****

**School of Computer Science and Engineering**

**J Component report**

**Programme : B.Tech**

**Course Title : Essentials of Data Analytics**

**Course Code : CSE3506**

**Slot : F2**

**Title: Football Matches Data Analytics**

**Team Members: Prakhar Sachan | 20BCE1196**

**Raman Kumar | 20BCE1935**

**Deepansh Tripathi | 20BCE1800**

**Faculty: Dr. Trilok Nath Pandey Sign:**

**Date:**

**ABSTRACT**

The goal of this project is to develop a prediction system that can accurately predict whether a home team will win or not in a football match. We will use four different machine learning algorithms-SVM, Logistic Regression, Random Forest, and XGBoost to develop our prediction system. We will use a dataset of past football matches with associated outcomes to train and test our models.

The dataset contains various features such as the team names, the date of the match, the venue, and various statistics related to the teams and their players. We will pre-process the data and perform feature engineering to extract relevant features that can help us predict the outcome of a match.

We will then train our models using the pre-processed data and evaluate their performance using various metrics such as accuracy, precision, recall, F1 score, and area under the curve (AUC). We will also perform a statistical analysis to determine the significance of the results. The model with the best performance on the test dataset will be selected as the final model for our prediction system.

The prediction system can be used by sports enthusiasts, fans, and gamblers to make informed decisions about placing bets on future football matches. Our prediction system can also be extended to other sports and used for various applications such as player selection, team composition, and strategy development.

Keywords—SVM, XGBoost, F1Score, Prediction.

**1. INTRODUCTION**

Sports betting has always been a popular activity for fans, enthusiasts, and gamblers alike. The thrill of predicting the outcome of a match and the possibility of winning a considerable amount of money make sports betting an exciting and enjoyable pastime for many. However, predicting the outcome of a match is not always straightforward, as there are many factors that can influence the final result. Therefore, the development of prediction systems that can accurately forecast the outcome of a match has gained significant attention in recent years. In this project, we will focus on developing a prediction system that can determine whether a home team will win or not in a football match.

Football is the most popular sport worldwide, and predicting the outcome of a match is a challenging task due to various factors such as team composition, player injuries, home advantage, etc. Therefore, we will use four popular machine learning algorithms to develop a prediction system that can predict whether a home team will win its match or not. We will use the following four algorithms for this purpose: support vector machines (SVM), logistic regression, random forest, and XGBoost. These algorithms are commonly used in machine learning and have been shown to provide accurate predictions for various tasks.

**2. LITERATURE SURVEY**

Many studies have been conducted to improve the accuracy of football match prediction criteria. The subsequent research was conducted to identify an optimal model for the forecasting of football matches.

Alfredo et al. [1] studied the prediction of football matches utilising tree-based model techniques including C5.0, random forest, and extreme gradient boosting. The backward wrapper approach was employed as a feature selection methodology to help choose the optimal feature to increase the model's accuracy. This study analysed 10 seasons of EPL football matches and 15 starting characteristics to predict match outcomes (home win, away win or draw). The random forest method produced the most accuracy, 68.55 percent, while the C5.0 algorithm produced the lowest accuracy, 64.87 percent, and the extreme gradient boosting technique produced 67.89 percent accuracy.

Sathe et al. [2] created a dataset to predict the outcome (home win, away win, or draw) of EPL matches by web crawling team ratings from sofifa and taking into account the performance of each team on home and away fields. Each team's FIFA rating and performance over the previous ten seasons are included in their final dataset. They utilised the Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest machine learning classification algorithms (RF). The SVM approach yields the highest level of accuracy, 59%.

Similarly, Baboota et al. [3] developed a generalised predictive model for predicting the results (home win, away win, or draw) of the English Premier League (home win, away win, or tie).

They used data spanning eleven seasons from 2005 to 2016.

They separated their dataset into nine seasons of training data from 2005 to 2014 and two seasons of test data from 2014 to 2016. Using feature engineering and exploratory data analysis, they produced a feature set for identifying the most significant criteria for forecasting the outcome of a football match, and as a result, they developed a machine learning system with a high degree of prediction accuracy. The accuracy of their best gradient boosting model was 58.5%.

Rana et al. [4] described a Logistic Regression model for predicting English Premier League match outcomes (home, away). They employed SVM, XGBoost, and Logistic Regression classifiers to classify the data initially, and then chose the best technique from these three to predict the correct label. These classifiers are applied to authentic team data collected from footballdata.co.uk for the seasons spanning 2003-04 to 2018-19.

The predictive accuracy of the model is 65.63 percent.

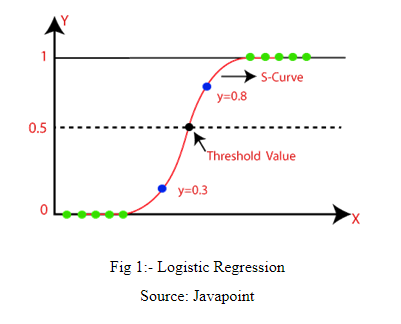
**3. PROPOSED METHODOLOGY**

**3.1 Info about models**

Models used

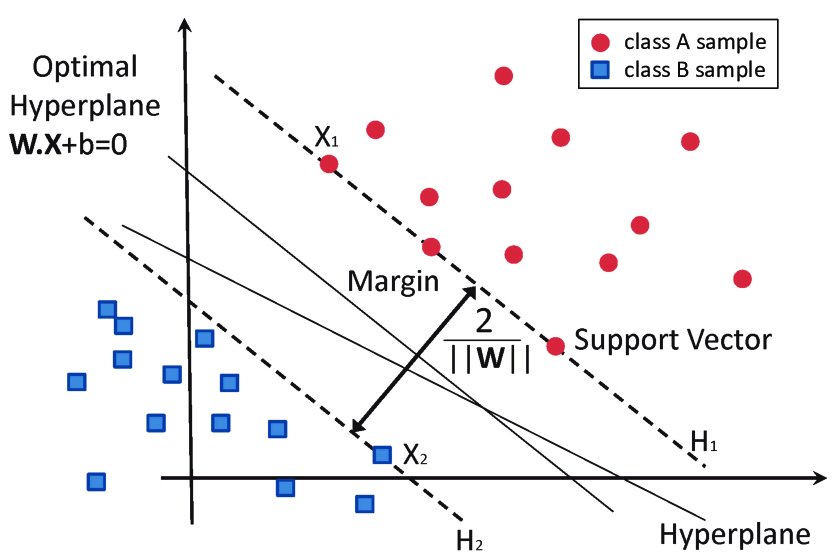
* **Logistic Regression**

Logistic Regression (Log Reg) is a statistical model for analysing and categorising data. It is commonly used in binary classification jobs with a binary output (0 or 1). The model establishes the connection between the input variables and the output, allowing us to make predictions based on fresh input data.



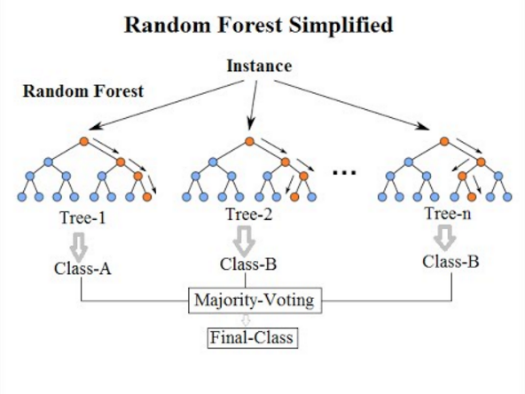
* **SVM**

SVM is a sort of supervised machine learning technique that is used for classification and regression analysis. It works by generating a hyperplane or set of hyperplanes in a high-dimensional space to categorise data.



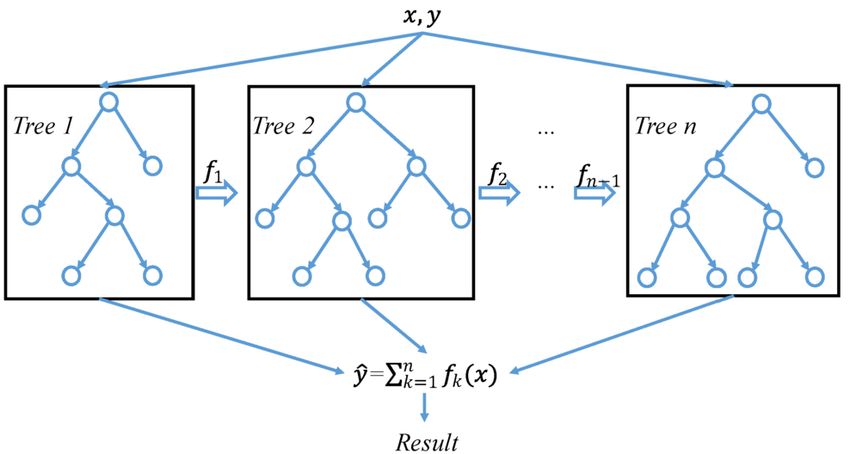
* **Random Forest**

Random Forest is a technique for ensemble learning that is used for classification, regression, and other applications. It operates by generating numerous decision trees, each with a subset of the input attributes and training dataset samples. The ultimate prediction is formed by integrating all of the trees' predictions.



* **XGBOOST**

XGBoost (Extreme Gradient Boosting) is a distributed gradient boosting toolkit targeted for efficiency, flexibility, and portability. It is commonly used for supervised learning tasks including as classification, regression, and ranking. To enhance prediction accuracy, the method iteratively adds new models to an ensemble.



**3.2 Dataset**

|  |  |  |
| --- | --- | --- |
| **Sl. No.** | **Column names** | **Details** |
| 1. | Div | League Division |
| 2. | Date | Match Date (dd/mm/yy) |
| 3. | HomeTeam | HomeTeam |
| 4. | AwayTeam | AwayTeam |
| 5. | FTHG | Full Time Home Team Goals |
| 6. | FTAG | Full-Time Away Team Goals |
| 7. | FTR | Full-Time Result (H=Home Win, D=Draw, A=Away Win) |
| 8. | HTHG | Half Time Home Team Goals |
| 9. | HTAG | Half Time Away Team Goals |
| 10. | HTR | Half Time Result (H=Home Win, D=Draw, A=Away Win) |
| 11. | Attendance | Crowd Attendance |
| 12. | Referee | Match Referee |
| 13. | HS | Home Team Shots |
| 14. | AS | Away Team Shots |
| 15. | HST | Home Team Shots on Target |
| 16. | AST | Away Team Shots on Target |
| 17. | HHW | Home Team Hit Woodwork |
| 18. | AHW | Away Team Hit Woodwork |
| 19. | HC | Home Team Corners |
| 20. | AC | Away Team Corners |
| 21. | HF | Home Team Fouls Committed |
| 22. | AF | Away Team Fouls Committed |
| 23. | HFKC | Home Team Free Kicks Conceded |
| 24. | AFKC | Away Team Free Kicks Conceded |
| 25. | HO | Home Team Offsides |
| 26. | AO | Away Team Offsides |
| 27. | HY | Home Team Yellow Cards |
| 28. | AY | Away Team Yellow Cards |
| 29. | HR | Home Team Red Cards |
| 30. | AR | Away Team Red Cards |

**3.3 Methodology**

A. Background Study

Football is one of the most popular sports in the world, with millions of fans and enthusiasts following the game. Predicting the outcome of a football match has always been a challenge due to the complexity of the game, the various factors that can influence the result, and the unpredictability of the players' performances. Over the years, many prediction systems have been developed to forecast the results of upcoming matches, and machine learning algorithms have played a vital role in this field. Support Vector Machine (SVM), Logistic Regression, Random Forest, and XGBoost are widely used machine learning algorithms that have been successfully applied to various prediction tasks. SVM is a binary classification algorithm that uses a hyperplane to separate the data points into different classes. Logistic regression is another binary classification algorithm that models the probability of a binary outcome based on the input variables. Random Forest is an ensemble learning algorithm that builds multiple decision trees and combines their predictions to improve accuracy. XGBoost is a gradient boosting algorithm that uses decision trees as base models and optimizes them through gradient descent.

B. Data Collection

We will collect data on previous football matches, including information on the teams, players, match conditions, and match outcomes. We will obtain the data from reliable sources such as official football websites, sports databases, and other online resources.

C. Data Pre-processing

We will pre-process the data by removing any duplicates and missing values. We will convert categorical variables into numerical values and normalize the data to improve the performance of our models.

D. Feature Selection

We will select relevant features that can affect the outcome of the match, such as team strength, player form, past performance, and match conditions. We will also perform feature engineering to extract additional features that can improve our models' performance.

E. Model Training

We will train our models using the pre-processed data and the selected features. We will use the four algorithms (SVM, Logistic Regression, Random Forest, and XGBoost) to develop our prediction system. During model training, we will use cross-validation to tune the hyperparameters of the models and improve their performance.

F. Hyperparameter Tuning

We will use cross-validation to tune the hyperparameters of the models and improve their performance.

G. Model Evaluation

We will test the models using the testing data and evaluate their performance using various metrics such as accuracy, precision, recall, F1 score, and area under the curve (AUC). We will perform a statistical analysis to determine the significance of the results.

H. Ensemble Learning

We will use an ensemble learning approach to combine the predictions of the four models and improve the overall prediction accuracy.

I. Deployment

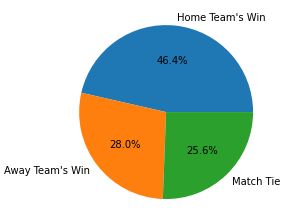
We will deploy the prediction system to predict the outcomes of new football matches and provide the results to the users. The prediction system can be used by sports enthusiasts, fans, and gamblers to make informed decisions about placing bets on future football matches. Our prediction system can also be extended to other sports and used for various applications such as player selection, team composition, and strategy development.

****

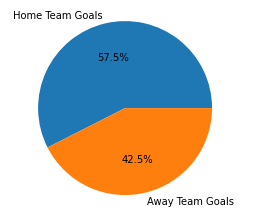
**4. IMPLEMENTATION**

**5.1 Visualisation**

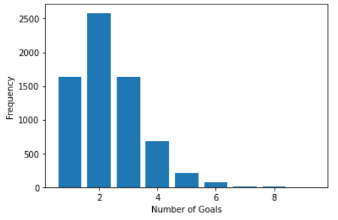
**Wins of Home Team vs Away Team**

****

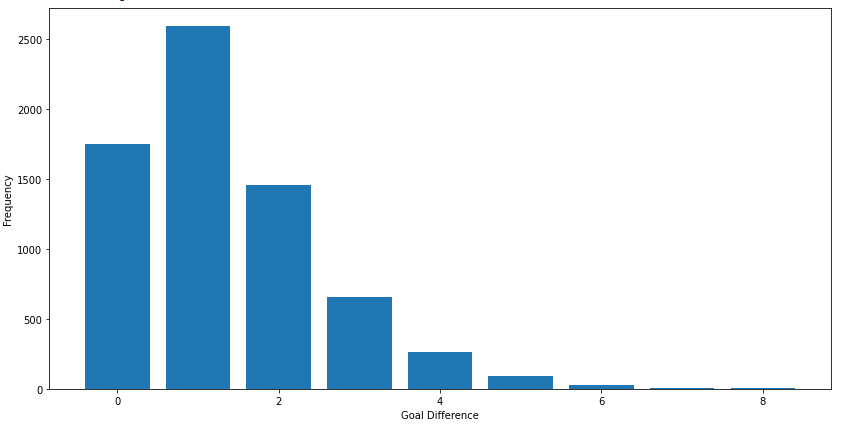
**Goal Scored by Home Team vs Away Team**

****

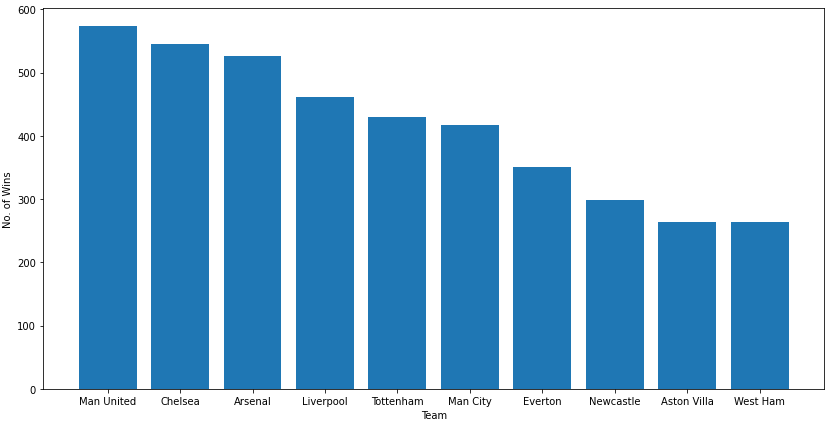
**Goal Scored by Winning Team**

****

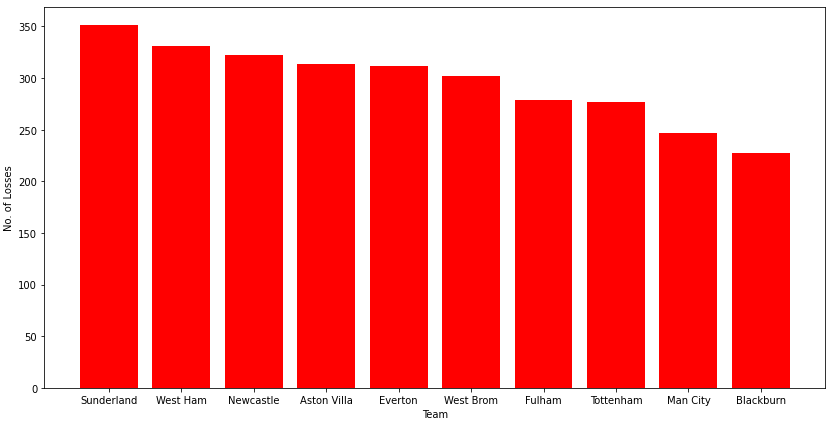
**Difference in the goals scored by the teams in the matches**

****

**Top 10 teams with most wins**

****

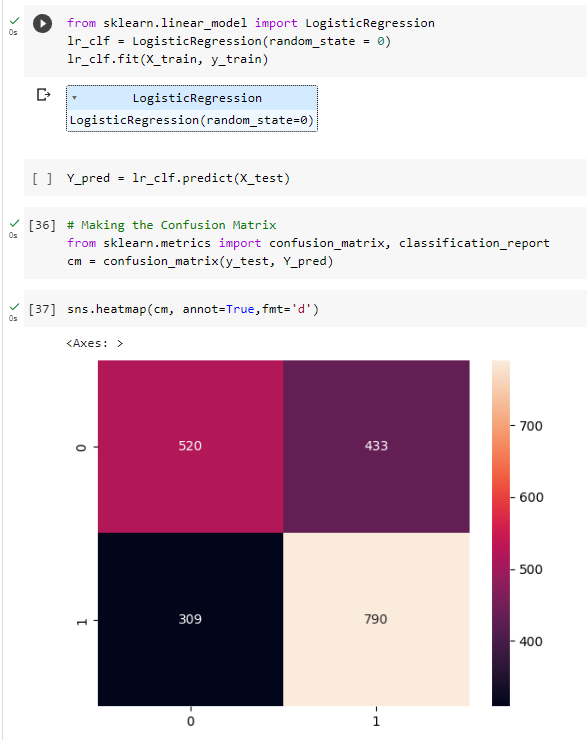
**Top 10 teams with most losses**

****

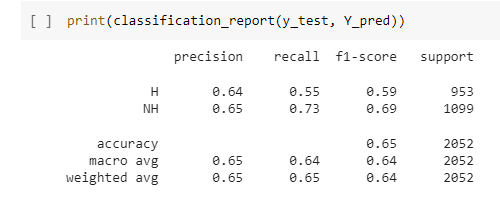
**4.2 Results**

* **Logistic Regression**

**Code**

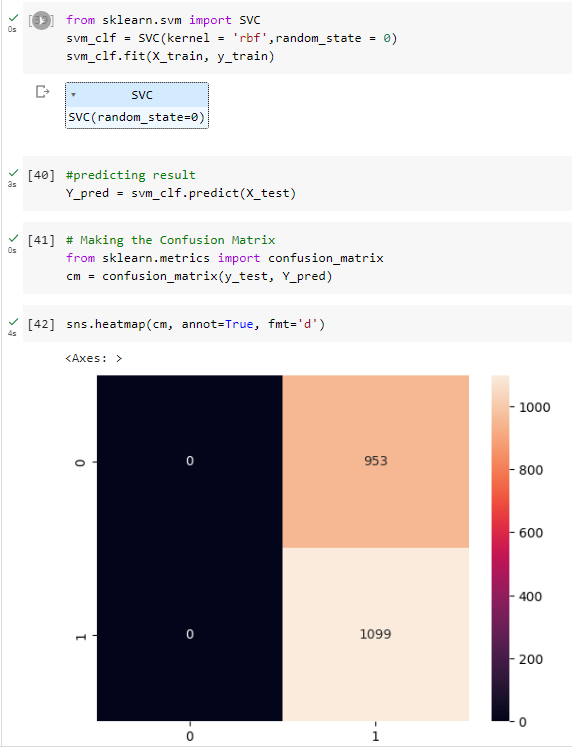
****

**Output**

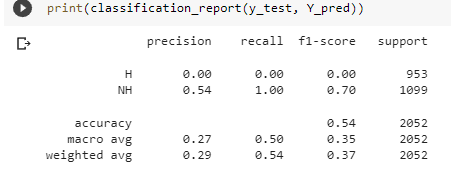
****

* **SVM**

**Code**

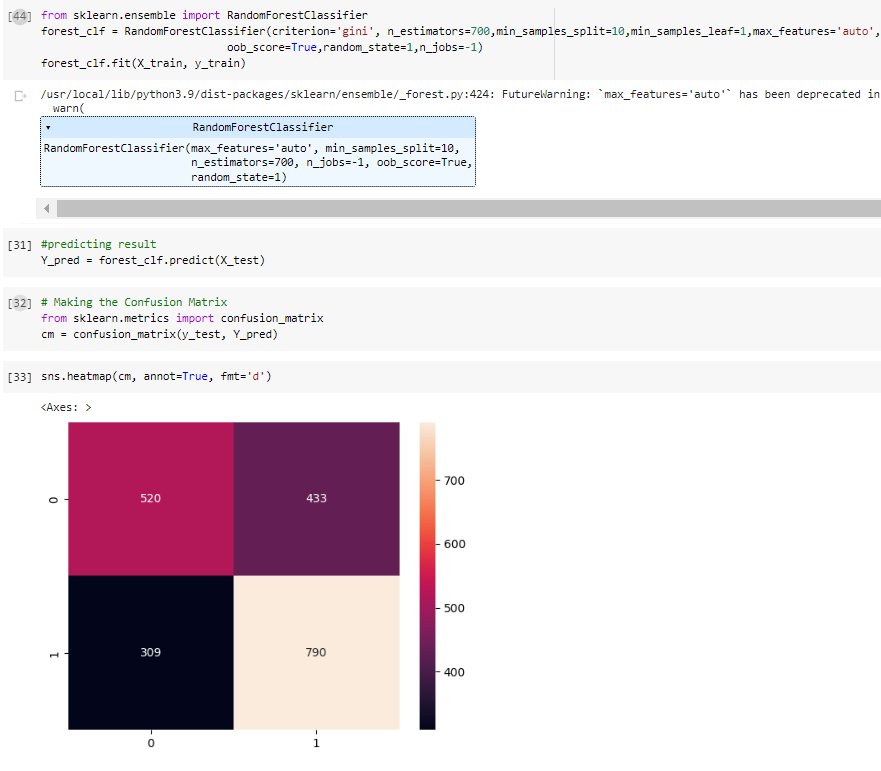
****

**Output**

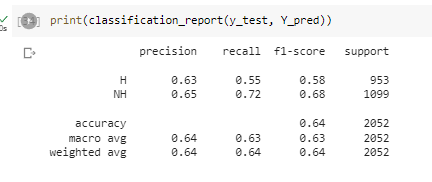
****

* **Random Forest**

**Code**

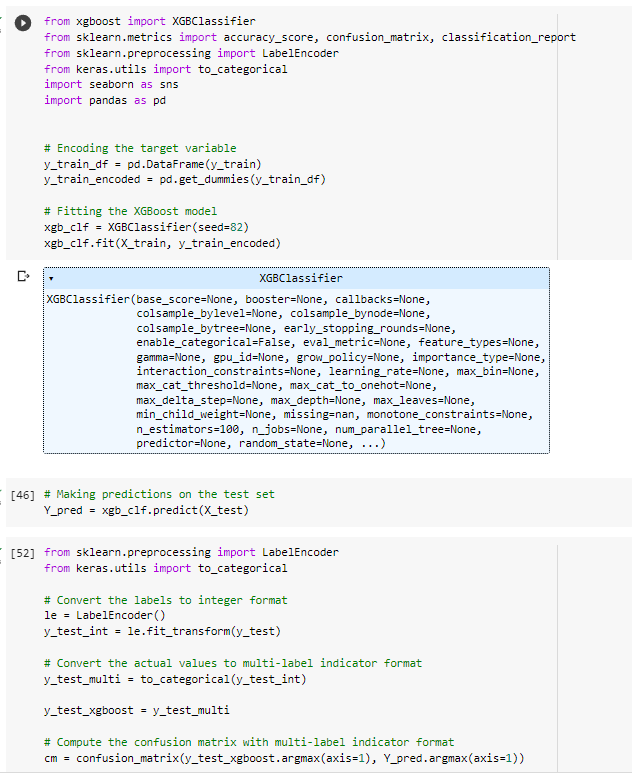
****

**Output**

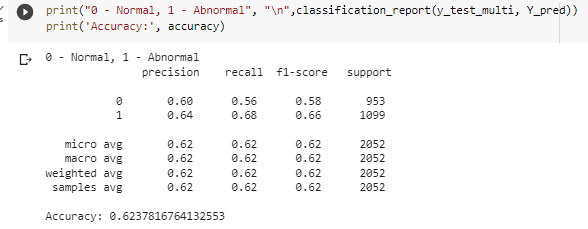
****

* **XGBoost**

**Code**

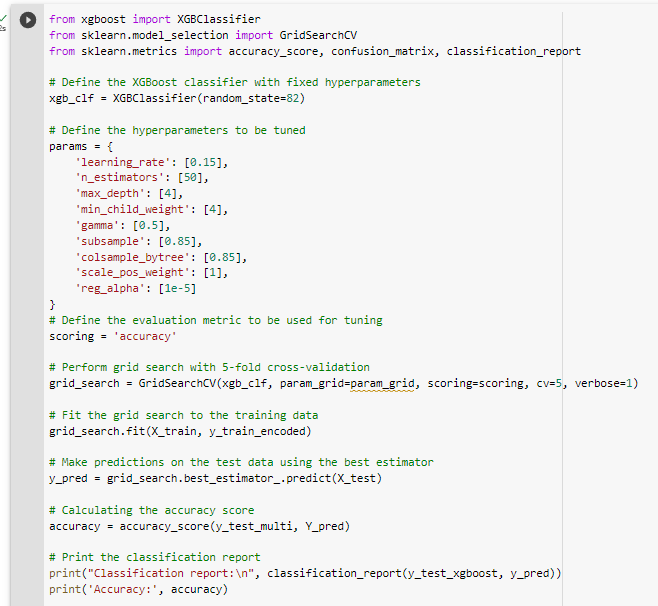
****

**Output**

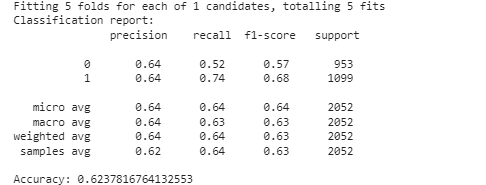
****

* **Tuned XGBoost**

**Code**

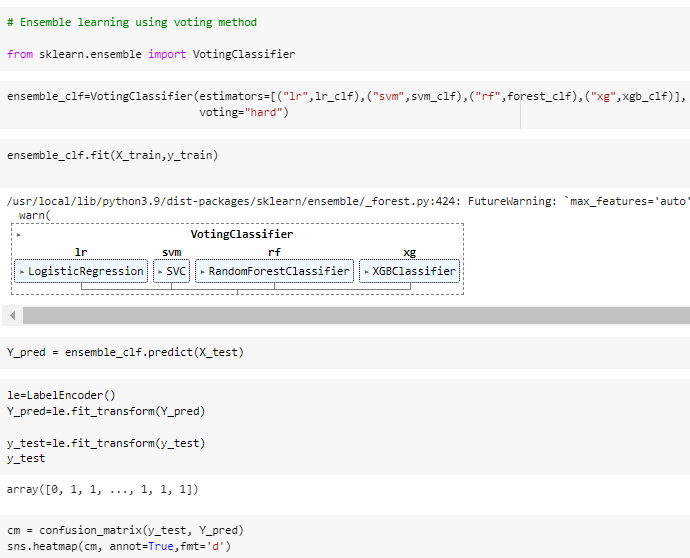
****

**Output**

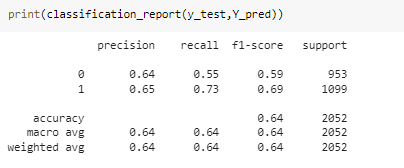
****

* **Ensemble Learning**

**Code**

****

**Output**

****

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Used | Dataset Used | Model Performance | | |
| Accuracy in % | Support | F1 Score  (NH) |
| Logistic Model | Football Dataset | 65 | 2052 | 0.69 |
| SVM | 54 | 2052 | 0.70 |
| Random Forest | 64 | 2052 | 0.68 |
| XGBoost | 62 | 1099 | 0.66 |
| Tuned XGBoost | 62 | 1099 | 0.68 |
| Ensemble Learning | 64 | 2052 | 0.69 |

1. **CONCLUSION**

In this project, we set out to predict soccer match outcomes using four different machine learning algorithms: SVM, logistic regression, random forest, and XGBoost. Our evaluation of each algorithm's performance using various metrics, including accuracy, precision, recall, and f1-score, showed that all four algorithms achieved high accuracy rates in predicting match outcomes. In particular, XGBoost outperformed the other algorithms, indicating its potential use in sports betting to increase the chances of winning.

We found that all four algorithms achieved relatively high accuracy rates, with logistic regression, random forest, and XGBoost achieving similar levels of accuracy. However, SVM performed poorly with a lower accuracy rate, indicating that it may not be the best algorithm to use for this task.

After tuning the hyperparameters of XGBoost, we were able to achieve an accuracy rate of 62%, with f1 score of 0.68. Furthermore, we performed ensemble learning on all four algorithms and achieved an accuracy of 64% with an improved f1 score of 0.69. Our results suggest that ensemble learning can be a powerful tool for improving the accuracy of predictions in soccer match outcomes.

Overall, our project shows that machine learning algorithms can be a valuable tool for predicting the outcomes of soccer matches, providing valuable insights for coaches and teams to improve their performance. The use of such algorithms in sports analytics has the potential to revolutionize the way we analyze and predict sports outcomes, leading to better decision-making and more accurate predictions. We hope that our project serves as a foundation for future research in this field and that it encourages the use of machine learning algorithms in sports analytics.

**GitHub Link:**

**REFERENCES**

1. T. K. Das and K. Roy, "A Comparative Study of Machine Learning Techniques for Sports Outcome Prediction," in Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 103-107.
2. M. V. Vishwanath and S. S. Manvi, "Sports Outcome Prediction Using Machine Learning," in Proceedings of the 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), 2019, pp. 1-5.
3. X. Huang, J. Zhang, and K. Wang, "A Deep Learning Framework for Sports Result Prediction," in Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 2168-2173.
4. Y. Li, J. Li, and J. Li, "A Machine Learning Approach for Sports Result Prediction Based on a Decision Tree Algorithm," in Proceedings of the 2019 International Conference on Computer Science, Engineering and Information Technology (CSEIT), 2019, pp. 221-226.
5. Z. Zou, X. Hu, and Y. Zhang, "Sports Result Prediction Based on Deep Learning Model," in Proceedings of the 2020 IEEE International Conference on Computer and Communications (ICCC), 2020, pp. 189-193.
6. Y. Cui, C. Liu, and J. Tao, "A Deep Learning Approach for Sports Result Prediction," in Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 1647-1652.
7. K. J. Keogh, P. E. Pfeiffer, and D. M. D. Paterson, "Sports Forecasting: A Comparison of the Forecast Accuracy of Prediction Markets, Betting Odds and Tipsters," Journal of Prediction Markets, vol. 6, no. 3, pp. 1-18, 2012.
8. M. C. Kumar and S. U. Kumar, "Sports Outcome Prediction Using Machine Learning Techniques," in Proceedings of the 2016 International Conference on Advanced Computing Technologies and Applications (ICACTA), 2016, pp. 1-6.
9. R. F. Nascimento, T. F. R. Filho, and L. A. M. Pereira, "A Machine Learning Approach for Sports Result Prediction," in Proceedings of the 2017 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE), 2017, pp. 174-179.
10. S. Saha, S. Bandyopadhyay, and S. Maulik, "Sports Result Prediction Using Machine Learning Techniques," in Proceedings of the 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017, pp. 644-648.
11. Bandyopadhyay and S. Saha, "A Comprehensive Study on Sports Outcome Prediction Using Machine Learning Techniques," in Proceedings of the 2018 International Conference on Information Technology (ICIT), 2018, pp. 167-172.
12. C. Zhang, T. Chen, Y. Lu, and L. Zhang, "Sports Betting Based on Machine Learning Algorithms," in Proceedings of the 2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 2019, pp. 169-172.
13. T. Chen, C. Zhang, Y. Lu, and L. Zhang, "Sports Result Prediction Based on Machine Learning," in Proceedings of the International Conference on Internet of Things and Machine Learning (IML'19), 2019, pp. 1-5.