A CAPSTONE PROJECT REPORT ON

IPL FIRST INNINGS SCORE PREDICTOR USING MACHINE ALGORITHMS

**NAME: N S ASWIN**

**EMAIL:nsaswin17@gmail.com**

ABSTRACT

ABSTRACT India's most popular sport is cricket and is played across all over the nation in different formats like T20, ODI, and Test. The Indian Premier League (IPL) is a national cricket match where players are drawn from regional teams of India, National Team and also from international team. Many factors like live streaming, radio, TV broadcast made this league as popular among cricket fans. The prediction of the outcome of the IPL matches is very important for online traders and sponsors. We can predict the match between two teams based on various factors like team composition, batting and bowling averages of each player in the team, and the team's success in their previous matches, in addition to traditional factors such as toss, venue, and day-night, the probability of winning by batting first at a specified match venue against a specific team. In this paper, we have proposed a model for predicting outcome of the IPL matches using Machine learning Algorithms namely SVM, Random Forest Classifier (RFC), Logistic Regression and K-Nearest Neighbor. Experiment

Data set:

<https://github.com/Aswin173/Capstone-Stone-Project.git>

Source Code:

<https://github.com/Aswin173/Capstone-Stone-Project.git>

# ACKNOWLEDGE

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of my capstone project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, I have fortunate to have Mr. ANBU JOEL as my mentor. He has readily shared his immense knowledge in data analytics and guide me in a manner that the outcome resulted in enhancing my data skills.

I certify that the work done by me for conceptualizing and completing this

project is original and authentic.

Date: 1/7/2022 Name: N.S.ASWIN

TABLE OF CONTENTS :

|  |  |  |
| --- | --- | --- |
| CHAPTER NO: | TITLE | PG.NO |
| 1  2  3  3.1  3.2  4  4.1  5  5.1  6 | ABSTRACT  ACKNOWLEDGE  INTRODUCTION  LITERATURE SURVEY  DATASET FEATURES  FEATURE SELECTION  ALGORITHMS  IMPLEMENTATION OF GUI  SCORE PREDICTION  PROGRAM OF THE PROJECT  DEPLOYMENT PROGRAM  CONCLUSION | 2  3  4  6  8  9  10  11  12  13  23  25 |

1. INTRODUCTION

Cricket is an outdoor game which is played by bat and bowl which includes 2 teams of 11 players each. Cricket is a teamwork game and is played mostly in three formats and occupies the 2 spots in the list of the most popular sport around the World. Like in any sport, there are many factors that plays an importantrole in deciding the winner of the match. Selection of a team is based on the player performance and other considerations like pitch factor, team size, venue etc. There are many variables and constraints which makes The Analysis of Cricket Match Difficult. There are three different formats of Cricket namely - Tests, Twenty-twenty (T20) and One Day International (ODI). Cricket is not only a nation game but also an international game. In this game, every ball is crucial because every ball can change the whole match in Cricket [15, 16]. Indian Premier League (IPL) is a national cricketmatch where players are drawn from regional teams of India, National Team and also from international team. It is based on 20-20 format and is owned by Celebrities, Businessmen and others and the entire IPL is controlled by Board of Control for Cricket in India (BCCI). For the current year (2021)there are total of 8 Teams in IPL namely, Royal Challengers Bangalore (RCB), Rajasthan Royals (RR), Chennai Super Kings (CSK), Mumbai Indians (MI), Kolkata Knight Riders (KKR), Delhi Capitals (DC), Punjab Kings (PK) and SunRisers Hyderabad (SRH). The motivation behind this paper includes the answers to following questions: “What is the probability of winning the game at a particular venue based on decision to field/bat first on winning the toss?”, “Most dismissals by a bowler in a match?”, “Does Home Ground have any effect on the result of the game? In this paper We are trying to find out the match winner of an IPL match based on the stadium they are choosing and the toss decision using machine learning techniques like Linear Regression , Ridge Regression, Lasso Regression and XGBoost . Remainder of the paper is organized as follows: The section 2 is the literature survey, section 3 deals with the problem definition and the architecture. Section 4 deals with the experimental results. Section 5 talks about the conclusion.

2. LITERATURE SURVEY

Ahmad et al. [1], predicted the emerging players from batsman as well as from the bowlers using machine learning techniques. Song et al. [2] predicted estimation of the location of a moving ball based on the value of the cricket sensor network. Roy et al. [3] predicted ranking system which is based on the social network factors and their evaluation in the form of composite distributed framework using Hadoop framework and MapReduce programming model is used for processing the data. Priyanka et al. [4], predicted the outcome of IPL-2020 based on the 2008-2019 IPL datasets using Data Mining Algorithms with an accuracy of 82.73%. Kansal et al. [5], predicted player evaluation in IPL based on the 2008-2019 datasets using Data Mining Technique. Data mining algorithms are used which gives evaluation using player statistics assessing a player's performance and determining his base price. Kaluarachchi et al. Passi et al. [9], predicted the performance of players based on the runs and the number of wickets. Both the type of problems is treated as classification problems where the list of runs, and list of wickets are classified in different ranges based on machine learning algorithms. Nigel Rodrigues et al. [10], predicted the value of the traits of the batsmen and the bowlers in the current match. This would help in selecting the players for the upcoming matches by using past performances of a player against a specific opposition team by using Multiple Random Forest Regression. Wright [11], predicted the possible fixture for a cricket match based on the various venue, teams, number of holidays between each match in a fair and efficient manner. A metaheuristic procedure is used to progress from the basic solution to a complex final solution by a technique, Subcost-Guided Simulated Annealing (SGSA). Maduranga et al. [12], predicted the outcome of any cricket match by using data mining algorithms and provided solutions for the approach used by other authors. Shetty et al. [13], predicted the capabilities of each player depending on various factors like the ground, pitch type, opposition team and several others by using machine learning techniques. This model helped them to select the best players of the game and predict outcomes of the match.

3 .DATASET FEATURES :

The approach over here we are using is ML based. So the basic requirements of an ML algorithm is dataset, training of that dataset using the algorithm and testing the model. So, we have imported dataset from Kaggle. Later on calculating the accuracy and improving the accuracy by using Random Forest classifier for winning prediction and Linear Regression for score prediction Score Prediction:- For conducting our research, we collected data on all the IPL matches played in 2008. The dataset consists of 76015 numbers of rows. Dataset consists 15 columns over which we applied feature selection techniques and selected 8 features in which 7 are input feature and 1 is our target variable. The attributes selected were bat team, bowl team, overs, runs, wickets, runs in prev 5, wickets in previous 5 for score prediction.

3.1 Dataset Attributes and their values

|  |  |
| --- | --- |
| Attributes | Values |
| Batting Team | Batting Team Name among 8 teams in IPL |
| Bowling Team | Bowling Team Name among 8 teams in IPL |
| Overs | Value > 5 Over |
| Runs | 0-300 |
| Wickets | 0-10 |
| Run Scored in last 5 overs | 0-300 |
| Wickets fall in last 5 overs | 0-10 |
| Total Runs | 0-300 |

3.1 : FEATURE SELECTION :

Feature Selection is process in which we select an optimal set of features from input features set by using feature selection techniques. By removing redundant features, we reduce dimension of data and we can improve time and space complexity of data. Feature selection improves the performance of model and saving time and space.

The first step is to import the dataset using the pandas library and then further preprocess the dataset by checking the null values and replace it with the mean or median values of the respective column. The categorical data in the columns is mapped into numerical values. After that the feature selection techniques are applied to the dataset and select only the optimal set of features. The set of input features (X) and the output (Y) are defined in the dataset. The input features are independent of each other and the output feature depends on the input features. Then the library is imported, and the ML Model is created, and the train-split-test method is used to separate the data into training data and testing data and then train the ML Model with training data and predict using test data. Then the accuracy of the model is calculated by simply taking the ratio of the predicted testing data and the actual testing data. This method is repeated with each ML Algorithm and the accuracy of each algorithm is calculated. Finally, the accuracy of the algorithms is compared and then which of the algorithms is the best for this dataset is determined.

3.2 : ALGORITHMS :

We tried to use two machine learning techniques: regression and classification. Selected algorithms from each technique were trained then. Regression: Regression analysis uses various algorithm for the computation and based on that it predicts the continuous value. There are certain set of variables are used for the input and the continuous range value is the target variable. Based on the application different regression algorithms are used. There are different regression techniques. Out of which the linear, ridge and lasso algorithms are used for predicting the score . Linear regression: To predict the continuous values, Linear regression is used. Certain known parameters are given to the machine learning algorithms, it predicts the continuous values as output. It cannot used for the classification problems. The proposed model predicts the score using the Linear Regression.

Ridge regression: Ridge regression is also used to predict the continuous values. When the variables used for the prediction greater than the observations of when multicollinearity present in the data, ridge regression is used set has multicollinearity (correlations between predictor variables). Lasso regression: Lasso regression is a type of linear regression that used for predicting the continuous values. Shrinkage is used in the lasso regression. When data values focus towards central point shrinkage occurs. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models.

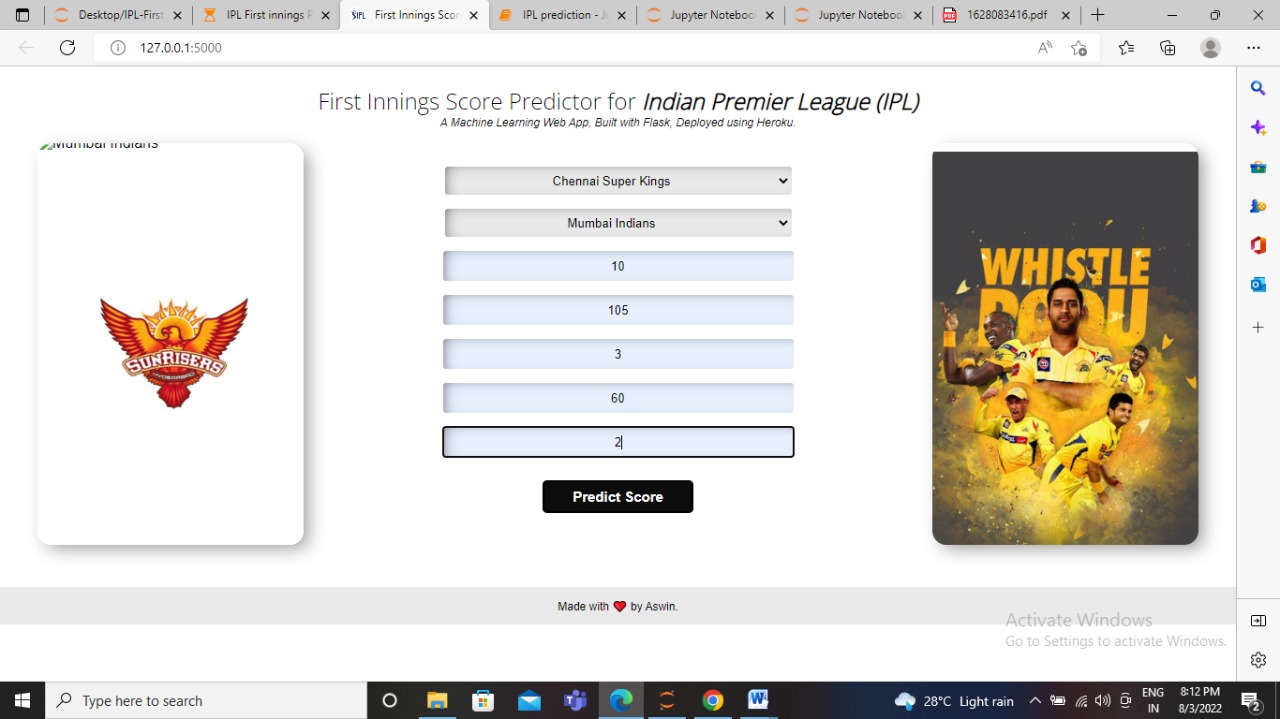
4. IMPLEMENTATION OF THE GUI

The Graphical User Interface is developed for the machine learning models using the Flask Framework. For the backend of the site Python is used. The site can be used to predict the IPL match score with the help of last 5 overs of the data. We can also predict the Winner of the match with the data of just Toss Winner and Toss Decision. All the input information necessary for the model for the prediction is provided to the model. The calculation is not stored in the system because all calculations computed at real time. We implemented it that way as we can add change more attribute to the system with minor changes to the program. A. Score Prediction The GUI required at least 5 overs of the data to predict the score as shown in the Fig 1.1. Model require the input data of Batting team, Bowling team, Over, Runs, Wickets, Run Scored in last 5 overs, Wickets fall in last 5 overs to predict the score of the match as shown.

4.1 : Score Prediction:

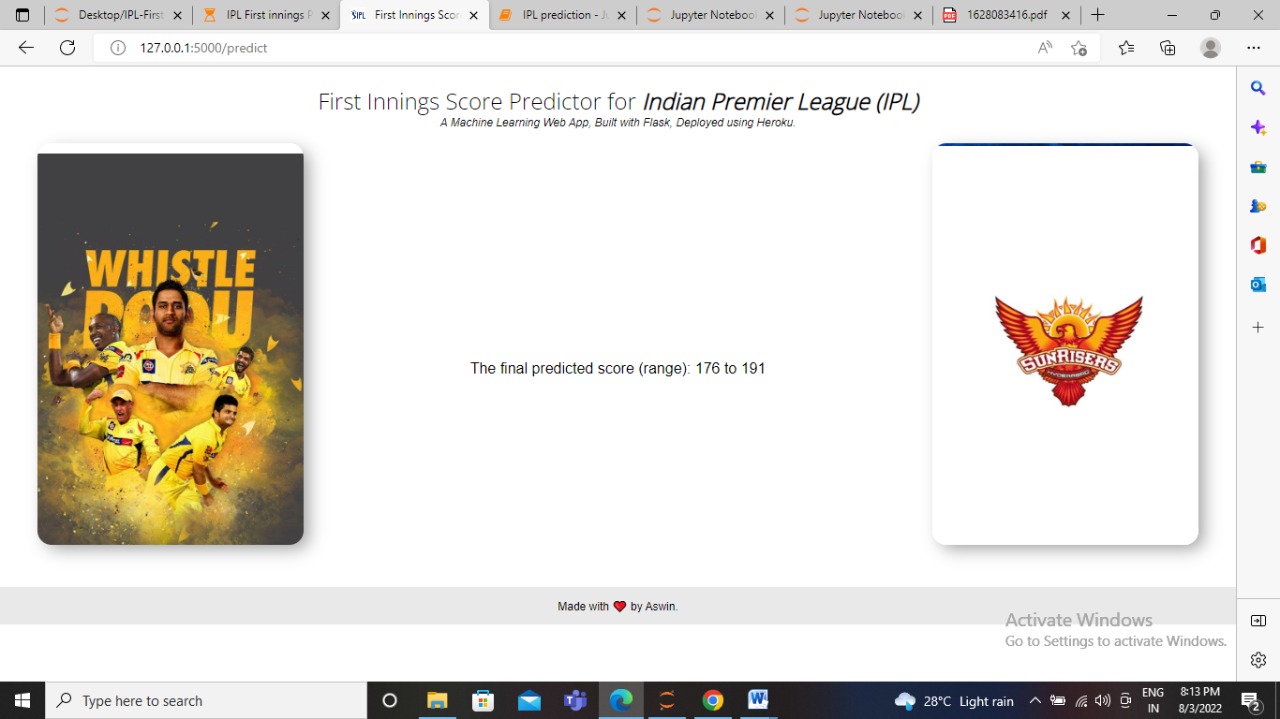
Step 1:

Input to Score Prediction.

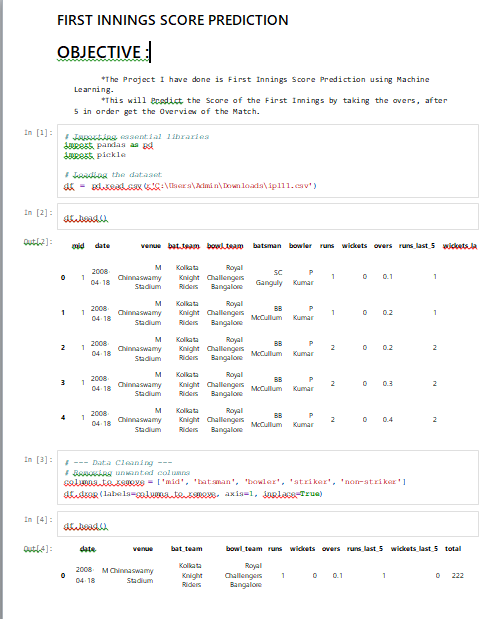


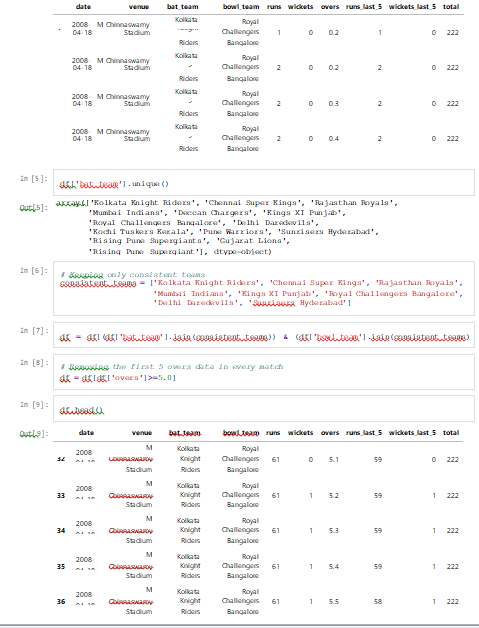
Step 2:

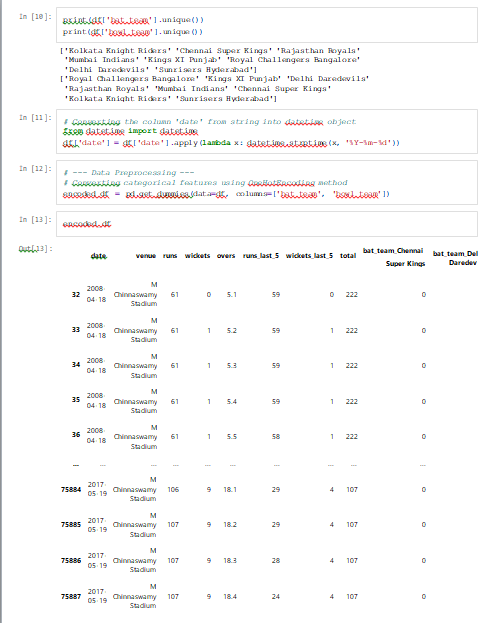
Score Prediction Output:

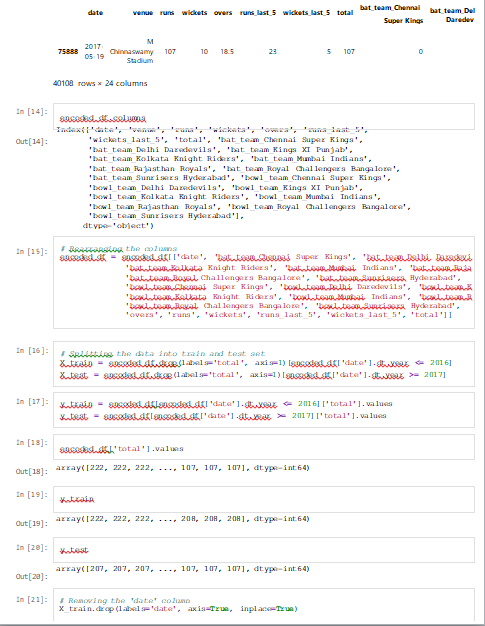


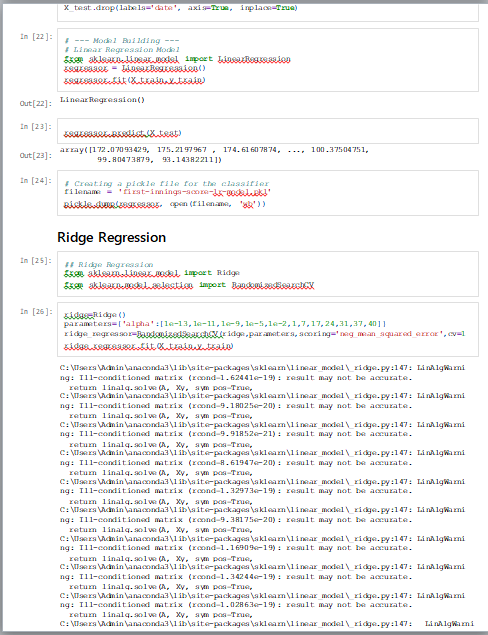
5 . PROGRAM OF THE PROJECT:

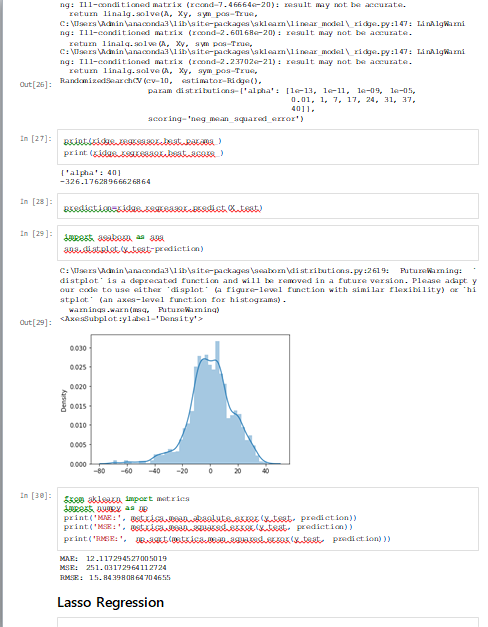






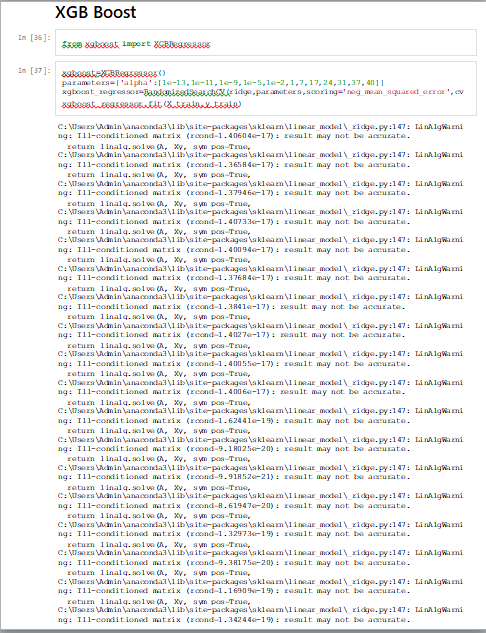


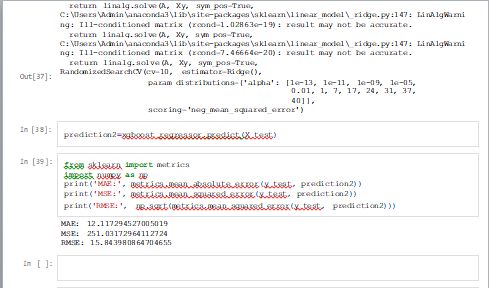






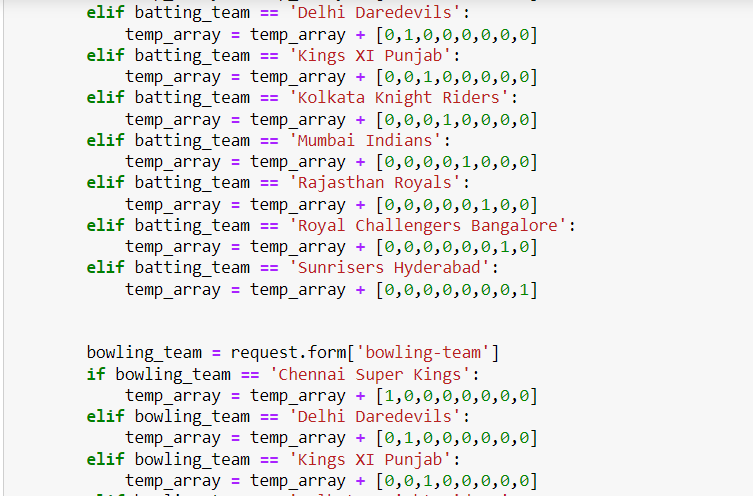


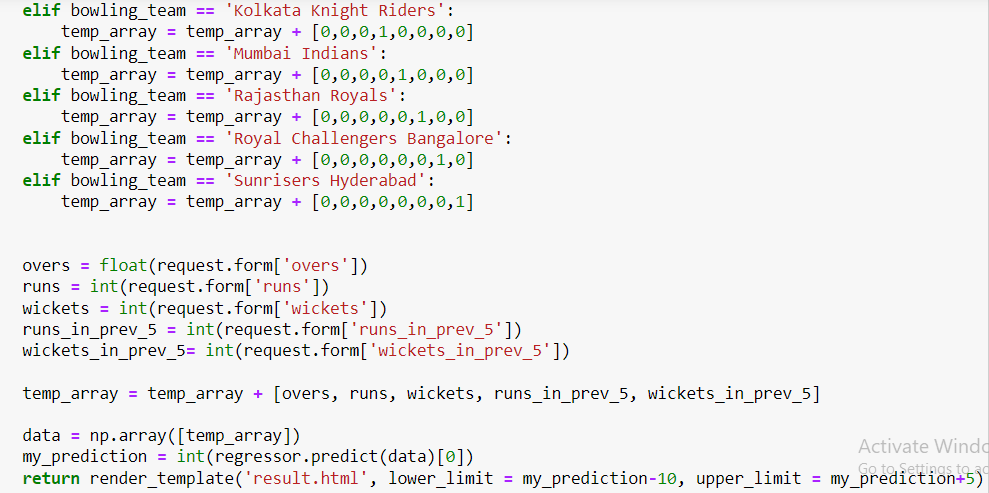


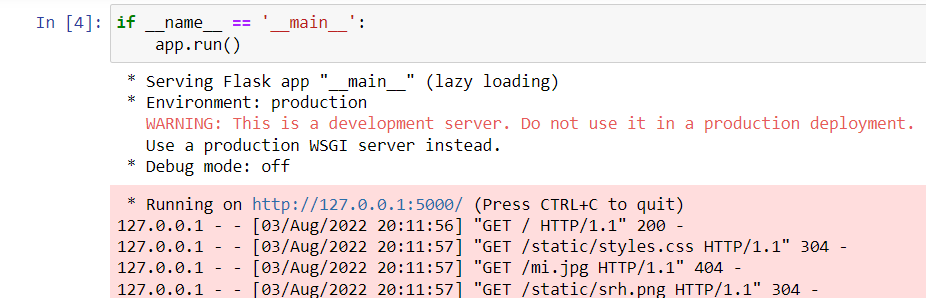


5.1 : DEPLOYMENT PROGRAM:









6.CONCLUSION :

This paper will give the important information regarding IPL score prediction and winning prediction system, that which parameters are required also the classifiers and algorithms. it helps in mathematical operation. Using all the information we have developed a website. for that the important work we have to do for the model is comparative analysis of machine learning techniques that is for score prediction the regressions and for winning prediction the analysis of classifiers. In Score Prediction analysis accuracy of Linear Regression is more than Ridge and Lasso Regression.So I used Linear Regression For Deployment Model.