**CRICKET BATSMAN STRIKE RATE PREDICTION**

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

**BACHELOR OF TECHNOLOGY**

in

**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

by

V.SHIVA CHAITHANYA 2203A52060

Under the guidance of

**Mr. D. RAMESH**

Assistant Professor, Department of CSE.

**

**Department of Computer Science and Artificial Intelligence**

**

**Department of Computer Science and Artificial Intelligence**

**CERTIFICATE**

This is to certify that project entitled **“CRICKET BASTMAN STRIKE RATE PREDICTION**" is the bonafied work carried out by **A.KRANTHI,V.RAKESH,B.TEJA** as a Course Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING** during the academic year 2022-2023 under our guidance and Supervision.

**Mr. D. RAMESH Dr. M. Sheshikala**

Asst. Professor, Assoc. Prof .& HOD (CSE)

S R University, S R University,

Ananthasagar ,Warangal Ananthasagar ,Warangal

**ACKNOWLEDGEMENT**

We express our thanks to Course co-coordinator **Mr.** **S.Naresh Kumar, Asst. Prof.** for guiding us from the beginning through the end of the Course Project. We express our gratitude toHead of the department CS&AI, **Dr. M.Sheshikala, Associate Professor** for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, **Dr C. V. Guru Rao**, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

**ABSTRACT**

Cricket Batsman Strike Rate Prediction using Machine Learning Technique in Big data analytics has started to play an important role in the healthcare practices and research. heart attack prediction will be found primarily on real-time processing, distributed and real-time classification and distribution, storage so; databases can be easily modified by the doctors. If you know all the attributes related to our health we can check easily how much chance to the Heart attack risk, using the system applications. It was recently used to train classification models. After that using extract the features that is condition to be find to be classified by Decision Tree (DT).Compared to existing; algorithms provides better performance. After classification, performance criteria including accuracy, precision, F-measure is to be calculated. If you are concern about the heart attack risks, you might be referred to a heart specialist. Some attributes are Heart Attack risk factors including which is the High blood pressure, high cholesterol and diabetes, increases your risk even more. Hence we are also checking your symptoms of heart attack and take about prevention.

**Table of Contents**

**Chapter No. Title Page No.**

1. Introduction
   1. Overview 1
   2. Problem Statement 3
   3. Existing system 7
   4. Proposed system 9
   5. Objectives 12
   6. Architecture 14
2. Literature survey
   * 1. Document the survey done by you 15
3. Data pre-processing

3.1. Dataset description 18

3.2. Data cleaning 19

3.3. Data augmentation 20

3.4. Data Visualization 21

1. Methodology
   1. Procedure to solve the given problem 22
   2. Overview Technology 23

4.3. Software description 24

1. Results and discussion 25
2. Conclusion and future scope ………
3. References

**INTRODUCTION**

* 1. **OVERVIEW**

Cricket is being played in many countries all around the world. There are lots of domestic and international tournaments being held in many countries which play cricket. Cricket is a game played between two team comprising of 11 players in each team. The result is either a win, loss or tie. However, sometime due to bad weather conditions the game is also washed out as cricket is a game which cannot be played in a rain. Moreover, the game is also extremely unpredictable because at every stage of the game the momentum shifts to one of the team between the two. A lot of times the results gets decided on the last ball of the match where the game gets really close. Considering all these unpredictable scenarios of this unpredictable game, there is a huge interest among the spectators to do some predictions either at the start of the game or during the game.

**1.2. PROBLEM STATEMENT**

In this to design a system that can be provide the score and winning prediction in cricket match, the system can analyze multiple parameters like winning toss, batting side, DL approach, home ground advantage, player wise performance etc. while declaring a time for particular championship it is very important to select the best team so that the chances of the team winning the championship become easy. This problem had to be solved to generate the best players from both the team for the best battle. To solve this Problem we have collected historical data of all some team like (India, Australia, New Zealand, South Africa etc.), and using prediction algorithm like Naïve Bayes algorithm we are predicting the best starting players for both the team can be used in fantasy league for winning the maximum points.

**1.3. EXISTING SYSTEM**

Cricket winning can be predicted like all other games. We need to find the best attributes or factors that influences the match outcome. The result of a cricket match depends upon more of in-game and more of pre-game attributes. Pre-game attributes like pitch, Team strength, weather, venue etc. and in-game attributes like run rate, total run, strike rate, wickets in hand etc. influences a match result predominantly. Below are the attributesthat decides outcome ofthe cricket match.

In this system, the input details are obtained

from the patient. Then from the user inputs,

using ML techniques heart disease is analyzed.

Now, the obtained results are compared with the

results of existing models within the same

domain and found to be improved. The data of

heart disease patients collected from the UCI

laboratory is used to discover patterns with NN,

DT, Support Vector machines SVM, and Naive

Bayes. The results are compared for

performance and accuracy with these

algorithms. The proposed hybrid method returns

results of 87% for F-measure, competing with

the other existing methods.

In this system, the input details are obtained

from the patient. Then from the user inputs,

using ML techniques heart disease is analyzed.

Now, the obtained results are compared with the

results of existing models within the same

domain and found to be improved. The data of

heart disease patients collected from the UCI

laboratory is used to discover patterns with NN,

DT, Support Vector machines SVM, and Naive

Bayes. The results are compared for

performance and accuracy with these

algorithms. The proposed hybrid method returns

results of 87% for F-measure, competing with

the other existing methods.

In this system, the input details are obtained

from the patient. Then from the user inputs,

using ML techniques heart disease is analyzed.

Now, the obtained results are compared with the

results of existing models within the same

domain and found to be improved. The data of

heart disease patients collected from the UCI

laboratory is used to discover patterns with NN,

DT, Support Vector machines SVM, and Naive

Bayes. The results are compared for

performance and accuracy with these

algorithms. The proposed hybrid method returns

results of 87% for F-measure, competing with

the other existing methods.

In this system, the input details are obtained

from the patient. Then from the user inputs,

using ML techniques heart disease is analyzed.

Now, the obtained results are compared with the

results of existing models within the same

domain and found to be improved. The data of

heart disease patients collected from the UCI

laboratory is used to discover patterns with NN,

DT, Support Vector machines SVM, and Naive

Bayes. The results are compared for

performance and accuracy with these

algorithms. The proposed hybrid method returns

results of 87% for F-measure, competing with

the other existing methods.

In this system, the input details are obtained

from the patient. Then from the user inputs,

using ML techniques heart disease is analyzed.

Now, the obtained results are compared with the

results of existing models within the same

domain and found to be improved. The data of

heart disease patients collected from the UCI

laboratory is used to discover patterns with NN,

DT, Support Vector machines SVM, and Naive

Bayes. The results are compared for

performance and accuracy with these

algorithms. The proposed hybrid method returns

results of 87% for F-measure, competing with

the other existing methods.

**1.4. PROPOSED SYSTEM**

There has been a lot of related study to this problem in various different sports. The paper I have used for references are all related work that had been done on this problem. The paper by Trawinski described the prediction of results using a fuzzy classification system. This paper was predicting the results for basketball games. I had used the attribute selection technique mentioned in this paper for my project. The attribute selection technique proposed in this paper was done using WEKA so it was good reference point for me too. The wrapper method algorithm and the ranker method algorithm implemented in this paper was also used in my project.

begin



Diagnosis of heart diseases is a significant and boring task and also an important duty in medical science, which

requires extreme attention. However there is some tools for data extraction and analysis. Also existence of huge

set of medical data leads to correct diagnosis of disease. Using medical data including age, sex, blood pressure,

and blood sugar, it is possible to increase the possibility of heart diseases prediction

Diagnosis of heart diseases is a significant and boring task and also an important duty in medical science, which

requires extreme attention. However there is some tools for data extraction and analysis. Also existence of huge

set of medical data leads to correct diagnosis of disease. Using medical data including age, sex, blood pressure,

and blood sugar, it is possible to increase the possibility of heart diseases prediction

Diagnosis of heart diseases is a significant and boring task and also an important duty in medical science, which

requires extreme attention. However there is some tools for data extraction and analysis. Also existence of huge

set of medical data leads to correct diagnosis of disease. Using medical data including age, sex, blood pressure,

and blood sugar, it is possible to increase the possibility of heart diseases prediction



Data collection

classification



Testing data

Training data



Result

End

Test the model

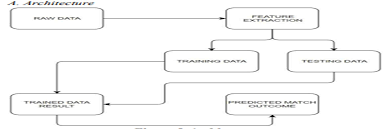
Classification techniques

**1.5. OBJECTIVES**

Cricket is played by two teams of 11, with one side taking a turn to bat a ball and score runs, while the other team will bowl and field the ball to restrict the opposition from scoring. The main objective in cricket is to score as many runs as possible against the opponent. Before the match begins, the captain of both teams will toss a coin, with the winner of the toss being able to decide which team bats and fields first.Each cricket match consists of periods known as innings, and the number of innings that each team has will be determined before the match, usually one or two. During an inning, one team bats the ball while the other attempts to field. Both teams take turns alternating between batting and fielding.

**1.6.ARCHITECTURE**

The architecture of the proposed system is as displayed in the figure below. The major components of the architecture are as follows:, raw data,feature extraction,training data,testing data,training data result,predicted match outcome.



**2.LITERATURE SURVEY**

**2.1.1. Document the survey done by you**

An extensive online search produced very few articles related to players’ performance prediction in the game of cricket. A very small number of researchers have studied the performance of cricket players. Muthuswamy and Lam.

[1] predicted the performance of Indian bowlers against seven international teams against which the Indian cricket team plays most frequently. They used back International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.8, No.2, March 2018 20 propagation network and radial basis network function to predict how many runs a bowler is likely to concede and how many wickets a bowler is likely to take in a given ODI match. Wikramasinghe

[2] predicted the performance of batsmen in a test series using a hierarchical linear model. Barr and Kantor

[3] defined a criterion for comparing and selecting batsmen in limited overs cricket. They defined a new measure P(out) i.e. probability of getting out and used a twodimensional graphical representation with Strike Rate on one axis and P(out) on another. Then they define a selection criterion based on P(out), strike rate and batting average of the batsmen. Iyer and Sharda

[4] used neural networks to predict the performance of players where they classify batsmen and bowlers separately in three categories – performer, moderate and failure. Based on the number of times a player has received different ratings, they recommend if the player should be included in the team to play World Cup 2007. Jhanwar and Paudi

[5] predict the outcome of a cricket match by comparing the strengths of the two teams. For this, they measured the performances of individual players of each team.

**3.DATA PRE-PROCESSING**

**3.1.1 DATASET DESCRIPTION**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sno** | | **Attributes** | | **Description** | |
| 1.  2.  3.  4.  5.  6.  7.  8.  9. | | Team  Player  Filter  Matches  Runs  Highest    Average  Strike rate  Output | | Name of the team  In which team the player is playing  On which team the player have scored the runs and played matches  How many matches the player is played  Runs that the player have scored  Among all the matches played from that bastman and scored highest runs  All the matches played by the batsman can take average  The number of the batter divided by the number of balls faced  Target variable | |
|  | |  | |  | |

**3.2 DATA CLEANING**

The data obtained from the many websites was already cleaned. So, I did not have to do any sort of cleaning on the data. , I had to tackle the missing values data and the data for the matches which were washed outdue to rain. Those matches data were present in the .json file. after combiningthe two data into a single data set, I had to manually fill in the data about the 7 missing teams from the http://espncricinfo.com/ website. Moreover there were 10 matches which were washed out so those instance were filled with null values and they were discarded from the data set. So, the final dataset had 5219 instances with 21 attributes.

**3.4 DATA VISUALISATION**

An important factor of data visualization project is to visualize the data in a visually appealing format, for the users to draw insights from the data. Graphically represented data are easy to be interpreted. The graphs contain all the details of the features that were extracted from the huge datasets. We have done this using amCharts, A Python library for data visualization. We have further used different packages to get the exact analysis and visualization for teams and their players.

The processing of these data for each individual happens in the backend program which is written with the help of flask. Pandas, an open source for data analysis is used as the data processing tool and to provide input output for both csv files.

**DATASET**

****

**4. METHODOLOGY**

**4.1 PROCEDURE TO SOLVE THE GIVEN PROBLEM**

In this project of strike rate analysis and prediction we use three approaches:

1.Ridge regression

2.Lasso regression

3.Linear regression

**1.Ridge Regression**

Ridge regression is a model tuning method that is used to analyse any data that suffers from multi collinearity. This method performs L2 regularization. When the issue of multi collinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values. Ridge regression is a technique which is used for analyzing multiple regression where the data suffers from multi collinearity.

The problem which arises due to multi collinearity is that the basic linear regression model (least square estimates) becomes unbiased and the variance becomes so large that the predicted values are far from the true value. The advantage of using the ridge regression is to avoid over fitting. It works in the same way as the linear regression but it just adds an extra term (α) which helps in the reduction of overfitting. The goal of any machine learning model is to generalize the pattern which it needs to be predicted, i.e. The model should work best on both training as well as test data. Over fitting occurs when the trained model performs well on the training data and performs poorly on the testing dataset. Fig 4.2: Ridge Regression Formula and Examples of Ridge Regression: The ridge regression formula is of the form: Examples where ridge regression may be used

• Ridge regression can be used for the analysis of prostate-specific antigen.

• It method used for the analysis of multi collinearity in multiple regression data

**2.Lasso Regression**

Lasso regression is a type of linear regression that uses shrinkage. The word “LASSO” stands for Least Absolute Shrinkage and Selection Operator. It is a statistical formula for the regularization of data models and feature selection. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of muti collinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination. Fig 4.3: Lasso Regression Formula and Examples of LASSO Regression: The formula which is used to calculate Lasso Regression is Where λ is the lasso penalty factor. Examples where multiple linear regression may be used

• For analyzing the prostate-specific antigen and the clinical measures among the patients who were about to have their prostates removed

• For predicting the effect on sales and spending a certain amount of money on advertising

**3. Linear regression**

Linear regression also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.Thegoalofmultiplelinearregressionistomodelthelinearrelationshipbetweenthe explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

A multiple linear regression analysis is carried out to predict the values of a dependent variable, *Y*, given a set of p explanatory variables (x1,x2,….,x*p)*. In these notes, the necessary theory for multiple linear regression is presented and examples of regression analysis with censusdataaregiventoillustratethistheory.Thiscourseonmultiplelinearregressionanalysis is therefore intended to give a practical outline to the technique. Complicated or tedious algebrawillbeavoidedwherepossible,andreferenceswillbegiventomoretheoreticaltexts on this technique. Important issues that arise when carrying out a multiple linear regression analysis are discussed in detail including model building, the underlying assumptions, and interpretation of results. However, before we consider multiple linear regression analysis we begin with a brief review of simple linear regression.

**Formula and Calculation of Linear Regression**

*yi*=*β*0+*β*1*xi*1+*β*2*xi*2+...+*βxi*+*ϵ*

where,for*i*=*n*observations:

*yi*=dependentvariable

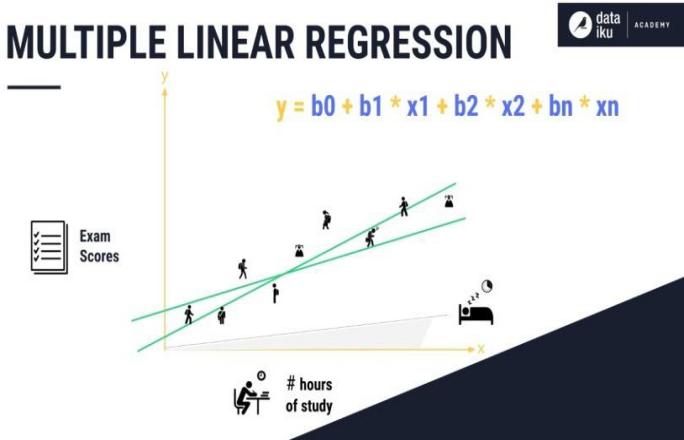
*xi*=explanatoryvariables

*β*0=y-intercept(constantterm)

*βp*=slopecoefficientsforeachexplanatoryvariable

*ϵ*=the model’s error term (also known as the residuals)Exampleswheremultiplelinearregressionmaybeusedinclude:

* Tryingtopredictanindividual’sincomegivenseveralsocio-economiccharacteristics.
* Trying to predict the overall examination performance of pupils in ‘A’ levels, given the values of a set of exam scores at age 16.
* Tryingtoestimatesystolicordiastolicbloodpressure,givenavarietyofsocio-economic and behavioural characteristics (occupation, drinking smoking, age etc).



**4.2.OVERVIEW TECHNOLOGY**

In our program we used python 3 programming language. Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development, as well as for use as a scripting or glue language to connect existing components together. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. The Python interpreter and the extensive standard library are available in source or binary form without charge for all major platforms, and can be freely distributed.

**4.3.SOFTWARE DESCRIPTION**



**Software requirements:**

**Operating system:** Windows

**Platform**: google co-lab

**Programing language:** python

**5. RESULTS**



**DESCRIPTION:**

import pandas as pd

dd=pd.read\_csv('/content/all\_batsmen\_data[1].csv’) print(dd)

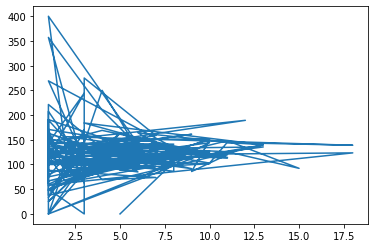
from matplotlib import pyplot as pp

x1=dd['Matches']

y=dd['Strike Rate']

a=pp.plot(x1,y)

print(a)

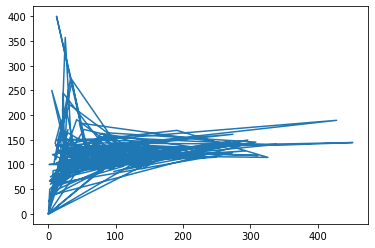


x2=dd['Runs']

y=dd['Strike Rate']

a=pp.plot(x2,y)

print(a)

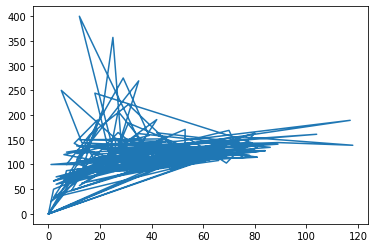


x3=dd['Highest']

y=dd['Strike Rate']

a=pp.plot(x3,y)

print(a)

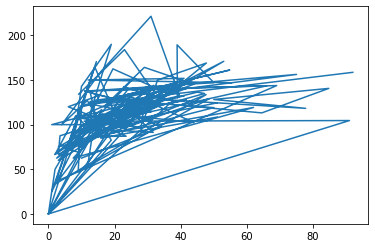


x4=dd['Average']

y=dd['Strike Rate']

a=pp.plot(x4,y)

print(a)



**CODE:**

import numpy as np

from matplotlib import pyplot as plt

e=[]

import numpy as np

import pandas as pd

import numpy as np

from sklearn import preprocessing

df = pd.read\_csv(“/content/all\_batsmen\_data[1].csv”)

print(df.isnull())

df=df.fillna(0)

y=df.iloc[:,7:8]

x=df.iloc[:,4:7]

print(x,'\n',y)

from sklearn.model\_selection import train\_test\_split

X\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.15,random\_state=2)

X\_train,x\_test,y\_train,y\_test=train\_test\_split(x,y,test\_size=0.10,random\_state=99)

print(X\_train.shape)

print(y\_train.shape)

print(x\_test.shape)

print(y\_test.shape)

import numpy as np

from sklearn.metrics import mean\_squared\_error,r2\_score

from sklearn.linear\_model import Lasso

model\_lasso= Lasso(alpha=0.01)

model\_lasso.fit(x,y)

pred\_train\_lasso= model\_lasso.predict(x\_test)

print(mean\_squared\_error(y\_test,pred\_train\_lasso))

from sklearn.linear\_model import Ridge

rr= Ridge(alpha=0.01)

rr.fit(X\_train, y\_train)

pred\_train\_rr= rr.predict(X\_train)

print(np.sqrt(mean\_squared\_error(y\_train,pred\_train\_rr)))

print(r2\_score(y\_train, pred\_train\_rr))

from sklearn.linear\_model import LinearRegression

model = LinearRegression()

model.fit(x, y)

model = LinearRegression().fit(x, y)

r\_sq = model.score(x, y)

print("coefficient of determination:", r\_sq)

print("intercept: ",model.intercept\_)

print("coefficients:",model.coef\_)

print(X\_train)

print(y\_train)

from sklearn import linear\_model

from sklearn.metrics import mean\_squared\_error,mean\_absolute\_error

reg\_all=linear\_model.LinearRegression()

reg\_all.fit(X\_train,y\_train)

y\_pred=reg\_all.predict(x\_test)

#Rsquare=reg\_all.score(y\_pred,y\_test)

#print("Rsquare: %f" %(Rsquare))

#print(y\_test)

print("Intercept: ", reg\_all.intercept\_)

mse=mean\_squared\_error(y\_test,y\_pred)

#print("mse: %f" %(mse))

print(mse)

mae = mean\_absolute\_error(y\_test,y\_pred)

#print("mae: %f" %(mae))

print(mae)

rmse=np.sqrt(mean\_squared\_error(y\_test,y\_pred))

#print("rmse: %f" %(rmse))

print(rmse)

**RESULT:**

[200 rows x 8 columns]

Team Player Filter Matches Runs Highest Average Strike Rate

0 1 2 10.0 8 70 30 14.00 127.27

1 1 2 41.0 2 75 74 75.00 156.25

2 1 2 61.0 18 339 118 21.18 138.93

3 1 2 62.0 11 452 89 41.09 144.40

4 1 2 NaN 13 290 64 24.16 140.09

.. ... ... ... ... ... ... ... ...

195 15 153 41.0 5 15 12 15.00 100.00

196 15 154 63.0 8 50 20 8.33 86.20

197 15 155 63.0 8 29 11 7.25 90.62

198 15 156 63.0 8 151 58 21.57 104.86

199 15 157 63.0 5 0 0 0.00 0.00

Matches Runs Highest Average

0 8 70 30 14.00

1 2 75 74 75.00

2 18 339 118 21.18

3 11 452 89 41.09

4 13 290 64 24.16

.. ... ... ... ...

195 5 15 12 15.00

196 8 50 20 8.33

197 8 29 11 7.25

198 8 151 58 21.57

199 5 0 0 0.00

[200 rows x 4 columns]

Strike Rate

0 127.27

1 156.25

2 138.93

3 144.40

4 140.09

.. ...

195 100.00

196 86.20

197 90.62

198 104.86

199 0.00

[200 rows x 1 columns]

(180, 4)

(180, 1)

(20, 4)

(20, 1)

546.598292138884

47.879532597624554

0.09809244000914952

coefficient of determination: 0.11148724706322599

intercept: [103.44459817]

coefficients: [[-1.83394936 -0.02043467 0.88167642 -0.23215956]]

Matches Runs Highest Average

44 6 22 16 11.00

46 1 25 25 0.00

103 8 137 70 22.83

53 3 51 45 25.50

159 1 24 24 24.00

.. ... ... ... ...

68 10 130 43 18.57

168 1 53 53 53.00

185 5 70 26 14.00

35 7 166 46 55.33

129 8 169 45 24.14

[180 rows x 4 columns]

Strike Rate

44 137.50

46 357.14

103 112.29

53 104.08

159 114.28

.. ...

68 123.80

168 170.96

185 106.06

35 146.90

129 121.58

[180 rows x 1 columns]

Intercept: [106.74512506]

582.7336737439771

15.745189750450152

24.139877252048674

**6. CONCLUSION AND FUTURE SCOPE**

The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system.

The proposed working model can also help in reducing treatment costs by providing Initial

diagnostics in time. The model can also serve the purpose of training tool for medical students

and will be a soft diagnostic tool available for physician and cardiologist. General physicians can

utilize this tool for initial diagnosis of cardio-patients. There are many possible improvements

that could be explored to improve the scalability and accuracy of this prediction system. As we

have developed a generalized system, in future we can use this system for the analysis of

different data sets. The performance of the health’s diagnosis can be improved significantly by

handling numerous class labels in the prediction process, and it can be another positive direction

of research. In DM warehouse, generally, the dimensionality of the heart database is high, so

identification and selection of significant attributes for better diagnosis of heart disease are very

challenging tasks for future research

The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system.

The proposed working model can also help in reducing treatment costs by providing Initial

diagnostics in time. The model can also serve the purpose of training tool for medical students

and will be a soft diagnostic tool available for physician and cardiologist. General physicians can

utilize this tool for initial diagnosis of cardio-patients. There are many possible improvements

that could be explored to improve the scalability and accuracy of this prediction system. As we

have developed a generalized system, in future we can use this system for the analysis of

different data sets. The performance of the health’s diagnosis can be improved significantly by

handling numerous class labels in the prediction process, and it can be another positive direction

of research. In DM warehouse, generally, the dimensionality of the heart database is high, so

identification and selection of significant attributes for better diagnosis of heart disease are very

challenging tasks for future research

Selection of the right players for each match plays a significant role in a team’s victory. An accurate prediction of how many runs a batsman is likely to score and how many wickets a bowler is likely to take in a match will help the team management select best players for each match. In this paper, we modeled batting and bowling datasets based on players’ stats and characteristics. Some other features that affect players’ performance such as weather or the nature of the wicket could not be included in this study due to unavailability of data. Four multiclass classification algorithms were used and compared. Random Forest turned out to be the most accurate classifier for both the datasets with an accuracy of 90.74% for predicting runs scored by a batsman and 92.25% for predicting wickets taken by a bowler. Results of SVM were surprising as it achieved an accuracy of just 51.45% for predicting runs and 68.78% for predicting wickets. Similar studies can be carried out for other formats of the game i.e. test cricket and T20 matches.

**7.REFERENCES**

[1] A. L. Samuel, “Some studies in machine learning using the game of checkers. iirecent progress,” in Computer Games I, pp. 366–400, Springer, 1988.

[2] A. Bandulasiri, “Predicting the winner in one day international cricket,” Journal of Mathematical Sciences & Mathematics Education, vol. 3, no. 1, pp. 6–17, 2008.

[3] Indian Premier League Official Website

[4] P. Langley, W. Iba, K. Thompson, et al., “An analysis of bayesian classifiers,” in Aaai, vol. 90, pp. 223–228, 1992.

[5] S. Kampakis and W. Thomas, “Using machine learning to predict the outcome of English county twenty over cricket matches,” arXiv preprint arXiv:1511.05837, 2015.

[6] L. Passfield and J. G. Hopker, “A mine of information: can sports analytics provide wisdom from your data?,” International journal of sports physiology and performance, vol. 12, no. 7, pp. 851–855, 2017.

[7] T. H. Davenport, “What businesses can learn from sports analytics,” MIT Sloan Man-agement Review, vol. 55, no. 4, p. 10, 2014.