ABSTRACT

Cricket is a very familiar and exciting sport that people of all age groups are insane to see and play. For many it's a billion-dollar market as they speculate financially, hoping to be able to earn profit in the form of gambling and various other ways. In this project, a model using machine learning algorithms is proposed to predict the score of each match and winning team based on past datasets available from 2008 to 2019 IPL matches in Kaggle. This proposed methodology includes the following steps like Pre-processing of collected datasets, Feature selection from raw data, Conversion of categorical data into numerical data, Partitioning of samples into training and test samples, Training, and classification. Few machine learning algorithms like Support Vector Machine, Random Forest, Naive Bayes were already used in previous papers. In this project, algorithms like Lasso Regression, Ridge Regression, and Random Forest regression models are proposed for a score prediction, and SVM(Linear, RBF), Logistic Regression classifier is for the match-winning prediction. The accuracy of the above machine learning algorithms is used to predict the winner of an IPL match along with its Precision, Recall and F-Measure measured and the model with better accuracy is considered.

Keywords : IPL, Machine Learning, Match winner prediction, Score Prediction, SVM, kNN, Naive Bayes.

CONTENTS

ABSTRACT R5

LIST OF SYMBOLS R8

LIST OF FIGURES R9

LIST OF TABLES R10

LIST OF ABBREVIATION R11

CHAPTER 1 INTRODUCTION 1

1.1 Introduction 2

1.1.1 Indian Premier League (IPL) 2

1.1.2 Machine Learning 2

1.1.3 Flask 5

1.2 Motivation of the work 6

1.3 Problem Statement 6

1.4 Organization of Thesis 6

CHAPTER 2 LITERATURE SURVEY 7

2.1 A Perspective on Analyzing IPL Match Results using Machine Learning 8

2.2 Predictive Analysis of IPL Match Winner using Machine Learning Techniques. 10

2.3 Predicting Outcome of Indian Premier League(IPL) Matches Using Machine Learning 12

CHAPTER 3 METHODOLOGY 15

3.1 Proposed System 16

3.1.1 Data Acquisition 16

3.1.2 Data Cleaning 19

3.1.3 Feature Selection 20

3.1.4 Training Classification Methods 21

3.1.5 Testing Data 24

3.2 User Interface 24

CHAPTER 4 EXPERIMENTAL ANALYSIS AND RESULTS 25

4.1 System Configuration 26

4.1.1 Software Requirements 26

4.1.1.1 Introduction to Python 26

4.1.1.2 Introduction to Flask Framework 26
4.1.1.3 Python Libraries 27
4.1.2 Hardware Requirements 28
4.2 Code 29
4.3 Experimental analysis and Performance Measures 68
4.3.1 Performance Analysis and Models Comparison 68
4.3.1.1 Methods Comparison 69
4.4 Results 70

CHAPTER 5 CONCLUSION AND FUTURE WORK 73

5.1 Conclusion 74

5.2 Future Work 74

REFERENCES 75

LIST OF FIGURES

Fig.No.	Topic Name	Page No.
3.1	System Architecture	16
3.2	Sample Data Points of IPL Ball to Ball data acquired from Kaggle website	17
3.3	Sample Data Points of IPL Match data acquired from Kaggle website	18
3.4	Importing required packages	18
3.5	Reading the dataset from google drive	19
3.6	Encoding categorical values	20
3.7	Removing Null and duplicated values and dropping unnecessary features	21
3.8	Common function for training model	21
3.9	Support Vector Machine(SVM) training model	22
3.10	kNN model representation	22
3.11	k Nearest Neighbours (kNN) training model	23
3.12	Naive Bayes training model	24
4.1	Precision and Recall	69
4.2	Index Page	70
4.3	Match Prediction Page	70
4.4	Match Prediction Result Page	71
4.5	Score Prediction Page	71
4.6	Score Prediction Result Page	72

1.1 Introduction

1.1.1 Indian Premier League (IPL)

Sports have gained much importance at both national and international level. Cricket is one such game, which is marked as the prominent sport in the world. T20 is one among the forms of cricket which is recognized by the International Cricket Council (ICC). Because of the short duration of time and the excitement generated, T20 has become a huge success. The T20 format gave a productive platform to the IPL, which is now pointed as the biggest revolution in the field of cricket. IPL is an annual tournament usually played in the months of April and May. Each team in IPL represents a state or a part of the nation in India. IPL has taken T20 cricket's popularity to sparkling heights.

It is the most attended cricket league in the world and in the year 2010, IPL became the first sporting event to be broadcasted live. Till date, IPL has successfully completed 13 seasons from the year of its inauguration. Currently, there are 8 teams that compete with each other, organized in a round robin fashion during the stages of the league. After the completion of league stages, the top 4 teams in the points table are eligible to the playoffs. In playoffs, the winner between 1st and 2nd team qualifies for the final and the loser gets another opportunity to qualify for the finals by playing against the winner between 3rd and 4th team. In the end, the 2 qualified teams played against each other for the IPL title. The significance is that IPL employs television timeouts and therefore there is no time constraint in which teams have to complete the innings.

In this paper, we have examined various elements that may affect the outcome of an IPL match in determining the runs for each ball by considering the runs scored by the batsman in the previous ball as the labeled data. The suggested prediction model makes use of SVM and KNN to fulfill the objective of the problem stated. Few works have been carried out in this field of predicting the outcomes in IPL. In our survey, we found that the work carried out so far is based on Data Mining for analyzing and predicting the outcomes of the match. Our work novelty is to predict runs for each ball by keeping the runs scored by the batsman in the previous ball as the observed data and to verify whether our prediction fits into the desired model.

1.1.2 Machine Learning

Machine Learning is the preferred technique of predicting or classifying information to assist folks in creating necessary selections. Machine Learning algorithms are trained over instances or examples through that they learn from past experiences and analyse the historical knowledge. Simply building models isn't enough. you want to conjointly optimize and tune the model appropriately in order that it provides you with

correct results. Improvement techniques involve tuning the hyperparameters to succeed in Associate in Nursing optimum results.

As it trains over the examples, once more and once more, it will determine patterns to form selections additionally accurately. Whenever any new input is introduced to the cubic centimetre model, it applies its learned patterns over the new knowledge to form future predictions. Based on the ultimate accuracy, one will optimize their models by exploiting numerous standardized approaches. During this manner, the Machine Learning model learns to adapt to new examples and produce higher results.

Types of Learnings

Machine Learning Algorithms can be classified into 3 types as follows:

1. Supervised learning
2. Unsupervised Learning
3. Reinforcement Learning

1.1.2.1 Supervised Learning

Supervised learning is the preferred paradigm for machine learning. It is the simplest to know and therefore the simplest to implement. It is the task of learning a function that maps an input to an output supported example input-output pairs. It infers a function from labelled training data consisting of a group of coaching examples. In supervised learning, each example may be a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyses the training data and produces an inferred function, which may be used for mapping new examples. Supervised Learning is very similar to teaching a child with the given data and that data is in the form of examples with labels, we can feed a learning algorithm with these example label pairs one by one, allowing the algorithm to predict the right answer or not. Over time, the algorithm will learn to approximate the exact nature of the relationship between examples and their labels. When fully trained, the supervised learning algorithms are going to be ready to observe a replacement , never-before-seen example and predict an honest label for it.

Most of the sensible machine learning uses supervised learning. Supervised learning is where you've got input variable (x) and an output variable (Y) and you employ an algorithm to find out the mapping function from the input to the output.

$$Y = f(x)$$

The goal is to approximate the mapping function so well that once you have a new input file (x) that you simply can predict the output variables (Y) for the info . It's called supervised learning because the method of an algorithm learning from the training dataset is often thought of as an educator supervising the training process. Supervised learning is usually described as task oriented. It's highly focused on a singular task, feeding more and more examples to the algorithm until it can accurately

perform the on task. This is often the training type that you simply will presumably encounter, because it is exhibited in many of the common applications like Advertisement Popularity, Spam Classification, and face recognition.

Two types of Supervised Learning are:

(i) Regression:

Regression models a target prediction value supported by independent variables. It's mostly used for locating out the connection between variables and forecasting. Regressions are often wont to estimate/ predict continuous values (Real valued output). For instance , given an image of an individual then we've to predict the age based on the idea of the given picture.

(ii) Classification:

Classification means to group the output into a category . If the info is discrete or categorical then it's a classification problem. for instance , given data about the sizes of homes within the land market, making our output about whether the house "sells for more or but the asking price" i.e. Classifying houses into two discrete categories.

1.1.2.2 Unsupervised Learning

Unsupervised Learning may be a machine learning technique, where you are not got to supervise the model. Instead, you would like to permit the model to figure on its own to get information. It mainly deals with the unlabelled data and appears for previously undetected patterns during a data set with no pre-existing labels and with a minimum of human supervision. In contrast to supervised learning that sometimes makes use of human-labelled data, unsupervised learning, also referred to as self-organization, allows for modelling of probability densities over inputs.

Unsupervised machine learning algorithms infer patterns from a dataset without regard to known, or labelled outcomes. It's the training of machines using information that's neither classified nor labelled and allowing the algorithm to act on information without guidance. Here the task of the machine is to group unsorted information consistent with similarities, patterns, and differences with no prior training of knowledge . Unlike supervised learning, no teacher is as long as it means no training is going to be given to the machine. Therefore, machines are restricted to seek out the hidden structure in unlabelled data by our-self. For instance , if we offer some pictures of dogs and cats to the machine to categorize, then initially the machine has no idea about the features of dogs and cats so it categorizes them consistent with their similarities, patterns and differences. The Unsupervised Learning algorithms allow you to perform more complex processing tasks compared to supervised learning. Although, unsupervised learning is often more unpredictable compared with other natural learning methods.

Unsupervised learning problems are classified into two categories of algorithms:

(i) Clustering: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behaviour.

(ii) Association: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

1.1.2.3 Reinforcement Learning

Reinforcement Learning (RL) may be a sort of machine learning technique that permits an agent to find out in an interactive environment by trial and error using feedback from its own actions and experiences. Machine mainly learns from past experiences and tries to perform the absolute best solution to a particular problem. It's the training of machine learning models to form a sequence of selections . Though both supervised and reinforcement learning use mapping between input and output, unlike supervised learning where the feedback provided to the agent is the correct set of actions for performing a task, reinforcement learning uses rewards and punishments as signals for positive and negative behaviour. Reinforcement learning is currently the foremost effective tool thanks to the machine's creativity.

1.1.3 Flask

Flask is an API of Python that permits us to create web-applications. Flask was created by Armin Ronacher of Pocoo, a world group of Python enthusiasts formed in 2004. Flask's framework is more explicit than Django's framework and is additionally easier to find out because it's less base code to implement an easy web-Application. A Web-Application Framework or Web Framework is the collection of modules and libraries that helps the developer to write down applications without writing the low-level codes like protocols, thread management, etc. Flask is predicated on WSGI(Web Server Gateway Interface) toolkit and Jinja2 template engine. Python 2.6 or higher is required for the installation of the Flask.

Flask supports extensions which will add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and a number of other common framework related tools. Flask is additionally easy to start with as a beginner because there's little boilerplate code for getting an easy app up and running.

1.2 Motivation of the Work

The history of machine learning and technology have always been intertwined. Artistic revolutions which have happened in history were made possible by the tools to make the work. We are entering an age where machine learning is becoming increasingly present in almost every field.

As the audience of IPL is increasing daily, people are looking at trending technologies like data science, big data to deal with predictions. So we gathered the data from the past seasons and made an analysis on the data. We focused on the factors that are affecting the match winning and started to predict the match winner using those features.

1.3 Problem Statement

The main objective is to predict the IPL Match Result that would be beneficial for the franchises and authorities who are at a position of decision making. IPL has a large set of audience. In cricket, particularly IPL is most watched and loved by the people, where no one can guess who will win the match until the last ball of the last over. The main purpose of this research work is to find the best prediction model i.e. the best machine learning technique which will predict the match winner out of the two teams. The techniques used in this problem are k - Nearest Neighbour(kNN), Naïve Bayes and Support Vector Machine(SVM). The experimental study is performed on the dataset of the IPL's patients which is downloaded from kaggle. The prediction is evaluated using evaluation metrics like confusion matrix, precision, recall accuracy and f1-score.

1.4 Organization of Thesis

The chapters of this document describe the following:

Chapter-1 is about the introduction of our project where we have given clear insights about our project domain and other related concepts.

Chapter-2 specifies a literature survey where all different existing methods and models are examined.

Chapter-3 specifies the proposed system with a system architecture along with detailed explanations of each module.

Chapter-4 specifies the experimental analysis of our system along with performance measures and comparisons between different models. It also specifies about implementation along with sample code.

Chapter-5 gives the conclusion to our work with an insight for the future scope.

2.2 Predictive Analysis of IPL Match Winner using Machine Learning Techniques.

Title	Predictive Analysis of IPL Match Winner using Machine Learning Techniques
Authors	Ch Sai Abhishek Ketaki V Patil P Yuktha Meghana K S MV Sudhamani
Year of Publication	Dec,2019
Publishing Details	International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-2S, December 2019