# Cricket Prediction and Analysis Using Data Science with ML

*Gopi Krishna (PG Scholar)*
*department of Master Of Computer Applications*
*Vasireddy Venkatadri Institute of Technology(VVIT),*
*Namburu, Andhra Pradesh, India*

*Abstract*---**Cricket is one of the most-watched sport now-a-days. Winning in cricket depends on various factors like performances in the recent past matches, player performances, performance against the specific team and the current form of the team and the player. [6] In the past, lots of research has been done which measures the player's performance and predicts the winning percentage.**

**This article briefs about the factors that cricket game depends on and discusses various researches which predicted the winning of a team with the advent of statistical modeling in sports. [1] Cricket is one of the most mainstream group games in the world. With the help of this article, we predicting the outcome of 2019 ICC One Day International (ODI) cricket match using the supervised learning approach from the team composition perspective. [16] Our work proposes that the relative group quality between the contending groups frames a particular component for foreseeing the victor, [2] Modeling the team strength boils down to modeling individual player 's batting and bowling performances.**

***Keywords:- Modeling Players, Modeling Teams, Data mining, Winner prediction, Classification algorithm.***

## I. INTRODUCTION

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.

[2] 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 significance challenge predicting the accurate results of a game.

[1] [3] The cricket game is played in three formats – T20s, ODIs and TEST MATCHEs. We focus research on ODIs, which is the most popular format of the game. To predict outcome of the ODI cricket matches, we play an approach where we first estimate the [13] bowling and batting potentials of the 22 players playing the match using their career statistics and participation in recent games.

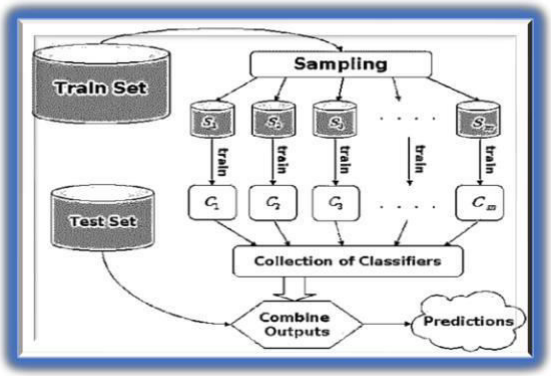


*Fig 1. Architecture*

TRAINING AND TEST DATA SETS:

- In reality, we have a wide range of information like money related information or client information.
- A calculation should make new expectations dependent on new information.
- You can recreate this by parting the dataset in preparing and test information.

- You can simulate this by splitting the dataset in training and test data.


*Fig 2. Prediction Process*

Prediction: Prediction modeling is constantly an enjoyable task. The significant time spent is to comprehend what the business needs and afterwards outline your concern. The subsequent stage is to tailor the answer to the requirements. As we take care of numerous issues, we comprehend that a system can be utilized to fabricate our first cut models. Not just this system gives you quicker outcomes, it likewise causes you to get ready for the following stages dependent on the outcomes.

Load Dataset — Data Understanding-:

*import pandas as pd*

*df = pd.read_excel("bank.xlsx")*

Data Transformation — Data Preparation-: Now we have data in format of pandas. Next, we look at the variable descriptions and the contents of the dataset using df.info() and df.head() respectively. The target variable (‘Yes’/’No’) is converted to (1/0) using the code below.

*df['target'] = df['y'].apply(lambda x: 1 if x == 'yes' else 0)*

Data understanding-: Exploratory statistics helps a modeler understand the data better. A couple of these stats are available in this framework. First, we check the missing values in each column in the dataset by using the below code.

*df.isnull().mean().sort values(ascending=False)*100*

Variable Selection — Data Preparation-: We follow the variable selection process, where the variables are selected based on a voting system. We use different algorithms to select features and then finally each algorithm votes for their selected feature. The last vote check is taken to choose the better feature for modeling.

## II . STRUCTURE PROCESSING

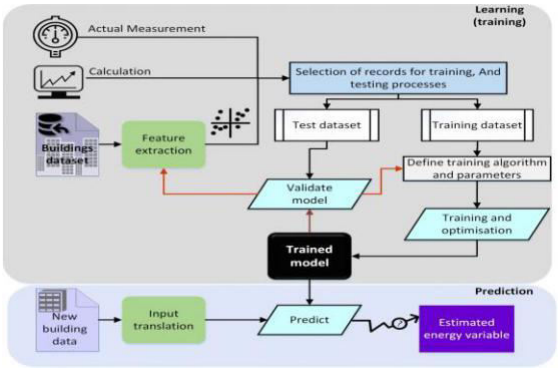


*Fig 3. Design of the System*

**Define project:** Define the task results, the deliverable, extent of the exertion, business goals, distinguish the informational datasets that will be utilized.

**Data Collection:** Data Mining for prescient investigation gets ready information from different hotspots for examination. This gives total perspective on client associations.

**Data Analysis:** Data Analysis is the way that towards reviewing, cleaning and the demonstrating information with the goal of finding valuable data, coming to end result.

**Insights:** Statistical Analysis empowers to approve the suspicions, theory and test them utilizing standard measurable models.

**Modelling:** Predictive demonstrating gives the capacity to consequently make precise prescient models about the future. There are some additionally choices to pick the best arrangement with the multi-modular assessment.

**Deployment:** Predictive model sending gives the choice to convey the systematic outcomes into the ordinary dynamic procedure to get results,

reports and yield via mechanizing the choices dependent on the demonstrating.

**Model-checking:** Models are overseen and observed to audit the model execution to guarantee that it is giving the outcomes anticipated.

## III. ALGORITHM

It is a strategy to break down an informational collection which has a needy variable and at least one autonomous Factors to foresee the result in a paired variable, which means it will have just two results.

The reliant variable is unmitigated in the nature. Subordinate variable is additionally alluded as target variable and the autonomous factors are known as the indicators. Calculated regression is a unique instance of direct relapse where we just foresee the result in a clear -cut variable. It predicts the likelihood of the occasion utilizing the log function.

The dependent variable is categorical in nature. Dependent variable is also referred as the target variable and the independent variables are called the predictors. Logistic Regression is a special case of linear Regression where we only predict the outcome in a categorical variable. It predicts the probability of the event using the log function.

We utilize the Sigmoid capacity/bend to foresee the straight out worth. The edge esteem chooses the outcome(win/lose).

Logistic Regression condition: $y = \beta0 + \beta1X1 + \beta2X2 \ldots\ldots + \beta nXn$

- Y represents the dependent variable that should be anticipated.
- β0 is the Y-block, which is essentially the point on the line which contacts the y-pivot.
- β1 is the incline of the line (the slant can be negative or positive relying upon the connection between the reliant variable and the free factor.)
- X here speaks to the free factor that is utilized to foresee our resultant ward esteem.

Sigmoid capacity: p = 1/1 + e-y

Apply sigmoid capacity on the Logistic Regression condition.

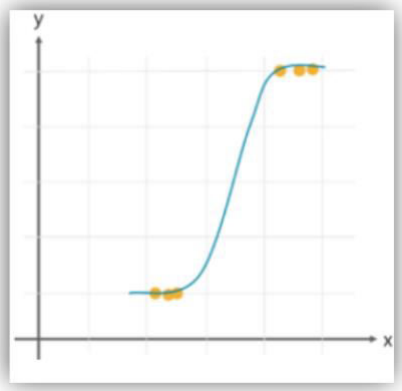


*Fig 4. Graphical representation of sigmoid curve*

## *IV. CONCLUSION*

In this article, we present a logistic regression calculation for anticipating the more precise outcome. we proposed the framework that gives the exact outcome, we utilized information handling strategies to stay away from the vacant information cells. it takes the player subtleties to assess the team execution. we can't appraise the exact aftereffect of running match between the teams since that may change in the latest possible time. with the additional player subtleties we accomplished the great outcomes and progressively precise.

## REFERENCES

[1] Duckworth, Frank C., and Anthony J. Lewis.”Afair method for resetting the target in interrupted one-day cricket matches.” Journal of the Operational Research Society 49.3 (1998): 220-227.

[2] Beaudoin, David, and Tim B. Swartz. ”The best batsmen and bowlers in one-day cricket.” South African Statistical Journal 37.2(2003):203.

[3] Lewis, A. J. ”Towards fairer measures of player performance in one-day cricket.” Journal of the Operational Research Society 56.7(2005):804-815.

[4] Swartz, Tim B., Paramjit S. Gill, and David Beaudoin. ”Optimal batting orders in one-day cricket.” Computers and research 33.7 (2006):1939-1950.

[5] Norman, John M., and Stephen R. Clarke. ”Optimal batting orders in cricket.” Journal of the Operational Research Society 61.6 (2010):980-986.

[6] Kimber, Alan. ”A graphical display for comparing bowlers in cricket.” Teaching Statistics. 15.3 (1993): 84-86.

[7] Barr, G. D. I., and B. S. Kantor. ”A criterion for comparing and selecting batsmen in limited overs cricket.” Journal of the Operational Research Society 55.12 (2004): 1266-1274.

[8] Van Staden, Paul Jacobus. ”Comparison of cricketers bowling and batting performances using graphical displays.” (20 10 Madan Gopal Jhanwar and Vikram Pudi

[9] Lemmer, Hermans ”THE ALLOCATION OF WEIGHTS within the CALCULATION OF BATTING AND BOWLING PERFORMANCE MEASURES.” South African Journal for Research in Sport, education and Recreation(SAJR SPER) 29.2 (2007).

[10] Kaluarachchi, Amal, and S. Varde Aparna. ”CricAI: A classification based tool to predict the **end in** ODI cricket.” 2010 Fifth International Conference on Information and Automation for Sustainability. IEEE, 2010.

[11] Sankaranarayanan, Vignesh Veppur, Junaed Sattar, and Laks VS Lakshmanan. ”Auto-play: a knowledge Mining Approach to ODI Cricket Simulation and Prediction.” SDM. 2014.

[12] Khan, Mehvish, and Riddhi Shah. ”Role of External Factors on Outcome of a 1 Day International Cricket (ODI) Match and Predictive Analaysis.”

[13] HOWSTAT Cricinfo, http://howstat.com/cricket/home.asp

[14] Barr, G. D. I., and R. van den Honert. ”Evaluating batsman’s scores in test cricket.” South African Statistical Journal 32.2 (1998): 169-183.

[15] Croucher, J. S. ”Player ratings in one-day cricket.” Proceedings of the fifth Australian conference on mathematics and computers in sport. Sydney, NSW: Sydney University of Technology, 2000.

[16] Lemmer, Hermanus H. ”The combined bowling rate as a measure of bowling performance in cricket.” South African Journal for Research in Sport, education and Recreation 24.2 (2002): 37-44.

[17] Barr, G. D. I., C. G. Holdsworth, and B. S. Kantor. ”Evaluating performances at the 2007 cricket **World Cup** .” South African Statistical Journal 42.2 (2008): 125.

[18] Pedregosa, Fabian, et al. ”Scikit-learn: Machine learning in Python.” Journal of Machine Learning Research 12.Oct (2011): 2825-2830.