DATA MINING AND WAREHOUSING MINI-PROJECT REPORT ON

“CLASSIFYING RESULTS OF GAME”

SUBMITTED BY

Sakshi Aher = A-26
Gayatri Kasav = A-34
Shital Bhopale=A-29

Dhanshri Surase=B -61

Under the guidance of

Mr.Vishal Patil

DEPARTMENT OF COMPUTER

ENGINEERING Academic Year 2021-22

Contents

1 Problem Statement

2 Abstract

3 INTRODUCTION

4 OBJECTIVE

5 Test Cases6 Result

7 Conclusion

8 References

9 List of Figures

10 List of Tables

1 Problem Statement

Consider a labeled dataset belonging to an application domain. Apply suitable data preprocessing steps such as handling of null values, data reduction, discretization. For prediction of class labels of given data instances, build classifier models using different techniques (minimum 3), analyze the confusion matrix and compare these models. Also apply cross validation while preparing the training and testing datasets.

2.Abstract

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels. For example, we can build a classification model to categorize bank loan applications as either safe or risky. Such analysis can help provide us with a better understanding of the data at large. In this project we use multiple classification models to analyze the outcome of hockey games played between various teams. Use suitable data preprocessing steps.We then compare performance of classification models to find which one is the best.

INTRODUCTION

We have been provided with the data regarding various aspects of the home team, the opposition team and their supporters for a number of hockey games. The Data fields are

1. Id − Unique id given to each game.

2. game seq Sequence of the game in the history of FIH (International Hockey Federation).

3. season end − Year in which the corresponding season ended.

4. date − Date on which the game was played.

5. season game seq − Sequence of the game in the corresponding season.

6. playoff − Whether the game is a playoffs game.

7. team id − Unique id for the home team.

8. Elo − Elo rating for the home team before the game.

9. opp team id − Unique id for the opposition team.

10. opp Elo − Elo rating for the opposition team before the game. 11. win equivalent − Equivalent number of wins for the home team in a season. 12. bet ratio − Fraction of bets placed on the home team. 13. home crowd − Number of supporters for the home team. 14. opp crowd − Number of supporters for the opposition team. 15. total crowd − Total number of attendees for the game. 16. game result − Win or loss for the home team (Win - 1, Loss - 0).

The train set contains 45000 records while the test set contains 13107 records. We drop the date column from our analysis. The null entries are as follows

Attribute	Null Count
Elo	9197
opp Elo	7006
win equivalent	12263

Table 1: Null Counts

We fill the null Elo and opp elo entries with the mean value of Elo and opp Elo ‗ attribute respectively i.e. 1501.184 1501.837 The boxplots of some attributes are as follows:

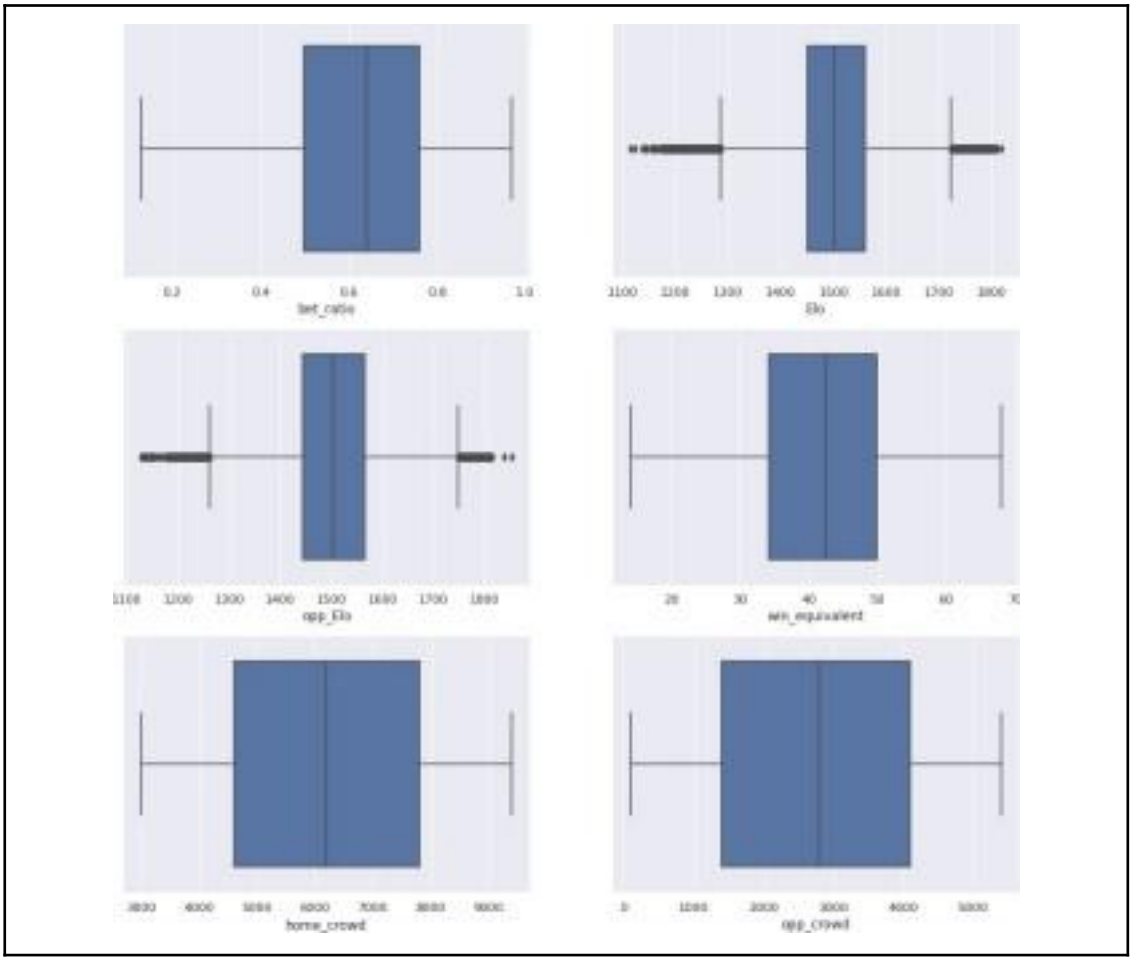

Figure 1: Boxplots

We drop rows with bet ratio below 0.13 and above 0.97. Similarly we drop rows with opp Elo below 1200 and above 1750, Elo below 1175 and above 1780 and win equivalent below 14 and above 68.

We extract the id from string team id and opp id fields. We drop columns Id, game seq,team id and opp id. We convert the playoffs to categorical data.

We have trained using two models: Logistic Regression, KNN classifier , XGBoost Classifier and RandomForest Classifier. We find that XGBoost Classifier performs better.We extract the id from string team id and opp id fields. We drop columns Id,game seq,team id and opp id. We convert the playoffs to categorical data.

2 OBJECTIVE

- To understand data preprocessing
- To perform classification on dataset and predict labels for test dataset

3 Test Cases

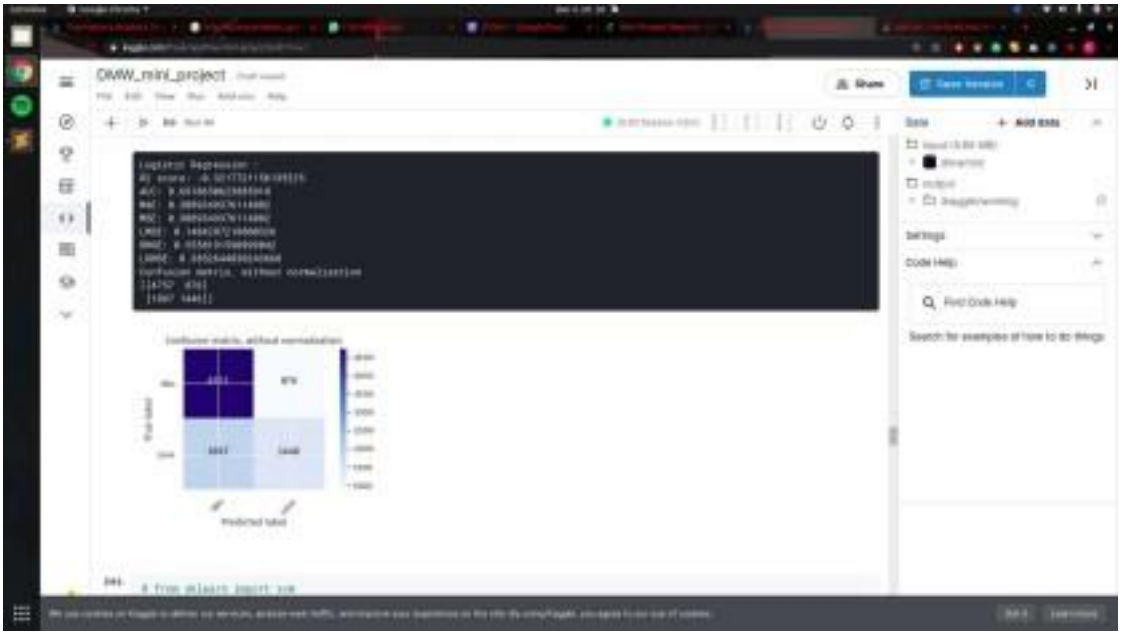

Figure 2: Output for logistic regression

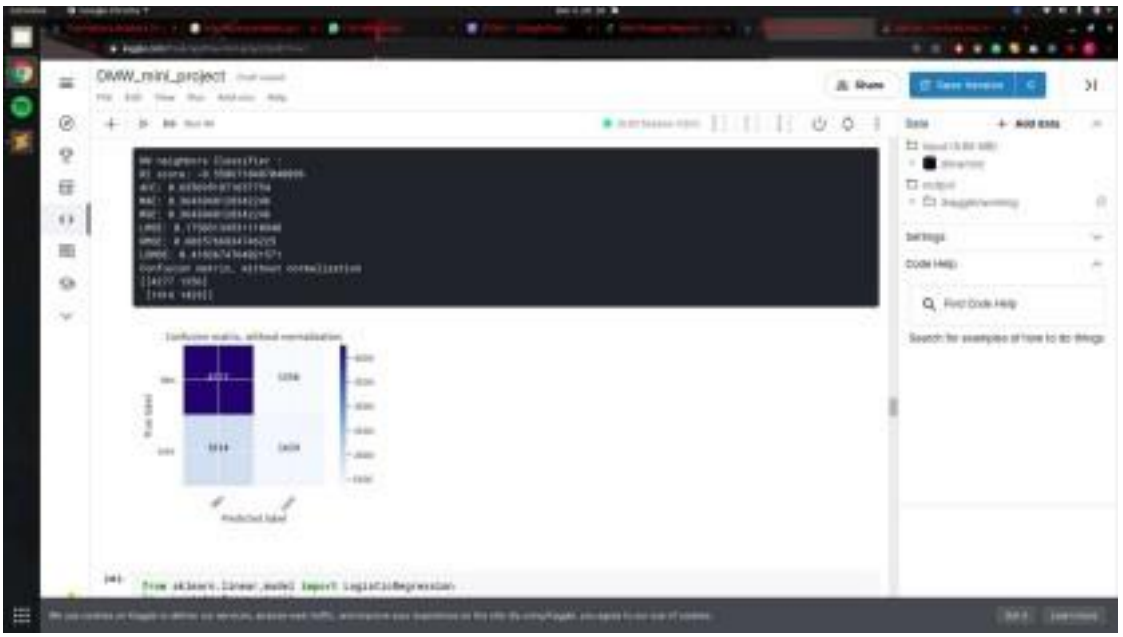

Figure 3: Output for K Neighbors classifier

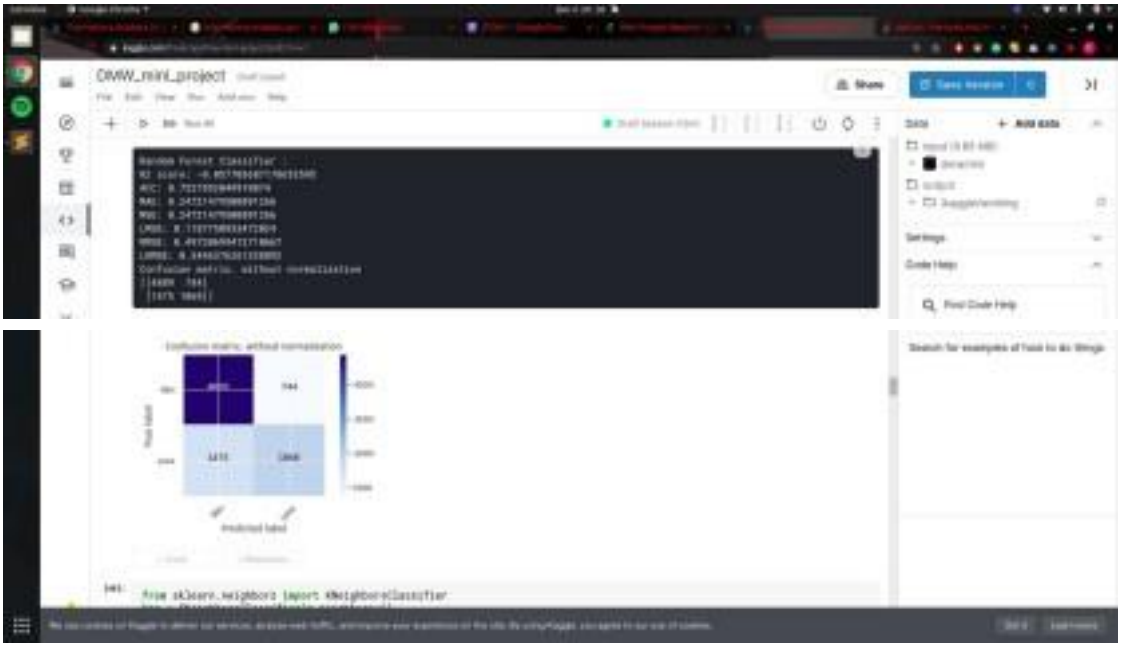

Figure 4: Output for random forest classifier

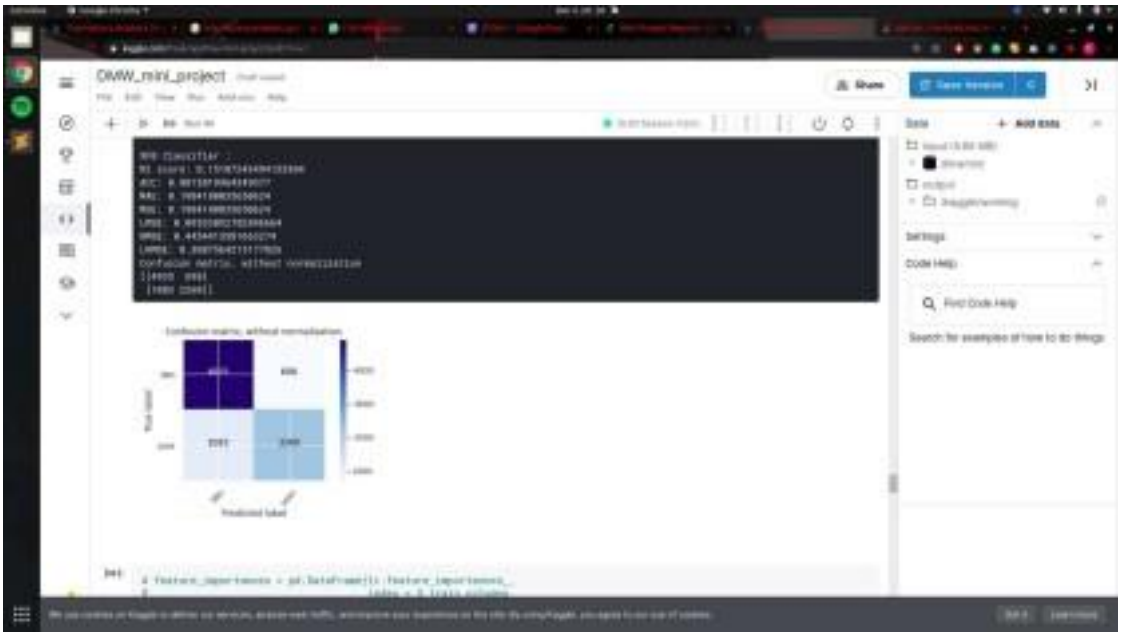

Figure 5: Output for XGBoost classifier

4 Result

The accuracy for the XGBoost classifier is around 80%. while that of other models is lesser. The following are confusion matrices of various model

outputs.

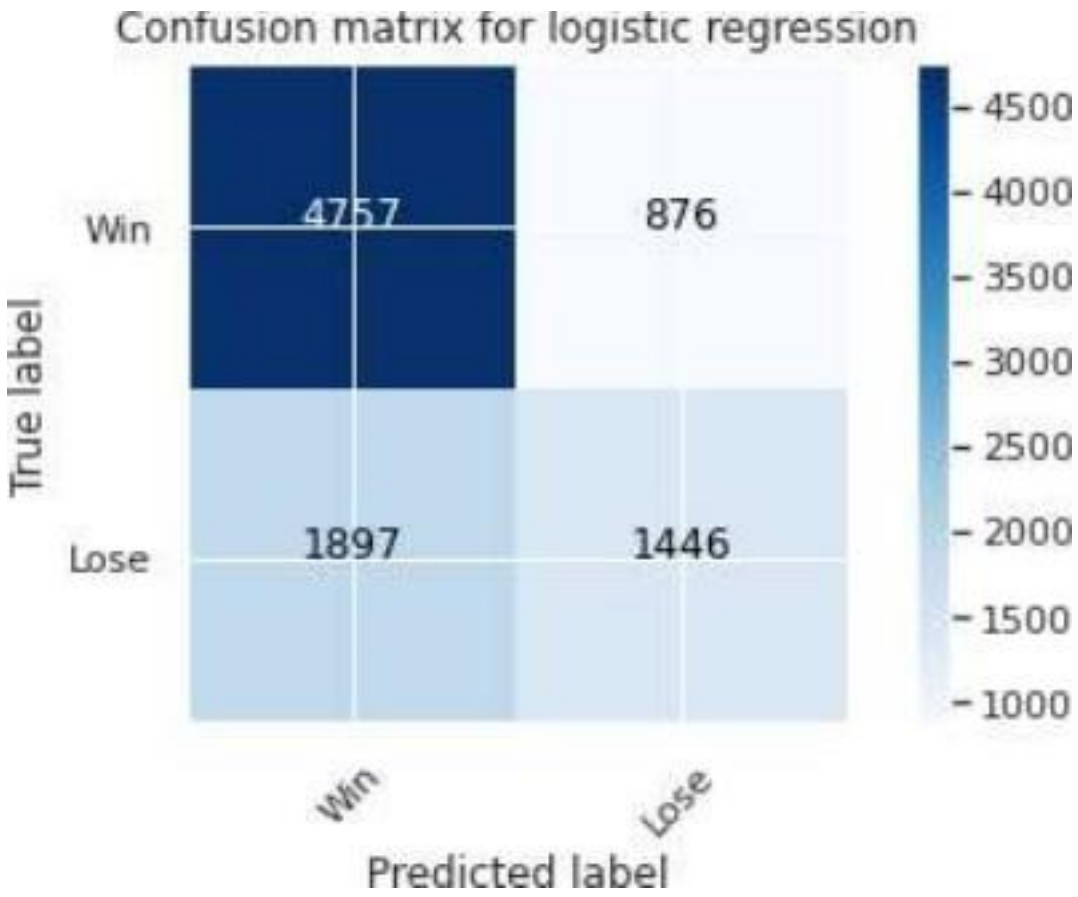

Figure 6: Logistic regression

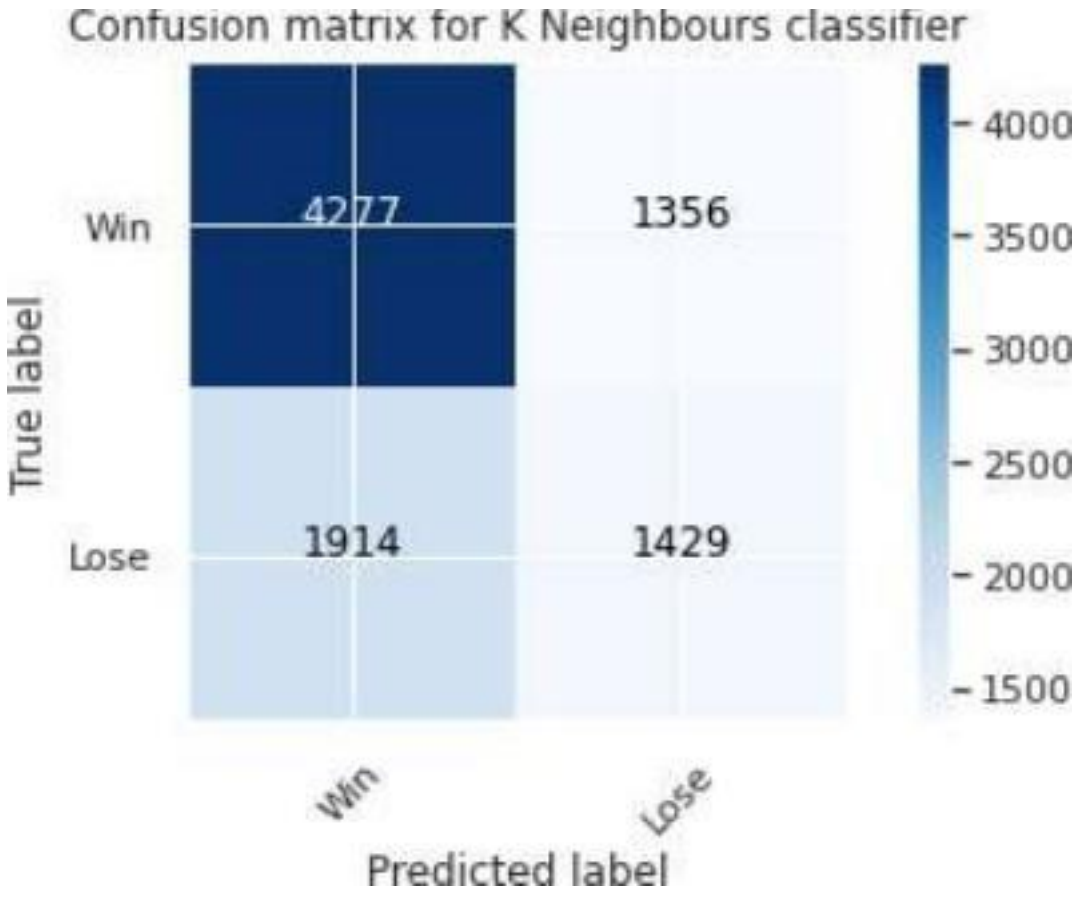

Figure 7: K Neighbours classifier

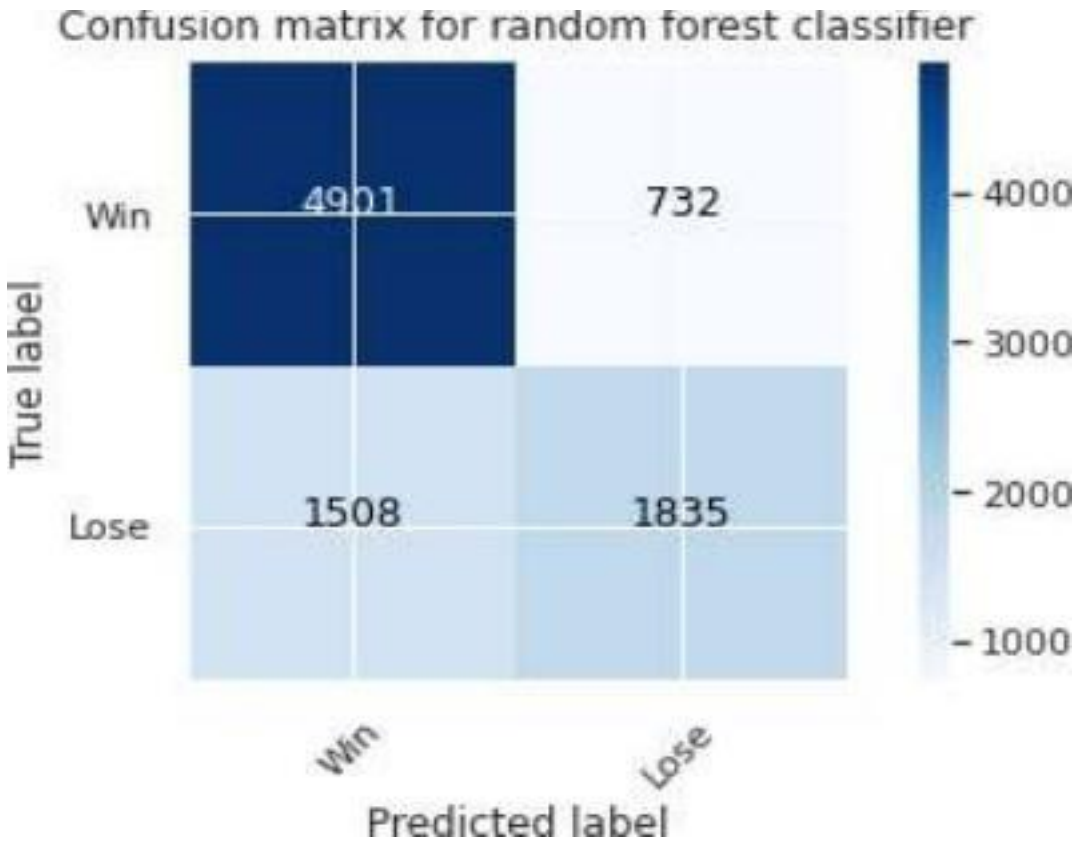

Figure 8: random forest classifier

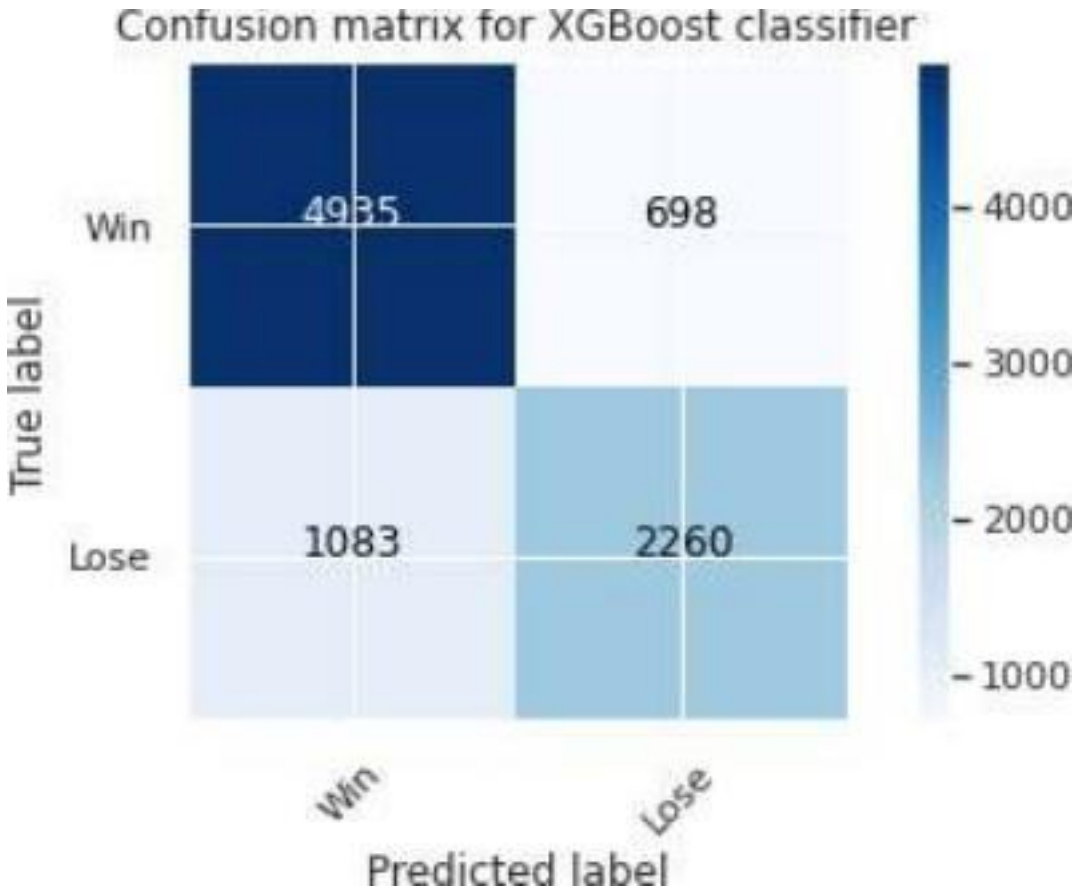

Figure 9: XGBoost classifier

7 Conclusion

We have analyzed the hockey game dataset and performed data pre-processing steps. We have experimented multiple classification models and found out the best performer among them. We have then used this model to make predictions on the test dataset.

References

[1] https://www.kaggle.com/c/datawiz19round1/data
[2] https://seaborn.pydata.org/index.html
[3] Jiawei Han,Micheline Kamber, Jian Pei, Data Mining Concepts and Techniques

List of Figures

1 Boxplots 4
2 Output for logistic regression .....
3 Output fo K Neighbours classifier……
4 Output for random forest classifier 7
5 Output for XGBoost classifier .....
6 Logistic regression 8
7 K Neighbours classifier 9
8 random forest classifier..... 10
9 XGBoost classifier .....

List of Tables

1 Null Counts 3