

PROFESSIONAL TRAINING REPORT
at
Sathyabama Institute of Science and Technology
(Deemed to be University)

**Submitted in partial fulfillment of the requirements for the
award For Bachelor of Technology Degree in**

Information Technology

By

DEEPIKA.M
(Reg. No.37120020)



DEPARTMENT OF INFORMATION TECHNOLOGY
SCHOOL OF COMPUTING

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade “A” by NAAC
JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119

AUGUST 2020



MA INSTITUTE OF SCIENCE AND TECHNOLOGY

(Established under Section 3 of UGC Act, 1956)

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai -600 119

SCHOOL OF COMPUTING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of DEEPIKA.M(37120020) who carried out the project entitled “CRICKET WINNING TEAM PREDICTION SYSTEM “Under our supervision from June 2020 to August 2020.

Internal Guide

Mrs .S.GOWRI.,M.E.,(Ph.D.,)

Head of the Department

DR.S.SUBASHINI M.E., Ph.D.,

Submitted for Viva voce Examination held on

Internal Examiner External Examiner

~ 2 ~

DECLARATION

I DEEPIKA.M hereby declare that the Project Report entitled CRICKET WINNING TEAM PREDICTION SYSTEM was done by me under the guidance of Dr.S.Gowri,M.E, Ph.D at SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY is submitted in partial fulfillment of the requirements for the award of Bachelor of Technology degree in INFORMATION TECHNOLOGY.

DATE:

PLACE: CHENNAI SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

The satisfaction and elation that accompany the successful completion of any task would be incomplete without the mention of the people who have made it possible. It is our great privilege to express our gratitude and respect to all those who have guided me during the course of my Professional Training.

First and foremost, we would express our sincere gratitude to our beloved Founder Chancellor **Col. Dr. JEPPIAAR, M.A., B.L., Ph.D.**, and Chancellor **Thirumathi. REMIBAI JEPPIAAR**. I extend my sincere thanks to our Pro Chancellor **Dr. Mariazeena Johnson, B.E., M.B.A., M.Phil., Ph.D.**, and the Vice President **Dr. Marie Johnson, B.E., M.B.A., M.Phil., Ph.D.**, and for providing me the necessary facilities for the completion of the professional training. I also acknowledge our Vice Chancellor **Dr. S.Sundar Manoharan Ph.D.**, and the Pro Vice Chancellor **Dr. T. Sasipraba, M.E., Ph.D.**, for their constant support and endorsement.

I would like to express our gratitude to our Registrar **Dr. S. S. Rau, Ph.D.** and Controller of Examinations **Dr. Igni Sabasti Prabu M.E., Ph.D.**, for their valuable support offered to complete my professional training successfully.

I like to express our gratitude to **Dr. R. Subhashini, M.E., Ph.D.**, Head of the Department of Information Technology, Sathyabama University for having been a constant source of support and encouragement for the completion of the professional training. I would also like to express our sincere thanks to our internal guide **Mrs.S.Gowri., M.E.,Ph.D.**, who guided me in the preparation of the report.

Also, I thank the **Almighty** and my **Parents** for supporting me in the completion of the professional Training.

TRAINING CERTIFICATE



~ 5 ~



Certificate of Internship

To,
Deepika Manimaran
Bangalore

This is to certify that **Deepika Manimaran** interned with **Zolostays Property Solutions Pvt. Ltd.**, under the guidance of **Saket Kumar, Sales Manager**.

Here are the details of the internship:

Name of the Department	Sales
Last Designation Held	Sales Intern
Date of Joining	1-Jul-2020
Date of Leaving	31-Aug-2020

In case of any queries, please feel free to reach out to hr@zolostays.com.

Yours faithfully,
For **Zolostays Property Solutions Pvt. Ltd.**



Akhil Sakri
Director
Date: 31-Aug-2020
Place: Bangalore

Ph: 88 8010 8010
www.zolostays.com
CIN U74900KA2015PTC080643

ZOLOSTAYS PROPERTY SOLUTIONS PRIVATE LIMITED

Registered Address
No. 1190, 22nd Cross, HSR Layout,
Sector 3, Bangalore, KA-560102, India

~ 6 ~

ABSTRACT

In recent years, machine learning has been widely used in detection and achieved favorable performance. Cricket is one of the most popular team games in the world. Cricket is the second most watched sport in the world after soccer, and enjoys a multi-million dollar industry. With the advent of statistical modeling in sports, predicting the outcome of a game has been established as a fundamental problem. The complex rules prevailing in the game, along with the

various natural parameters affecting the outcome of a cricket match present significant challenges for accurate prediction. The aim is to investigate ensemble learning method based techniques for winning team forecasting by prediction results in best accuracy. data cleaning/preparing and data visualization will be done on the entire given dataset. Our analysis provides a comprehensive guide to sensitivity analysis of model parameters with regard to performance in prediction of winning team or not by accuracy calculation. Additionally, to compare and discuss the performance of various machine learning algorithms with ensemble learning method from the given dataset with GUI based evaluation of classification report, identify the confusion matrix and to categorizing data from priority and the result shows that the effectiveness of the proposed voting classifier technique can be compared with best accuracy with precision, Recall and F1 Score.

~ 7 ~

TABLE OF CONTENT

CHAPTER TITLE	PAGE
1. INTRODUCTION	13
1.1 DOMAIN OVERVIEW	14
1.2 PROJECT DEFINITION	16
1.3 OBJECTIVES	16
1.4 DATASET	17
2. SYSTEM ANALYSIS	18
2.1 EXISTING SYSTEM	19
2.2 PROPOSED SYSTEM	20
2.3 PROJECT REQUIREMENTS	21
3. SYSTEM DESIGN	23
3.1 GENERAL	24
3.2 WORK FLOW DIAGRAM	24
3.3 USE CASE DIAGRAM	25
3.4 CLASS DIAGRAM	25
3.5 ACTIVITY DIAGRAM	27
3.6 SEQUENCE DIAGRAM	28
3.7 ERD	28
4. MODULES	29
4.1 LIST OF MODULES	30
4.2 DATA ANALYSIS OF VISUALIZATION	35
4.3 ALGORITHM	39
4.4 GUI APPLICATION	40

~ 8 ~

5. IMPLEMENTATION	41	5.1ALGORITHM DESCRIPTION	42
5.2PYTHON PACKAGES	42	5.3LOGISTIC REGRESSION	43
5.4DECISION TREE			
5.5RANDOM FOREST	44	5.6KNN	45
6. SCREENSHOTS	47		
7. CONCLUSION	49	7.1 CONCLUSION	
7.2FUTURE WORK			
8. REFERENCE	51		

~ 9 ~

LIST OF FIGURES

FIGURE NO	FIGURE NAME	PAGENO
1.1	PROCESS OF MACHINE LEARNING	15
3.2	WORK FLOW DIAGRAM	24
3.3	CLASS DIAGRAM	25
3.4	ACTIVITY DIAGRAM	27
3.5	SEQUENCE DIAGRAM	28
3.6	ERD	28
4.9	DENSITY PLOT	36

~ 10 ~

LIST OF ABBREVIATIONS

ABBREVIATIONS EXPANSIONS

CT COMPRESSOR TRACKING TIME FP FALSE
POSITIVE
FN FALSE NEGATIVE
TP TRUE POSITIVE
TN TRUE NEGATIVE
KNN K NEAREST NEIGHBOUR

~ 11 ~

CRICKET WINNING TEAM PREDICTION SYSTEM

ABSTRACT

In recent years, machine learning has been widely used in detection and achieved favorable performance. Cricket is one of the most popular team games in the world. Cricket is the second most watched sport in the world after soccer, and enjoys a

multi-million dollar industry. With the advent of statistical modeling in sports, predicting the outcome of a game has been established as a fundamental problem. The complex rules prevailing in the game, along with the various natural parameters affecting the outcome of a cricket match present significant challenges for accurate prediction. The aim is to investigate ensemble learning method based techniques for winning team forecasting by prediction results in best accuracy. The analysis of dataset by voting classifier ensemble learning method to capture several information's like, variable identification, uni-variate analysis, bi-variate and multi-variate analysis, missing value treatments and analyze the data validation, data cleaning/preparing and data visualization will be done on the entire given dataset. Our analysis provides a comprehensive guide to sensitivity analysis of model parameters with regard to performance in prediction of winning team or not by accuracy calculation. Additionally, to compare and discuss the performance of various machine learning algorithms with ensemble learning method from the given dataset with GUI based evaluation of classification report, identify the confusion matrix and to categorizing data from priority and the result shows that the effectiveness of the proposed voting classifier technique can be compared with best accuracy with precision, Recall and F1 Score.

~ 12 ~

CHAPTER 1 INTRODUCTION

1.1Domain overview:

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance. Data scientists use many different kinds of machine learning algorithms to discover patterns in python that lead to actionable insights. At a high level, these different algorithms can be classified into two groups based on the way they “learn” about data to make predictions: supervised and unsupervised learning. Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function from input variables(X) to discrete output variables(y). In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc.



Fig 1.1: Process of Machine learning

Supervised Machine Learning is the majority of practical machine learning uses supervised learning. Supervised learning is where we have input variables (X) and an output variable (y) and use an algorithm to learn the mapping function from the input to the output is $y = f(X)$. The goal is to approximate the mapping function so well that when you have new input data (X) that you can predict the output variables (y) for that data. Techniques of Supervised Machine Learning algorithms include logistic regression, multi-class classification, Decision Trees and support vector machines etc. Supervised learning requires that the data used to train the algorithm is already labeled with correct answers. Supervised learning problems can be further grouped into Classification problems. This problem has as goal the construction of a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for regression and categorical for classification. A regression model attempts to draw some conclusion from observed values. Given one or more inputs a regression model will try to predict the value of one or more outcomes. A classification problem is when the output variable is a category, such as "red" or "blue". The game of cricket is played in three formats - Test Matches, ODIs and T20s. We focus our research on ODIs, the most popular format of the game. To predict the outcome of ODI cricket matches, we propose an approach where we first estimate the batting and bowling potentials of the 22 players playing the match using their career statistics and active participation in recent games. We then use these player potentials to render the relative dominance one team has over the other. Taking two other base features into account, namely, toss decision and the venue of the match, along with the relative team strength, we adopt supervised learning algorithms to predict the winner of the match.

1.2.Problem Definition:

Statistical modeling has been used in sports since decades and has contributed significantly to the success on field. Cricket is one of the most popular sports in the world, second only to soccer. Various natural factors affecting the game, enormous media coverage, and a huge betting market have given strong incentives to model the game from various perspectives. However, the complex rules governing the game, the ability of players and their performances on a given day, and various other natural parameters play an integral role in affecting the final outcome of a cricket match. This presents significant challenges in predicting the accurate results of a game.

1.3.Objectives:

- To predict the results of the cricket match by not only taking factors like ground, venue but also by considering factors like team composition of the each player in the team, the performance of the team in their past few matches, the probability of winning by batting first at a specific venue against a specific team.

1.4.Preparing the Dataset:

The dataset is now supplied to machine learning model on the basis of this data set the model is trained. These datasets are merged to form a common dataset, on which analysis will be done.

Table shows details of the datasets:

Variable	Description
Scorecard	ODI scorecard
Team 1	ODI team A
Team 2	ODI team B

Margin	Winner team innings
--------	---------------------

~ 16 ~

Ground	Place of ground
Match Date	Date of ODI match
Winner	Winner team list
Host_Country	Host country name
Venue_Team1	Venue of ODI team A
Venue_Team2	Venue of ODI team B
Innings_Team1	Innings of ODI team A
Innings_Team2	Innings of ODI team B

~ 17 ~

CHAPTER 2 SYSTEM ANALYSIS

2.1.EXISTING SYSTEM

Sports video analysis has received increasing attention in recent years. Athlete tracking and action recognition are its two major issues that are highly related to each other; however, they are individually considered and processed in existing studies. In this paper, we propose a joint framework for athlete tracking and action recognition in sports videos. In athlete tracking, we propose a scaling and occlusion robust tracker, named SORT, to localize the position of the specific athlete in each frame. It follows the approach of Compressive Tracking (CT), but extends it in two aspects, i.e., scale refinement as well as occlusion recovery. For the former, an objectness method, Edge Box (EB), is adopted to generate proposals, which replace the fixed sampling boxes in CT and better fit the scales of the candidate objects. For the latter, a candidate obstruction based solution is presented, which brings in additional trackers to detect possible obstructions and relocate the target as occlusion ends. Regarding action recognition, we propose a Long-term Recurrent Region-guided Convolutional Network (LRRCN), which recognizes pre-defined actions by modeling discriminative temporal cues of the tracking results. We employ SPP-net to extract the robust feature of the tracked region of each frame. The features of all the frames are then fed into a stack of recurrent sequence models to capture the long-term region-level information. We extensively evaluate the proposed approach on a newly collected sports video benchmark and the off-the-shelf UIUC2 dataset, and the experimental results clearly show its effectiveness.

2.2.PROPOSED SYSTEM

Machine Learning and Data Mining are developing at a rapid pace with several new techniques being developed and old techniques being modified to enhance performance, keeping this in mind our work can be expanded to incorporate new methods of classification for outcome prediction and more features could be added along with the ones currently considered.

Exploratory Data Analysis of Winning team prediction:

Machine learning supervised classification algorithms will be used to give the ODI

dataset and extract patterns, which would help in predicting the likely winning team or not, thereby helping the ICC for making better decisions in the future. Multiple datasets from different sources would be combined to form a generalized dataset, and then different machine learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy. The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion need to be done. The correlation among attributes can be identified using plot diagram in data visualization process. Data preprocessing is the most time consuming phase of a data mining process. Data cleaning of loan data removed several attributes that has no significance about the behavior of a teams.

Advantages:

- Machine Learning is developing at a rapid pace with several new techniques being developed and old techniques being modified to enhance performance, keeping this in mind our work can be expanded to incorporate new methods of classification for outcome prediction.
- The goal of this problem is to predict the status of winning team of test dataset as accurate as possible and time limit can be set for the applicant to check whether won or loss. To provide quick, immediate and easy way to choose the deserving applicants.

2.3.Project Requirements

Requirements are the basic constraints that are required to develop a system. Requirements are collected while designing the system. The following are the requirements that are to be discussed.

1. Functional requirements

2. Non-Functional requirements

3. Environment requirements

A. Hardware requirements

B. software requirements

Functional requirements:

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirements analysis process. It lists requirements of a particular software system. The following details to follow the special libraries like sk-learn, pandas, numpy, matplotlib and seaborn.

Non-Functional Requirements:

Process of functional steps,

1. Problem define
2. Preparing data
3. Evaluating algorithms
4. Improving results
5. Prediction the result

Frontend: GUI

Backend: ML

Environmental Requirements:

1. Software Requirements:

Operating System : Windows

Tool : Anaconda with Jupyter Notebook

2. Hardware requirements:

Processor : i3 / i4

~ 21 ~

Hard disk : minimum 300 GB RAM :

minimum 4 GB

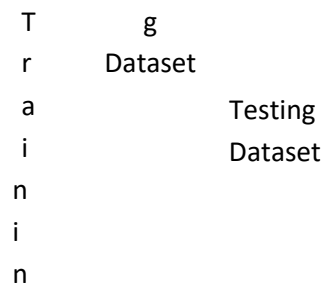
CHAPTER 3 SYSTEM DESIGN

3.1.General

Design is meaningful engineering representation of something that is to be built. Software design is a process design is the perfect way to accurately translate requirements in to a finished software product. Design creates a representation or model, provides detail about software data structure, architecture, interfaces and components that are necessary to implement a system.

3.2.Work flow diagram:





Classification ML Algorithms Best Model by Accuracy

Finding the winning team or not by best accuracy

Fig:3.2. Workflow Diagram

3.3.Use Case Diagram

Fig:3.3.Use case diagram

Use case diagrams are considered for high level requirement analysis of a system. So when the requirements of a system are analyzed the functionalities are captured in use cases. So, it can say that uses cases are nothing but the system functionalities written in an organized manner. Now the second things which are relevant to the use cases are the actors (Public).

3.4.Class Diagram:

fig:3.4.class diagram

~ 25 ~

Class diagram is basically a graphical representation of the static view of the system and represents different aspects of the application. So a collection of class diagrams represent the whole system. The name of the class diagram should be meaningful to describe the aspect of the system. Each element and their relationships should be

identified in advance Responsibility (attributes and methods) of each class should be clearly identified for each class minimum number of properties should be specified and because, unnecessary properties will make the diagram complicated. Use notes whenever required to describe some aspect of the diagram and at the end of the drawing it should be understandable to the developer/coder. Finally, before making the final version, the diagram should be drawn on plain paper and rework as many times as possible to make it correct.

3.5.Activity Diagram:

Fig:3.5. activity diagram

Activity is a particular operation of the system. Activity diagrams are not only used for visualizing dynamic nature of a system but they are also used to construct the executable system by using forward and reverse engineering techniques. The only missing thing in activity diagram is the message part. It does not show any message flow from one activity to another. Activity diagram is some time considered as the flow chart. Although the diagrams looks like a flow chart but it is not. It shows different flow like parallel, branched, concurrent and single.

3.6.Sequence Diagram:

Fig : 3.6.sequence diagram

Sequence diagrams model the flow of logic within your system in a visual manner, enabling you both to document and validate your logic, and are commonly used for both analysis and design purposes. Sequence diagrams are the most popular UML artifact for dynamic modeling, which focuses on identifying the behavior within your system.

3.7.Entity Relationship Diagram (ERD)

Fig:3.7.ERD

4.1.List of Modules:

- Data validation Process
- Exploration data analysis of pre-processing
- Data Visualization Process
- Comparing Algorithm with prediction in the form of best accuracy result □ Ensemble learning method by voting classifier
- GUI Application

Variable validation Process

Validation techniques in machine learning are used to get the error rate of the Machine Learning (ML) model, which can be considered as close to the true error rate of the dataset. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, to work with samples of data that may not be a true representative of the population of given dataset. To finding the missing value, duplicate value and description of data type whether it is float variable or integer. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters.

Fig: 4.1. Given data frame

The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. The validation set is used to evaluate a given model, but this is for frequent evaluation. It as machine learning engineers uses this data to fine tune the model hyper parameters. Data collection, data analysis, and the process of addressing data content, quality, and structure can add up to a time-consuming to-do

list. During the process of data identification, it helps to understand your data and its properties; this knowledge will help you choose which algorithm to use to build your model. For example, time series data can be analyzed by regression algorithms; classification algorithms can be used to analyze discrete data. A number of different data cleaning tasks using Python's Pandas library and specifically, it focus on probably the biggest data cleaning task, missing values and it able to more quickly clean data. It wants to spend less time cleaning data, and more time exploring and modeling.

Fig.4.2: Data-types and missing values of each column

Data Validation/ Cleaning/Preparing Process:

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the uni-variate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

Fig:4.3. Team wise win/ loss lists

Fig:4.4. Team wise - Innings win/ loss lists

Some of these sources are just simple random mistakes. Other times, there can be a deeper reason why data is missing. It's important to understand these different types of missing data from a statistics point of view. The type of missing data will influence how to deal with filling in the missing values and to detect missing values, and do some basic imputation and detailed statistical approach for dealing with missing data. Before, joint into code, it's important to understand the sources of missing data. Here are some typical reasons why data is missing:

~ 31 ~

- User forgot to fill in a field.
- Data was lost while transferring manually from a legacy database.
- There was a programming error.
- Users chose not to fill out a field tied to their beliefs about how the results would be

used or interpreted.

Fig:4.5. Total matches Vs Winning Teams

Important points:

- ❑ Finding minimum range of winning team
- ❑ Finding maximum range of winning team

Variable identification with Uni-variate, Bi-variate and Multi-variate analysis:

- ❑ import libraries for access and functional purpose and read the given dataset
- ❑ General Properties of Analyzing the given dataset
- ❑ Display the given dataset in the form of data frame
- ❑ show columns
- ❑ shape of the data frame
- ❑ To describe the data frame

~ 32 ~

- ❑ Checking data type and information about dataset
- ❑ Checking for duplicate data
- ❑ Checking Missing values of data frame
- ❑ Checking unique values of data frame
- ❑ Checking count values of data frame

- ❑ Rename and drop the given data frame
- ❑ To specify the type of values
- ❑ To create extra columns

Fig:4.6 Min / Max range of team lists

Exploration data analysis of pre-processing:

- ❑ Create a function of home stadium winning teams
- ❑ Create a function of away stadium winning teams
- ❑ Create a function of home stadium loss teams
- ❑ Create a function of away stadium loss teams
- ❑ Encoding process(Pre-Processing technique): Label
 - encoder used ○ Before encoding dataset
 - After encoding dataset

4.2.Exploration data analysis of visualization:

Data visualization is an important skill in applied statistics and machine learning. Statistics does indeed focus on quantitative descriptions and estimations of data. Data visualization provides an important suite of tools for gaining a qualitative

~ 33 ~

understanding. It can be helpful when exploring and getting to know a dataset and can help with identifying patterns, corrupt data, outliers, and much more. With a little domain knowledge, data visualizations can be used to express and demonstrate key relationships in plots and charts that are more visceral and stakeholders than measures of association or significance. Data visualization and exploratory data analysis are whole fields themselves and it will recommend a deeper dive into some the books mentioned at the end.

Fig:4.7. Most winning Teams by percentage

- How to chart time series data with line plots and categorical quantities with bar charts.
- How to summarize data distributions with histograms and box plots. □ How to summarize the relationship between variables with scatter plots.

Data Visualization types:

- Introduction to Matplotlib
- Line Plot
- Bar Chart
- Histogram Plot
- Box and Whisker Plot
- Scatter Plot

There are many excellent plotting libraries in Python and it recommends exploring them in order to create presentable graphics. For quick and dirty plots intended for

~ 34 ~

your own use, it recommends using the matplotlib library. It is the foundation for many other plotting libraries and plotting support in higher-level libraries such as Pandas. The matplotlib provides a context, one in which one or more plots can be drawn before the image is shown or saved to file and context can be accessed via functions on pyplot. A bar chart is generally used to present relative quantities for multiple categories. The x-axis represents the categories and are spaced evenly. The y-axis represents the quantity for each category and is drawn as a bar from the baseline to the appropriate level on the y-axis. A box and whisker plot, or boxplot for short, is

generally used to summarize the distribution of a data sample. The x-axis is used to represent the data sample, where multiple boxplots can be drawn side by side on the x-axis if desired. The boxplot is a graphical technique that displays the distribution of variables. It helps us see the location, skewness, spread, tile length and outlying points. The boxplot is a graphical representation of the Five Number Summary. A heat map is a graphical representation of data where the individual values contained in a matrix are represented as colors. It is a bit like looking a data table from above. It is really useful to display a general view of numerical data, not to extract specific data point.

Fig:4.8. Relationship diagram for given data co-relations

Many machine learning algorithms are sensitive to the range and distribution of attribute values in the input data. Outliers in input data can skew and mislead the

~ 35 ~

training process of machine learning algorithms resulting in longer training times, less accurate models and ultimately poorer results. Even before predictive models are prepared on training data, outliers can result in misleading representations and in turn misleading interpretations of collected data. Outliers can skew the summary distribution of attribute values in descriptive statistics like mean and standard deviation and in plots such as histograms and scatterplots, compressing the body of the data. Finally, outliers can represent examples of data instances that are relevant to the problem such as anomalies in the case of fraud detection and computer security.

The three steps involved in cross-validation are as follows:

1. Reserve some portion of sample data-set.
2. Using the rest data-set train the model.
3. Test the model using the reserve portion of the data-set.

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. To achieving better results from the applied model in Machine Learning method of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format, for example, Random Forest algorithm does not support null values. Therefore, to execute random forest algorithm null values have to be managed from the original raw data set.

Fig:4.9. Density plot for each column

~ 36 ~

False Positives (FP): A person who will pay predicted as defaulter. When actual class is no and predicted class is yes. E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

False Negatives (FN): A person who default predicted as payer. When actual class is yes but predicted class in no. E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

True Positives (TP): A person who will not pay predicted as defaulter. These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN): A person who default predicted as payer. These are the correctly predicted negative values which means that the value of actual class is no

and value of predicted class is also no. E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

4.3.Comparing Algorithm with prediction in the form of best accuracy result

It is important to compare the performance of multiple different machine learning algorithms consistently and it will discover to create a test harness to compare multiple different machine learning algorithms in Python with scikit-learn. It can use this test harness as a template on your own machine learning problems and add more and different algorithms to compare. Each model will have different performance characteristics. Using resampling methods like cross validation, you can get an estimate for how accurate each model may be on unseen data. It needs to be able to use these estimates to choose one or two best models from the suite of models that you have created. When have a new dataset, it is a good idea to visualize the data using different techniques in order to look at the data from different perspectives. The same idea applies to model selection. You should use a number of different ways of looking at the estimated accuracy of your machine learning algorithms in order to choose the one or two to finalize. A way to do this is to use

~ 37 ~

different visualization methods to show the average accuracy, variance and other properties of the distribution of model accuracies.

Prediction result by accuracy:

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. We need the output of the algorithm to be classified variable data. Higher accuracy predicting result is logistic regression model by comparing the best accuracy.

True Positive Rate (TPR) = $TP / (TP + FN)$

False Positive Rate (FPR) = $FP / (FP + TN)$

Accuracy: The Proportion of the total number of predictions that is correct otherwise overall how often the model predicts correctly defaulters and non-defaulters.

Accuracy calculation:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same.

Precision: The proportion of positive predictions that are actually correct. (When the model predicts default: how often is correct?)

$$\text{Precision} = TP / (TP + FP)$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

~ 38 ~

Recall: The proportion of positive observed values correctly predicted. (The proportion of actual defaulters that the model will correctly predict)

$$\text{Recall} = TP / (TP + FN)$$

Recall(Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.

F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall.

General Formula:

$$F\text{- Measure} = 2TP / (2TP + FP + FN)$$

F1-Score Formula:

$$\text{F1 Score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

4.4.GUI Application:

Graphical User Interfaces with Tk

Tk/Tcl has long been an integral part of Python. It provides a robust and platform independent windowing toolkit, that is available to Python programmers using the `tkinter.tix` and the `tkinter.ttk` modules. The `tkinter` package is a thin object-oriented layer on top of Tk/Tk. To use `tkinter`, you don't need to write Tcl code, but you will need to consult the Tk documentation, and occasionally the Tcl documentation. `tkinter` is a set of wrappers that implement the Tk widgets as Python classes. In addition, the internal module `_tkinter` provides a threadsafe mechanism which allows Python and Tcl to interact. Most of the time, `tkinter` is all you really need, but a number of additional modules are available as well. The Tk interface is located

~ 39 ~

in a binary module named `_tkinter`. This module contains the low-level interface to Tk, and should never be used directly by application programmers. It is usually a shared library (or DLL), but might in some cases be statically linked with the Python interpreter.

~ 40 ~

CHAPTER 5 IMPLEMENTATION

~ 41 ~

5.1.Algorithm Explanation:

In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc. In Supervised Learning, algorithms learn from labeled data. After

understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data.

5.2.Used Python Packages:

sklearn:

- In python, sklearn is a machine learning package which include a lot of ML algorithms.

- Here, we are using some of its modules like train_test_split,

DecisionTreeClassifier or Logistic Regression and accuracy_score. NumPy:

- It is a numeric python module which provides fast maths functions for calculations.
- It is used to read data in numpy arrays and for manipulation purpose.

Pandas:

- Used to read and write different files.
- Data manipulation can be done easily with data frames.

Matplotlib:

- Data visualization is a useful way to help with identify the patterns from given dataset.
- Data manipulation can be done easily with data frames.

~ 42 ~

5.3.Logistic Regression:

It is a statistical method for analysing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0

(no, failure, etc.).

5.4. Decision Tree:

It is one of the most powerful and popular algorithm. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. Assumptions of Decision tree:

- At the beginning, we consider the whole training set as the root.
- Attributes are assumed to be categorical for information gain, attributes are assumed to be continuous.
- On the basis of attribute values records are distributed recursively.
- We use statistical methods for ordering attributes as root or internal node.

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two or more branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data. Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covered by the rules are removed. This process is continued on the training

~ 43 ~

set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance. Attributes in the top of the tree have more impact towards in the classification and they are identified using the information gain concept. A decision tree can be easily over-fitted generating too many branches and may reflect anomalies due to noise or outliers.

5.5. Random Forest:

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random

decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

The following are the basic steps involved in performing the random forest algorithm:

- ❑ Pick N random records from the dataset.
- ❑ Build a decision tree based on these N records.
- ❑ Choose the number of trees you want in your algorithm and repeat steps 1 and 2.
- ❑ In case of a regression problem, for a new record, each tree in the forest predicts a value for Y (output). The final value can be calculated by taking the average of all the values predicted by all the trees in forest. Or, in case of a classification problem, each tree in the forest predicts the category to which the

~ 44 ~

new record belongs. Finally, the new record is assigned to the category that wins the majority vote.

5.6.KNN:

k-Nearest Neighbor. The **k**-NN is an instance-based classifier. The underlying idea is that the likelihood that two instances of the instance space belongs to the same category or class increases with the proximity of the instance. Proximity or closeness can be defined with a distance or similarity function. With the data stored into items, we now start building our classifier. For the classifier, we will create a new function, Classify. It will take as input the item we want to classify, the items list and k, the

number of the closest neighbors. If k is greater than the length of the data set, we do not go ahead with the classifying, as we cannot have more closest neighbors than the total amount of items in the data set. (alternatively we could set k as the items length instead of returning an error message). We want to calculate the distance between the item to be classified and all the items in the training set, in the end keeping the k shortest distances. To keep the current closest neighbors we use a list, called neighbors. Each element in the list holds two values, one for the distance from the item to be classified and another for the class the neighbor is in. We will calculate distance via the generalized Euclidean formula (for n dimensions).

Support Vector Machines

A classifier that categorizes the data set by setting an optimal hyper plane between data. I chose this classifier as it is incredibly versatile in the number of different kernelling functions that can be applied and this model can yield a high predictability rate. Support Vector Machines are perhaps one of the most popular and talked about machine learning algorithms.

~ 45 ~

CHAPTER 6 SCREENSHOTS

~ 46 ~

~ 47 ~

7.1.CONCLUSION:

Applying machine learning for analysing cricket sports by considering historical game data, players performance, natural parameters, pre-game conditions and other features is beneficial for multiple stakeholders. In a dynamic format like T20, where the situation in a game changes on every ball, it becomes challenging to predict the match outcome. For predicting the final outcome of a T20 cricket match, we have investigated machine learning technology for the possibility of improving the prediction rate of the results of matches. Machine learning may slightly improve predicting the results based on pre-game conditions but at this stage it cannot be an acceptable solution due to lack of variables in the dataset, which can be considered as one of this research's limitations. In order for machine learning techniques to be productive, more data including live data streaming and statistics of players are needed. Considering the dynamics of the tournament, team players' data and statistics are required. It would be advantageous to predict the final score of the innings by analysing the run rate per over and also checking the probability of winning for each team depending on the actual run rate and the required run rate in the second innings.

7.2.Future Work:

□ To optimize the work to implement in Artificial Intelligence Environment (AIE). □ ICC wants to automate the detecting the winning team from eligibility process (real time) based on the past data.

□ To automate this process by show the prediction result in web application or desktop application.

8. References

- [1] Abdelhamid, N., Ayesh, A., Thabtah, F., 2012. An experimental study of three different rule ranking formulas in associative classification mining, in: Proceedings of the 7th IEEE International Conference for Internet Technology and Secured Transactions (ICITST-2012), pp. 795–800, UK.
- [2] Analytics, C., 2017. Anaconda software distribution. Computer software Vers, 2-2.
- [3] Bhatia, G., 2017. The richest sport in India just keeps getting richer. Retrieved December 25, 2018 from <https://www.cnbc.com/2017/09/27/indian-premier-league-cricket-a-rich-sport-is-getting-a-lot-richer.html>.
- [4] D. Böhning **Multinomial logistic regression algorithm** Ann. Inst. Stat. Math., 44 (1) (1992), pp. 197-200
- [5] L. Breiman **Random forests** Machine Learn., 45 (1) (2001), pp. 5-32
- [6] R. Bunker, F. Thabtah **A machine learning framework for sport result prediction** J. Appl. Comput. Informatic (2017), [10.1016/j.aci.2017.09.005](https://doi.org/10.1016/j.aci.2017.09.005). Elsevier
- [7] Cricinfo, n.d. Retrieved from <http://www.espn.cricinfo.com>.
- [8] R.O. Duda, P.E. Hart, D.G. Stork **Bayesian decision theory** Pattern Classification, 11 (4) (2001), pp. 99-102
- [9] E. Frank, Y. Wang, S. Inglis, G. Holmes, I.H. Witten **Using model trees for classification** Machine Learn., 32 (1) (1998), pp. 63-76
- [10] J.H. Friedman **Stochastic gradient boosting** Comput. Stat. Data Anal., 38 (4) (2002), pp. 367-378

