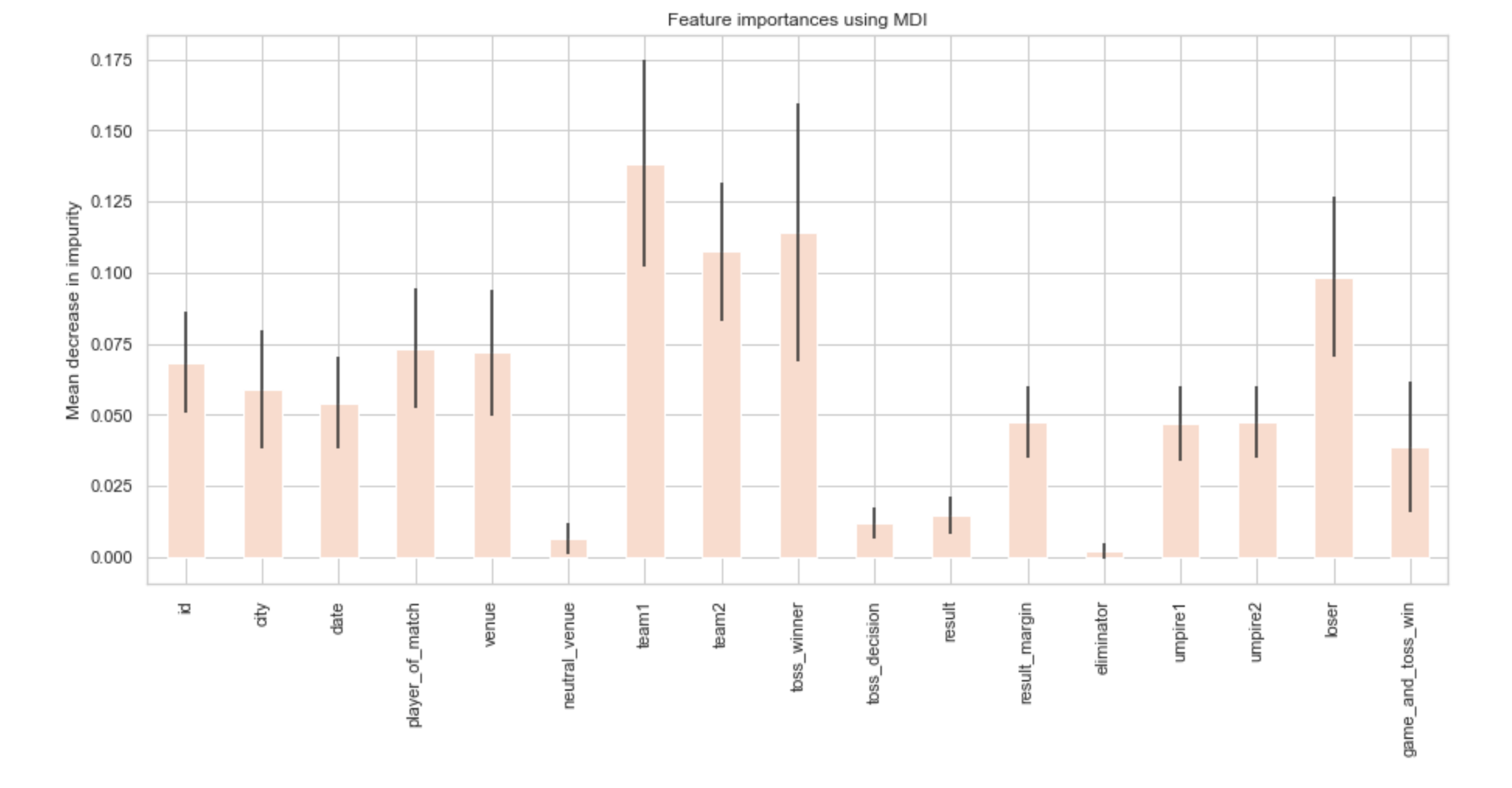
The accuracy is 94%, approximately 3.2 percent of losing teams are classified as winning. By using this example, we are attempting to demonstrate that accuracy is not a good metric when the data set is uneven. In such cases, using accuracy might lead to a false interpretation of the findings.

R^2 = coefficient of determination   
RSS = sum of squares of residuals   
TSS = total sum of squares   
R^2 = 1 - (RSS/TSS)



A good classifier should have a r2\_score of 1 (high). As the numerator sum of squares of residuals (RSS) becomes lower than precision tends to 1.





Observation reveals that traits such as team1, team2, loser and toss winner are more relevant than others.

## Discussion

In this report, we present the outcomes of a thorough examination of several elements of metrics used in the game together with practical visualizations, utilizing a detailed dataset of all the matches played in the history of the IPL. We had difficulties ranking the players, but we overcome them by modeling their strengths and weaknesses against specific opponents, their performance on a specific surface, etc. These are facts that can be especially useful and can provide the team with a significant competitive advantage. The players have also been ranked using machine learning methods according to the Player Ranking Index. Using the suggested technique, forecasts' accuracy rose to 94%.

## Conclusion

The goal is to create a model that uses the best machine learning strategy to predict who will win the game. Also, our objective is to provide reliable prediction with fewer features and tests. Only 17 essential characteristics are considered in this study. Different categorization methods were applied. Gradient boosting, Ada boost, Boost, and bagging are some of the techniques used. The data was pre-processed before being utilized in the model. In this paradigm, the Super Learner approach produced the greatest results. Other approaches, like clustering, association rules, and evolutionary algorithms, may be used to broaden this. Given the limitations of this study, it is necessary to create more complex models that combine diverse types of data to achieve greater accuracy for early match prediction.

## Contributions/References

<https://www.kaggle.com/datasets/patrickb1912/ipl-complete-dataset-20082020>

<https://ieee-dataport.org/documents/ipl-dataset>