DMW MINI PROJECT

Group members: 41149, 41151, 41152

Title: IPL match winner prediction

Objective: The objective is to predict result (Winner) of IPL match .

Problem statement: To predict IPL match winner with previous year matches dataset preprocessing it and applying suitable machine learning algorithms.

Outcome:

- To be able to apply various machine learning algorithms to predict IPL match winner.
- To understood the concepts behind ML classification algorithms.
- To preprocesse the data(removal of null entries,encoding etc) .

Software and hardware requirements:

- Ubuntu (Linux Distribution) / windows10/Fedora 20
- 8GB RAM
- Jupyter notebook

Theory concept with algorithm:

Preprocessing data:

Dataset which we are using is available on Kaggle, it consists of IPL data from year 2008 to 2019.Features include team names,toss winner ,toss decision etc and output label being match winner.

Before implementing machine learning algorithms on our data, we went through a series of preprocessing steps. These included:

- Removing unnecessary columns
- Removing the result where Results is no result and entries with null values present in any feature .
- Check whether team1, team2 and winner have same value or not
- Calculating Batting average for each team
- Merging batting average score into final dataset
- Calculating Bowling average for each team
- Merging Bowling average into dataset.
- Conversion of Winning Team and Toss winning Team labels to Integer
  0: indicates Team1 1: indicates Team2
- Changing some team's name and deleting teams which no longer play
- Applying one-hot encoding on string columns to convert it to binary integer values
- To scale the data using Standard Scaler

Algorithms:

1.Random forest Classification

Random forests is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance. Random forests has a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

2. Naïve Bayes :

It is a characterization method upheld Bayes' Theorem with a presumption of autonomy among indicators. In basic terms, a Naive Bayes classifier expect that the nearness of a particular component during a class is random to the nearness of the other element. Naive Bayes model is straightforward to create and particularly useful for very large data sets. Along with simplicity, Naive

Bayes is understood to outperform even highly sophisticated classification methods. Bayes theorem provides how of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:

Naive Bayes, more technically referred to as the Posterior Probability, updates the prior belief of an event given new information. The result is the probability of the class occurring given the new data.

$$P(c/x) = \frac{P(c) * P(x/c)}{P(x)}$$

- P(c/x) : Posterior Probability
- P(c) : Class Prior Probability
- P(x/c) : Likelihood
- P(x) : Predictor Prior Probability

3. Support vector machine (SVM):

Support Vector Machine is an administered ML calculation which can be utilized for each arrangement or relapse difficulties. Be that as it may, it's basically utilized in order issues. Inside the SVM algorithmic program, we plot each data thing as some degree in n-dimensional zone (where n is assortment of choices you have) with the value of each component being the cost of a particular facilitate. Then, we have a tendency to perform classification by finding the hyper-plane that differentiates the 2 categories. Support Vectors square measure merely the co-ordinates of individual observation. The SVM classifier may be a frontier that best segregates the 2 categories (hyper-plane/ line).

4. K nearest neighbours (KNN)

The k-nearest neighbour (K-NN) classifier is taken into account an example-based classifier, meaning that the training documents are used for comparison instead of a particular classification portrayal, similar to the classification profiles utilized by different classifiers. Accordingly, there is no genuine preparing stage. At the point when a substitution report must be sorted, the k most comparable archives (neighbours) are found and if an enormous enough extent of them have been allocated to a specific classification, the new record is moreover appointed to the current class, in

any case not. Furthermore, finding the nearest neighbours are frequently animated utilizing conventional ordering techniques. To choose whether a message is spam or ham, we look at the classification of the messages that are nearest there to it. The examination between the vectors is a continuous procedure.

Test cases:

Algorithm	Cross Validation Accuracy Scores	Standard deviation
Random forest classification	**64.71%**	**5.33%**
Naïve bayes	**62.80%**	**4.79%**
SVM	**62.94%**	**6.11%**
KNN	**61.14%**	**5.39%**

Conclusion:

We successfully implemented the prediction problem of IPL matches using ML classification algorithms.We successfully carried out data preprocessing (cleaning, encoding merging different datasets etc).The maximum accuracy attained was 64.71% with Random Forest classifier.