UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering, Built Environment and Information Technology
Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Department of Informatics

INDIVIDUAL ASSIGNMENT

Surname	Skhonde							
Initials	SC							
Student Number	2	0	5	8	7	0	5	9
Module Code	**INF**				7	9	1	
Assignment number	Assignment 1							
Name of Lecturer	Dr. WA NKONGOLO MIKE NKONGOLO							
Date of Submission	2024/09/20							

Declaration:

I declare that this assignment, submitted by me, is my own work and that I have referenced all the sources that I have used.

The University of Pretoria commits itself to produce academic work of integrity. I affirm that I am aware of and have read the Rules and Policies of the University, more specifically the Disciplinary Procedure and the Tests and Examinations Rules, which prohibit any unethical, dishonest or improper conduct during tests, assignments, examinations and/or any other forms of assessment. I am aware that no student or any other person may assist or attempt to assist another student, or obtain help, or attempt to obtain help from another student or any other person during tests, assessments, assignments, examinations and/or any other forms of assessment.

Signature of Student	Simphiwe Skhonde

Introduction

Cybergaming just refers to gaming which takes place on the internet or in a cyberspace. Cybersecurity is needed with its level of importance growing due to the rise in cybercrimes (Alotaibi, Furnell, Stengel & Papadaki, 2016). A study by Alotaibi *et al.* (2016) was conducted to determine the best way to create awareness on cybersecurity. This is important because the internet is a relatively new resource and being able to navigate it is very important. Many people use online platforms to perform a number of activities, two of which are banking and shopping online (Alotaibi *et al.*, 2016). The e-commerce business has grown significantly and can be potentially dangerous for customers as this is a way for criminals to get credit card information. There are also ways for cybercriminals to break hack banking accounts and get banking information. These are the reasons why cybersecurity awareness is pivotal in the current landscape through which consumer are required to navigate. Previous studies have found that although users may have good IT skills, they lack in cyber security awareness and engage in weak practices. The awareness programs suggested include using gaming technology. Games are seen as an effective educational tool as they are engaging, cost effective and transferrable (Alotaibi *et al.*, 2016).

This paper aims to identify gamer behaviour and engagement to determine the effectiveness of this cyber security awareness game.

The data used in this assignment was collected from University of Pretoria honours students. The entire INF791 class contributed to building the dataset. The “CyberVigilance” game was uploaded for students to download and run through visual studio to prompt the game. The students then played the game, download it as a csv file and upload it on the shared google drive.

The data went through cleaning and transformation to be used in various supervised machine learning algorithms in order to be able to find relationships between the different variables and understand the causal relationships in video game variables to better understand players and their needs as well as using the feedback to possibly come up with better cyber security awareness games.

Literature Review

The purpose of this analysis is focused on gamer engagement and the behavioural patterns of gamers. The engagement metrics will try to get more information to determine how well liked the game is. For example, the frequency of games played as well as the average number of games played. This will inform how well liked the game is and gives insight into the enjoyability of games, as the less number of times a user played the game, the less likely that they enjoyed the game. More and more video gaming is gaining popularity as a favourite pastime activity (Gosztonyi, 2023). To improve the games, developers and designers need to get more information about the players. This information can be determined using the data from player profiles and performance. The most popular topics in gaming studies are gaming habits of the youth and psychological and social effects from gaming (Gosztonyi, 2023). Due to the limited scope and data in this assignment, the gamer engagement and behavioural patterns of gamers will be analysed and discussed.

Cyber security refers to the protection of digital assets and systems from crime (Maurice Hendrix, Ali Al-Sherbaz & Victoria Bloom, 2016). A large number of individuals are affected through online scams and identity theft from cybercrimes (Maurice Hendrix *et al.*, 2016). The paper by Maurice Hendrix *et al.* (2016) explored Serious Games and their influence on inspiring behavioural changes in the players. Serious Games are games which have a purpose aside from entertainment alone(Maurice Hendrix *et al.*, 2016). They have gained popularity over the years and a number of studies have been conducted to determine their effectiveness in their objective which in this case is to increase cyber awareness. Most of the results from the papers assessed indicated a positive response to the game. The study by Alotaibi *et al.* (2016) indicated that the learning experience can significantly be improved by mobile gaming applications. This means that they can instigate positive behavioural changes (Busch, Mattheiss, Hochleitner, Hochleitner, Lankes, Fröhlich, Orji & Tscheligi, 2016). The games are aimed at the general public rather than specialised professionals as cyber crimes affect not only corporations and organisations but the average person as well(Alotaibi *et al.*, 2016).

It is important to also ensure that the games are constantly monitored and maintained as the threats can take on many forms are change as time goes by therefore to keep up with the cybercrimes, the cyber security awareness also needs to constantly be evolving and adapting to the crisis faced (Alotaibi *et al.*, 2016)

This assignment aims to investigate the engagement of players however, after determining the engagement it is important to find ways to constantly improve the engagement. One way to do this is to personalise the player experience (Busch *et al.*, 2016).

There are papers such as that by Busch *et al.* (2016) which tries to identify different personality types such as n skill-oriented archetypes aesthetic-oriented archetypes as well as goal-oriented archetypes. These types can be even further broken down into specific player types to really try and categorise the different players in the cyberspace. Based on player behaviour, experience and game analytics, a player can be classified accordingly to improve their experience and get offered the right quests and challenges to maximise their enjoyment. This field aims to create new user-driven game mechanics(Bakkes, Spronck & van Lankveld, 2012; Busch *et al.*, 2016).

Data Collection and Preparation

As mentioned earlier, the data was collected internally within the INF791 class group. This game was used in the paper by (Wa Nkongolo, 2024). The students all had to play the game to built a dataset. The results of the game were uploaded to a google drive where all the students had to download the uploaded files as a zip file which contained all the results.

The results of the game were downloaded as a CSV file and included fields Nickname, Defender Score, Attacker Score, Time (sec), Winner and Level. These were the player names, the score of the defender which was the player, the score of the attacker which was the software, the total time it took for the game to be completed, the outcome of who won and the level of the player respectively.

The zip file downloaded contained the csv files of all participating students. The files were all merged into one and cleaned to remove extra columns and make sure they were in the same format. The resulting file was uploaded to Jupyter where duplicated were removed and the Attacker Score, Defender Score and Time (sec) fields were plotted on histograms. After this point of analysing the data through the histogram plots characteristics about the data such as the skewness and distribution can be determined. This data will be the input to the several machine learning algorithms therefore it is important to understand it before pushing the data into the algorithms. This also informed which corrections and/or mathematical need to be made. The Time frame was overtly skewed to the right with outliers therefore the log transformation was used for this column using the following formula:

```
df2['Time(sec) Log'] = np.log(df2['Time(sec)']+1)
```

The Attacker Score was slightly skewed to the right and bimodal. The Yeo Johnson transformation was used with the following formula:

```
df2['Attacker Score Yeo'], _ = stats.yeojohnson(df2['Attacker Score'])
```

The Defender Score was also slightly skewed to the right. The sqare root function was best suited to transform this data with the following formula:

```
df2['Defender Score Sqrt'] = np.sqrt(df2['Defender Score'])
```

Although only the Naïve Bayes Algorithm is the only one where the data needs to be normally distributed because of the assumption, it is safer to transform all the fields to make sure as it does not negatively affect the other algorithms.

The Time, Attacker and Defender Scores were then standardized to have a mean of 0 and standard deviation of 1. This was again done mainly for the Naïve Bayes algorithm to ensure that the models produces from it would be the most

The Winner, Level and Nickname fields were then encoded as they were categorical variables and were not represented numerically and all the algorithms require numerical values as inputs.

Methodology

The software used was Jupyter lab which uses Python and the libraries used in this assignment include:

- Pandas
- Numpy
- Maplotlib
- Seaborn
- Scipy
- Sklearn

The EDA was performed to determine the mean and standard deviation and plot it on the graphs. The first thing that was done was to determine the distribution of the data. Some more analytics were performed to evaluate the engagement levels of the game. A pie chart was produced to show the percentage of participants who played once versus those who played multiple times. A detailed chart was also made to show the exact number of times participants played and the percentage of participants who played multiple times.

Distribution of the data:

Time:

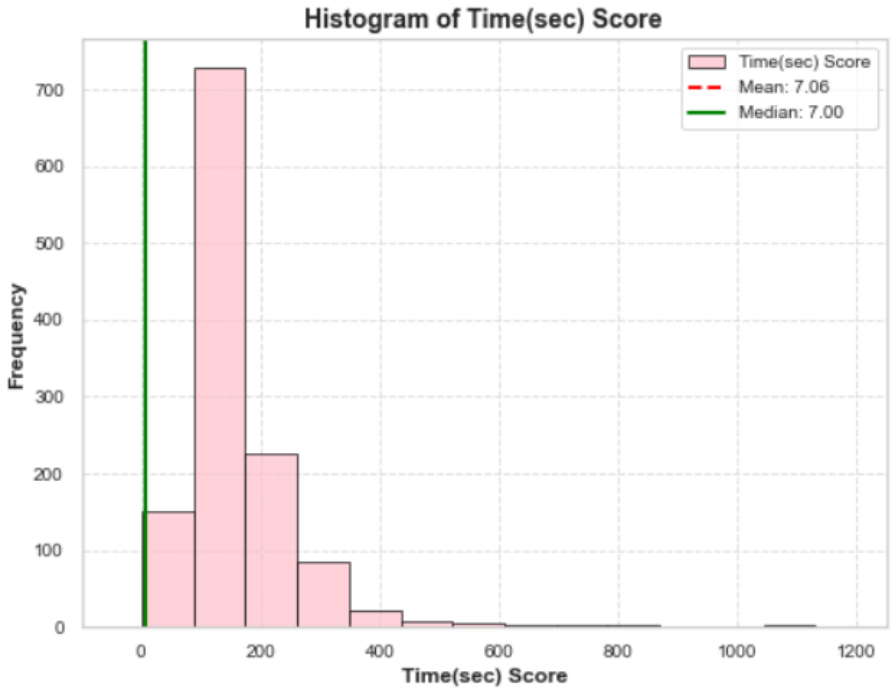


Figure 1: Time Unprocessed Data

Attacker Score:

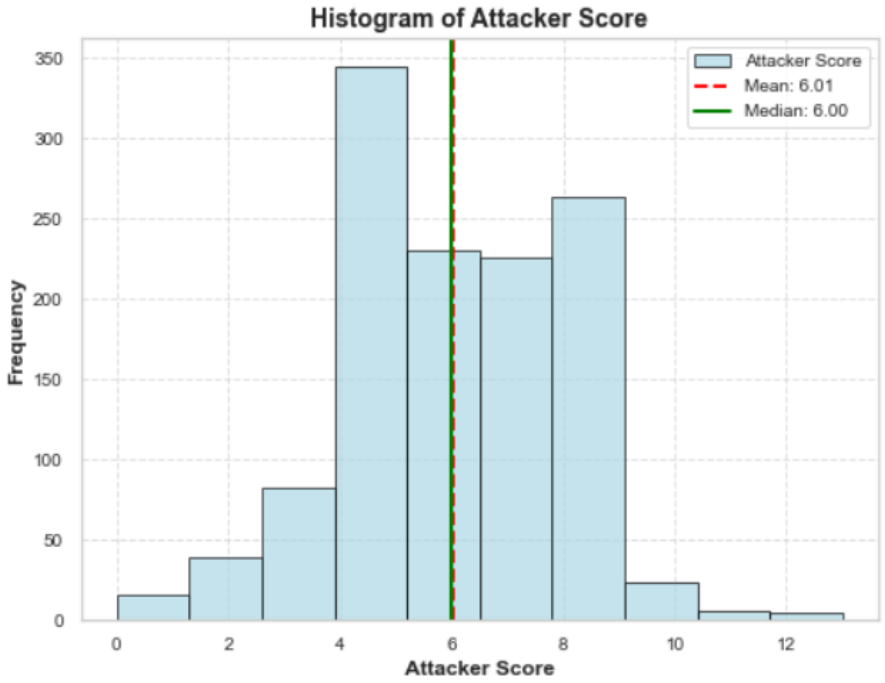


Figure 2: Attacker Score Unprocessed Data

Defender Score:

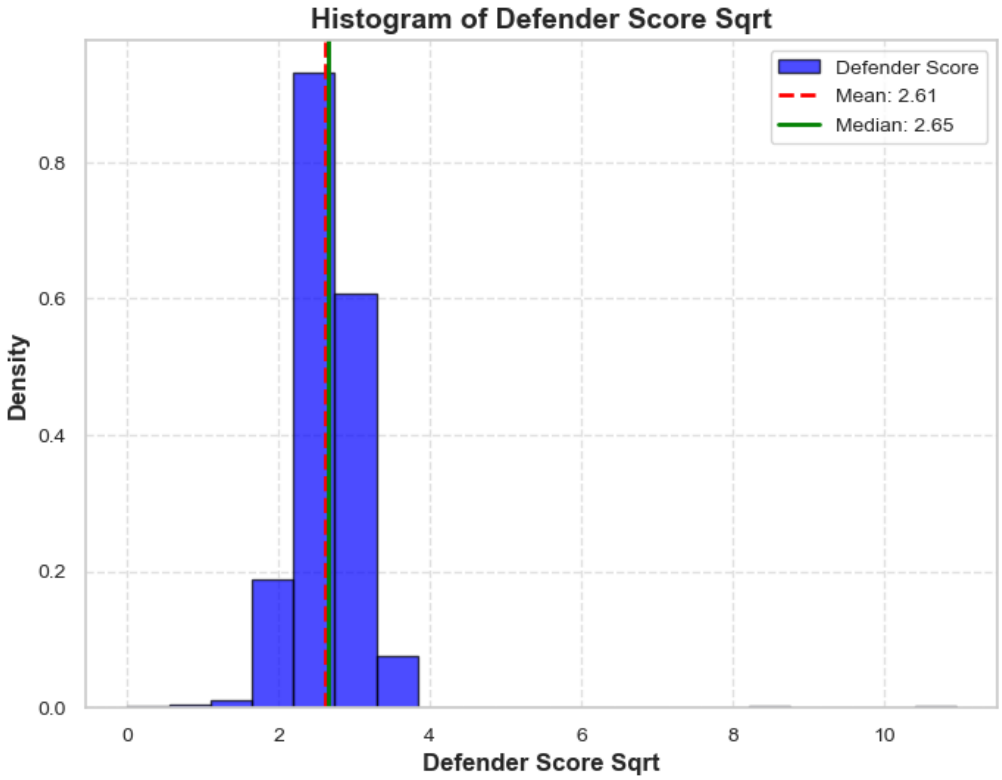


Figure 3: Defender Score Unprocessed Data

Transformed Plots

Time:

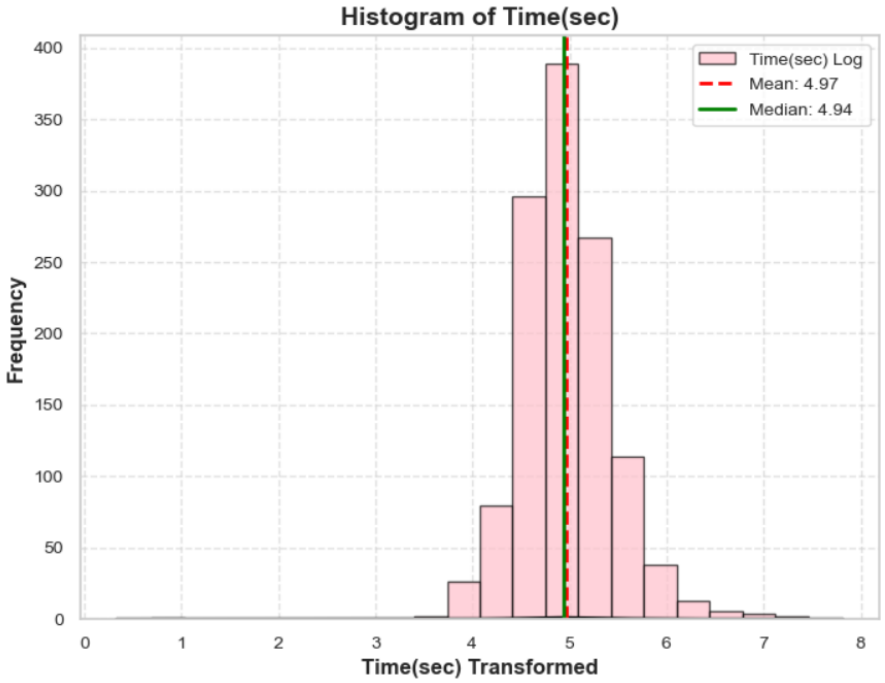


Figure 4: Transformed Time

Attacker Score:

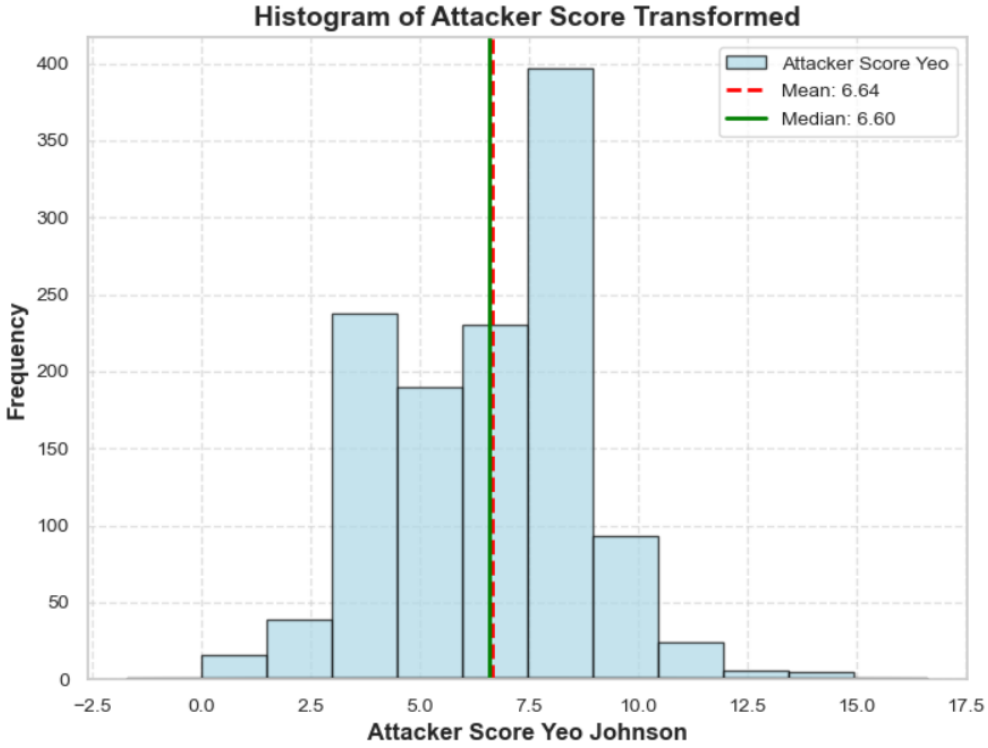
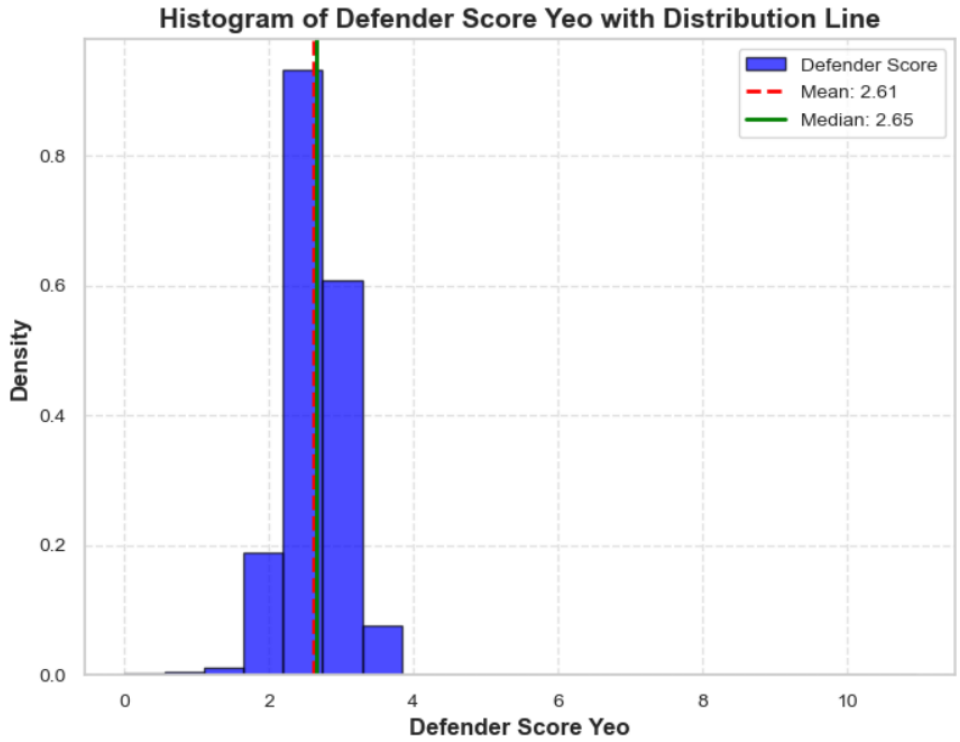


Figure 5: Transformed Attacker Score

Defender Score:



Normalised Plots

Time:

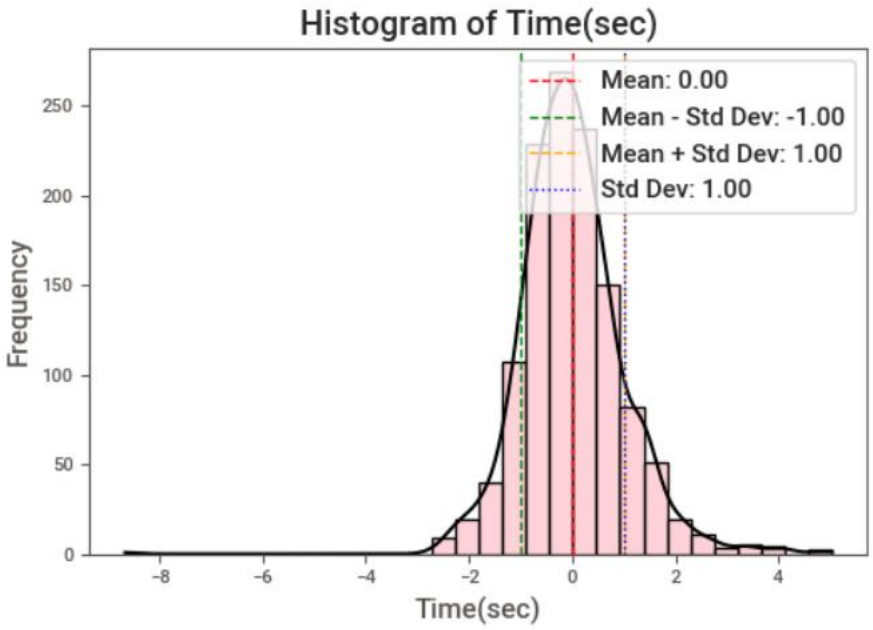


Figure 6: Normalised Time

Attacker Score:

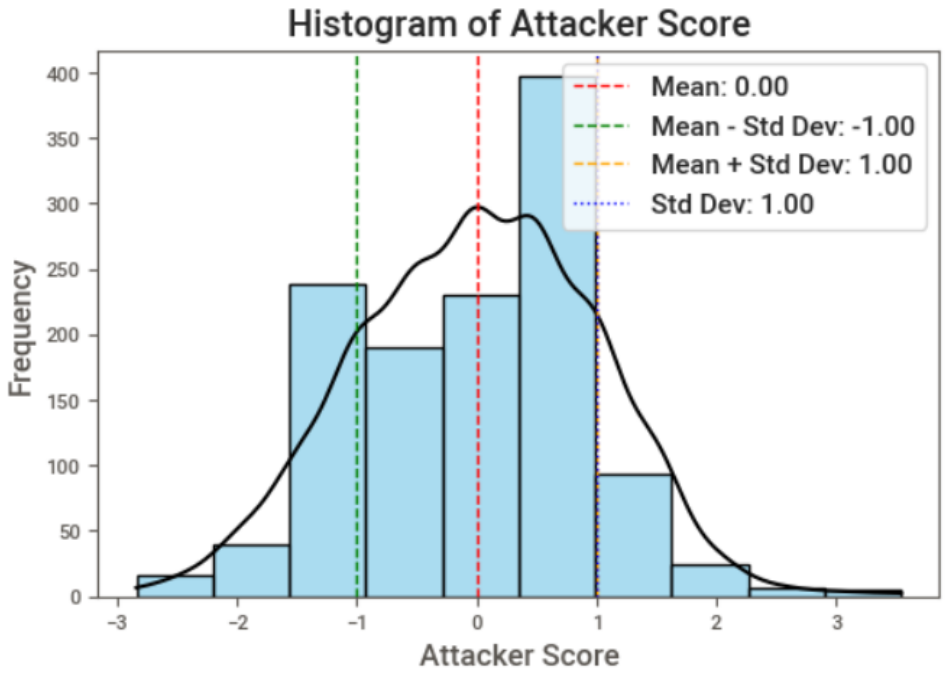


Figure 7: Normalised Attacker Score

Defender Score:

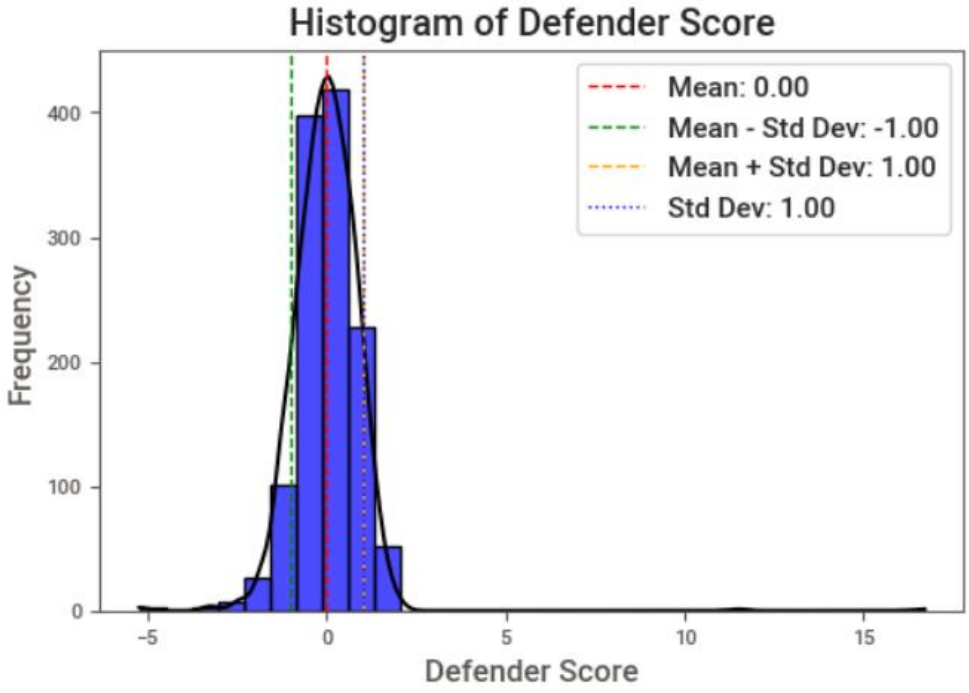


Figure 8: Normalised Defender Score

The win rate was also computed to determine the difficulty of the game which. The level which was linked to the outcome was also visualised. The defender (player) had the most number of wins however most players only played the game once therefore it would be nice to visualise the progression of players who played multiple times.

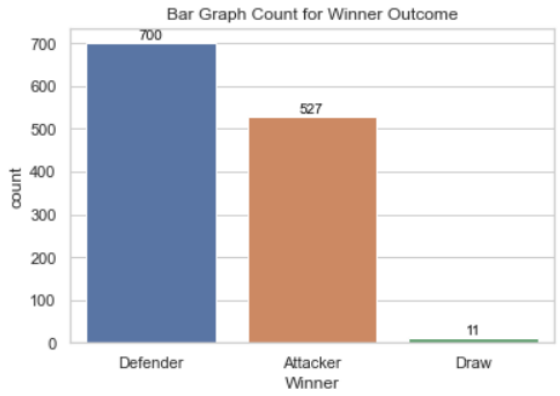


Figure 9: Winner Count

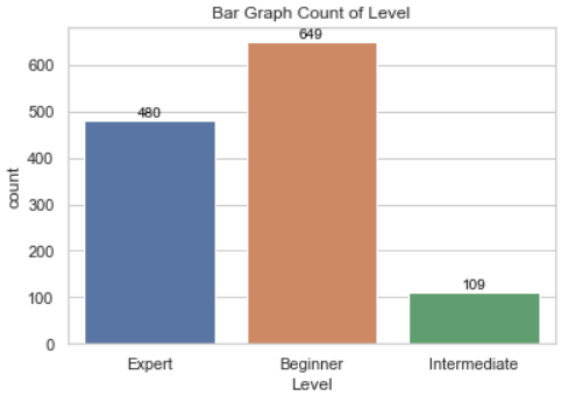


Figure 10: Level Count

The order of analysis was first performing the EDA to gain more information about the data and its suitability for the models. The different models are then performed and their results displayed. To check the precision and performance of the results, the evaluation metrics are then computed to rank the model performances.

The three machine learning algorithms used were Support Vector Machine (SVM), Random Forest and Naïve Bayes. These are three very common and popular models. SVM finds the most optimal hyperplane which separates the data into two classes. It works well with small datasets, high dimensional data and non-linear classification. It also handles outliers well.

Naïve Bayes calculates the class probability using the Bayes' theorem and assumes the class features are independent. It works well with large datasets and text classification.

Random Forest is an ensemble of decision trees as it take multiple trees and combines their predictions. It works well with large datasets as it uses a number of decision trees therefore the more data that can be fit into the model, the better it will work. The benefits include handling missing data well, reducing overfitting and providing feature importance.

For all three models, there was a test-train split in the data where 80% of the data was used to train the algorithm and the remaining 20% was used as the test dataset.

The model evaluation metrics used were the ROC which displays the trade off between the true positive and false positive rates at different thresholds, Lazy Predict which compares the performance of multiple machine learning models

Results

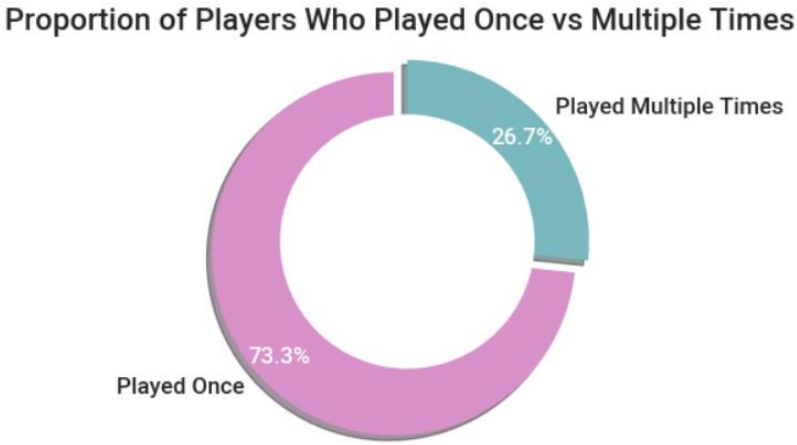


Figure 11: Single vs Multiple Time Players

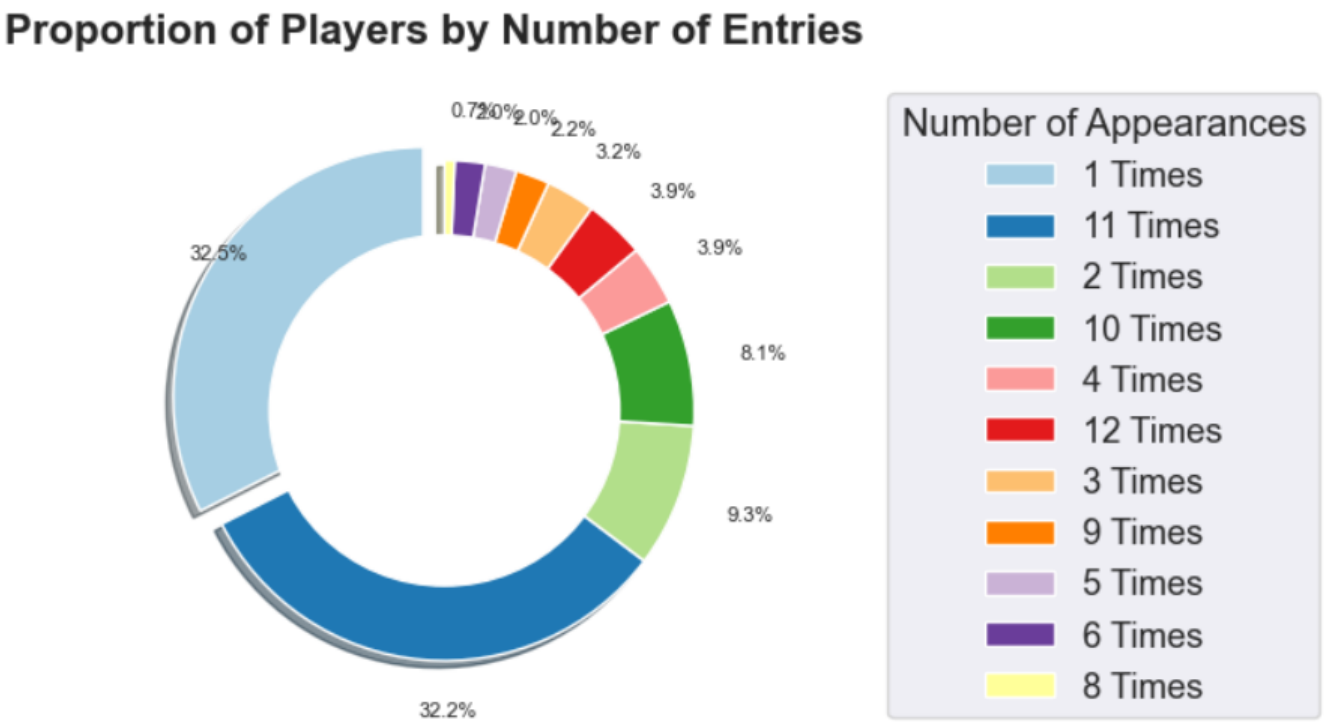
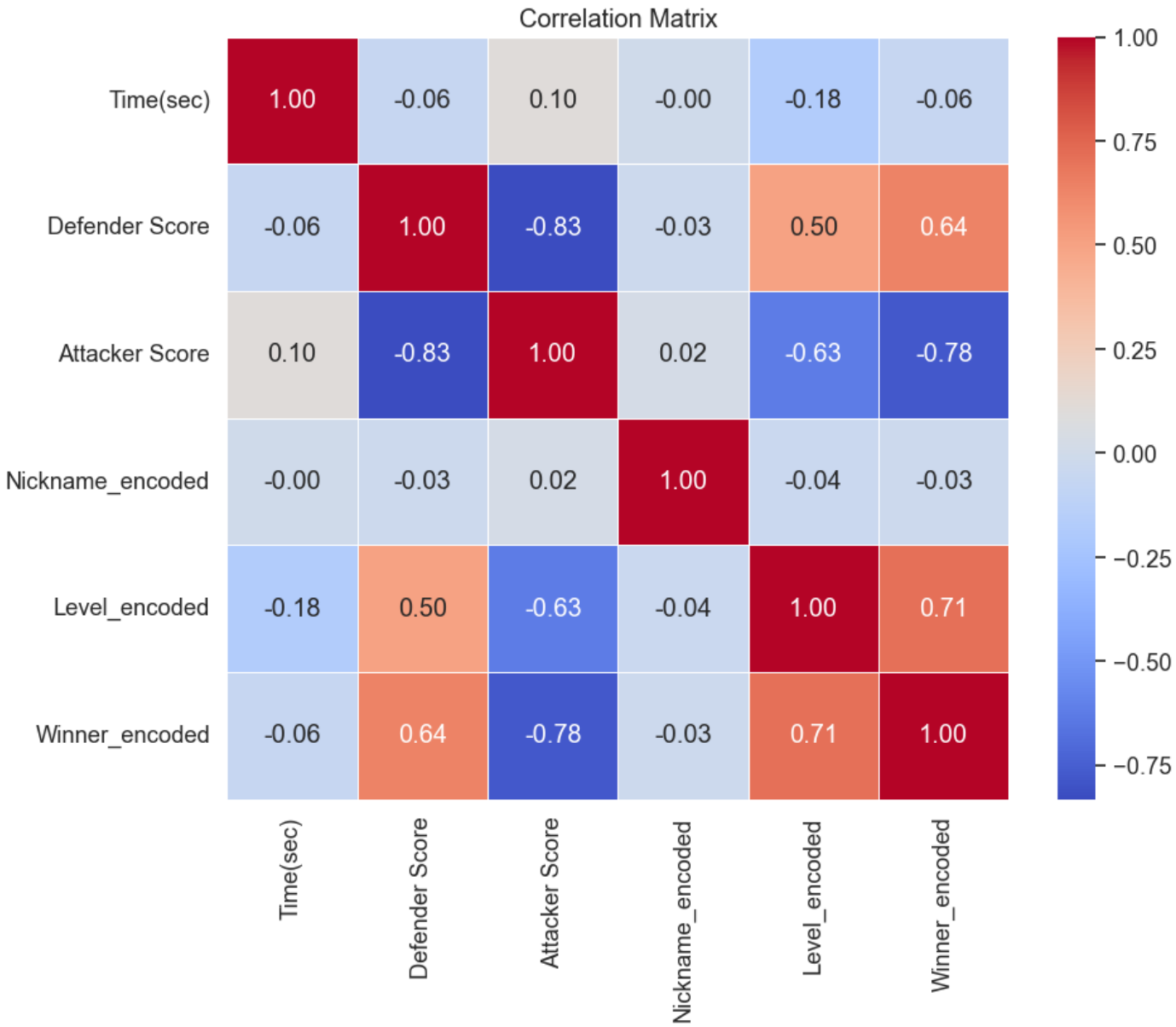


Figure 12: Number of Entries

The precision and recall values along with the confusion matrix were computed and displayed. The precision measures how many of the instances predicted as positive are actually positive. This means that a high precision score indicates accuracy in identifying positive cases. The recall value measures how many actual positive instances are identified correctly by the classifier therefore a high recall value indicates that the model actually identified most of the positive cases. The F1 score is another important indicator which combines the precision and recall outputs to balance them. This value can be viewed in the python notebook provided. The confusion matrix is another performance evaluation model which illustrates the classification of each class provided which is an easier way to visually identify errors and the performance of the models. Due to the size of the dataset, the models did not take much time to run and process.

Correlation Matrix



From the correlation matrix it can be observed that time is not highly correlated with the other variables and the nickname even less so however, the winner and level are highly correlated and the most correlated variables are the defender and attacker scores.

SVM

Accuracy of SVM : 0.984

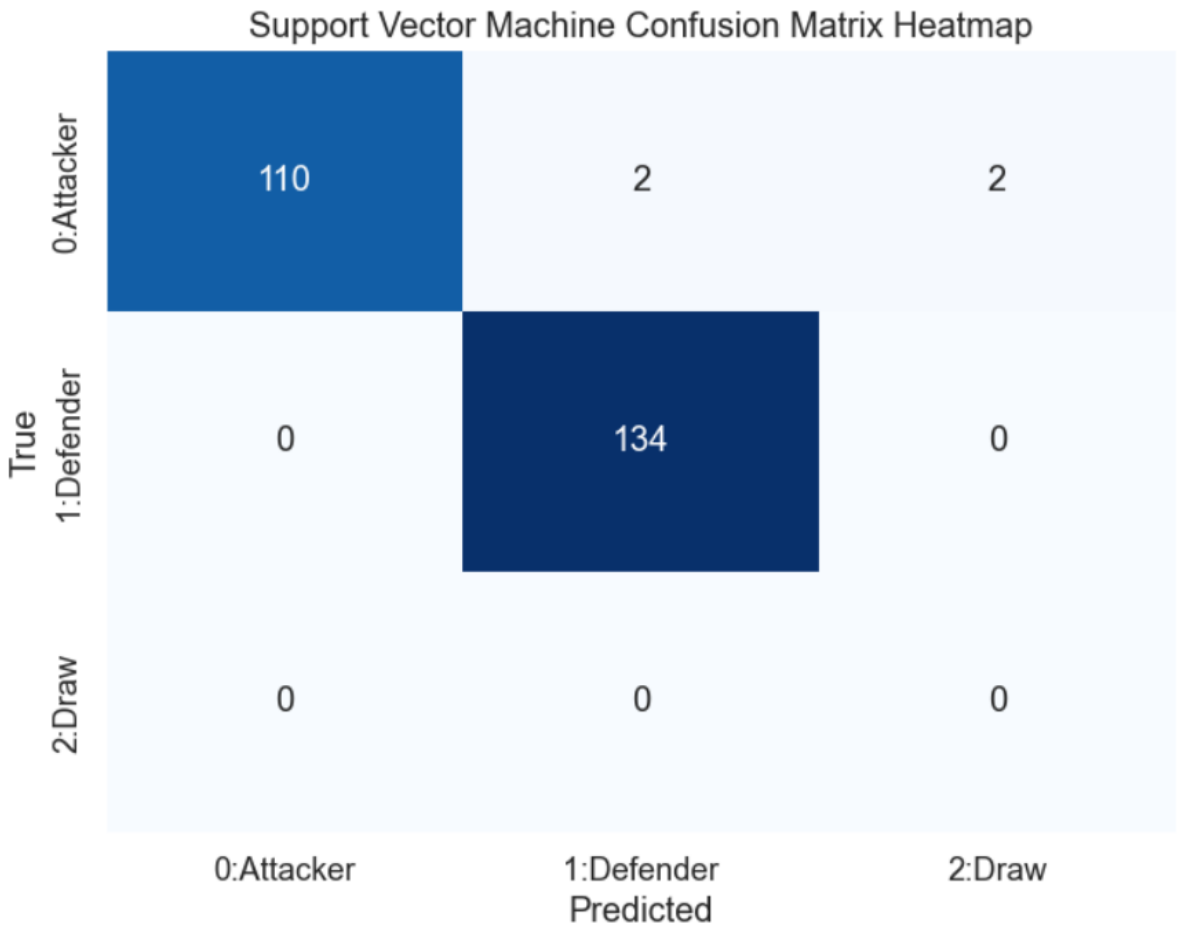


Figure 13: SVM Confusion Matrix Heatmap

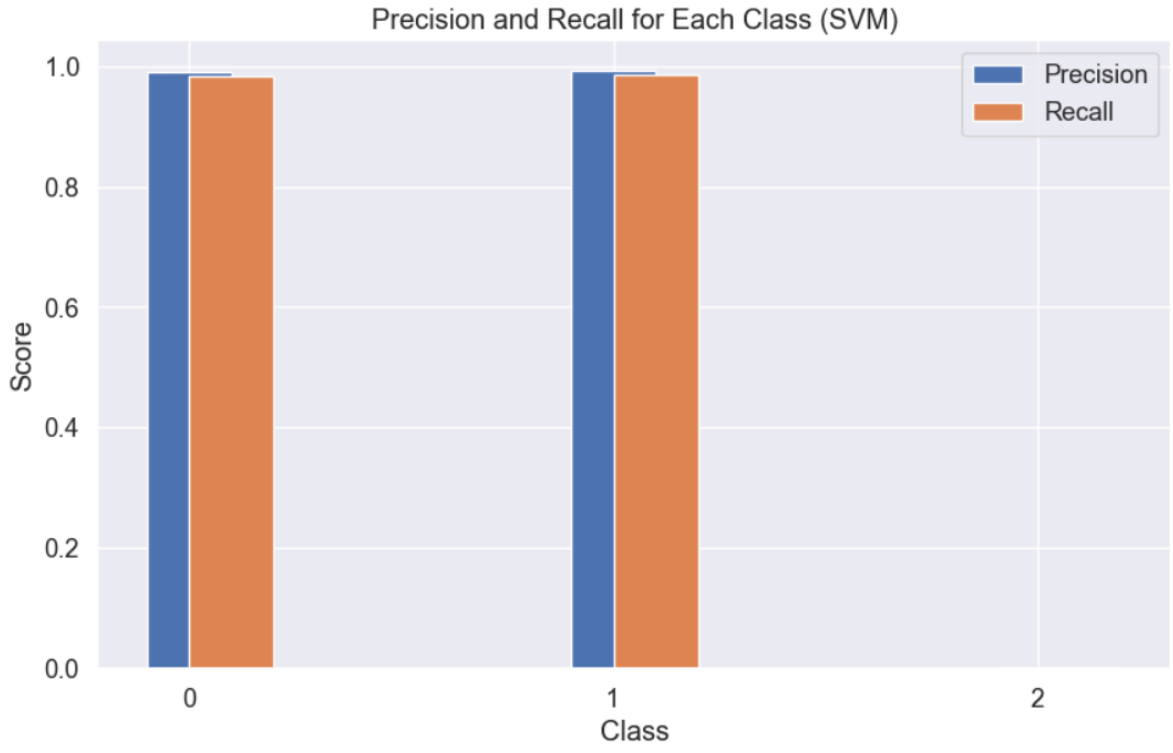


Figure 14: SVM Precision and Recall

Random Forest

Accuracy of Random Forest : 1.0

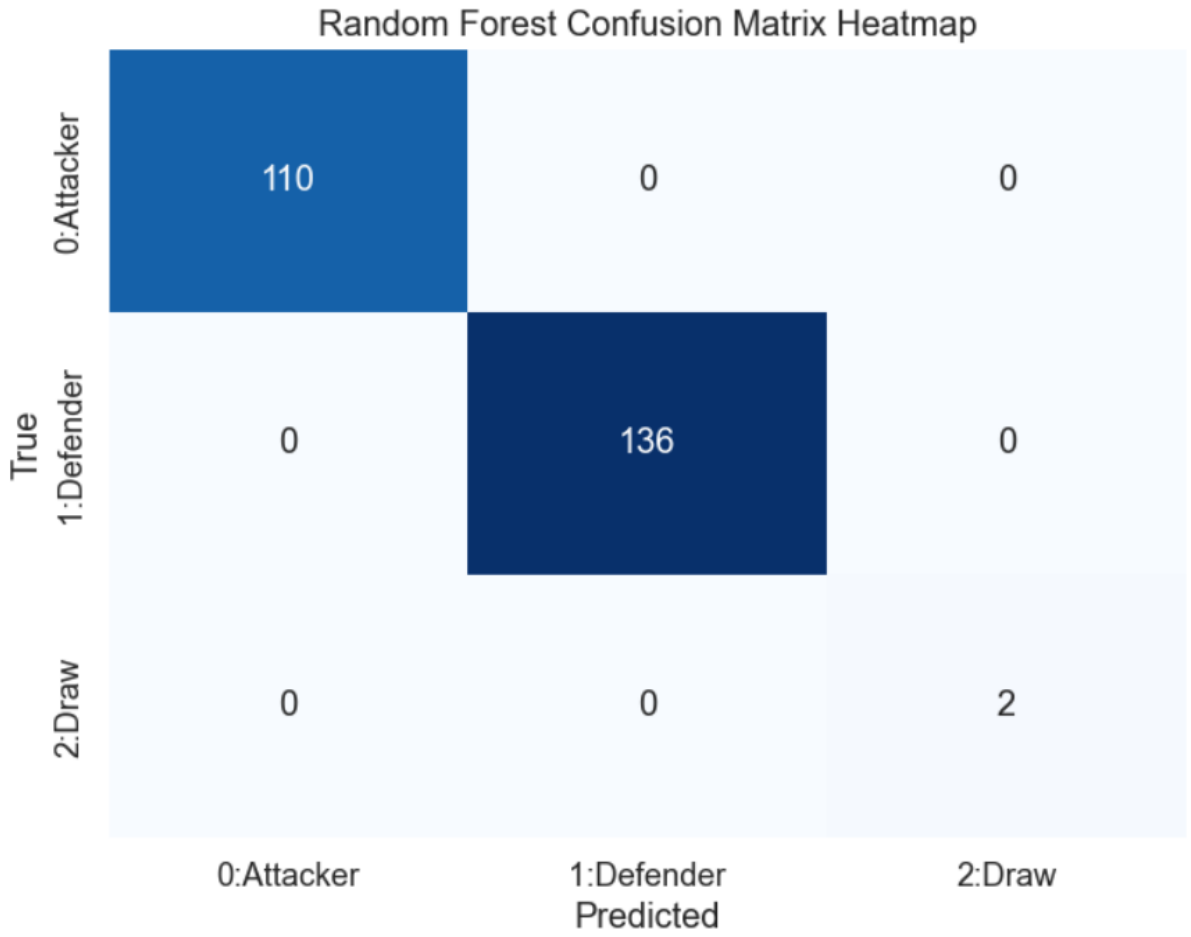


Figure 15: Random Forest Confusion Matrix Heatmap

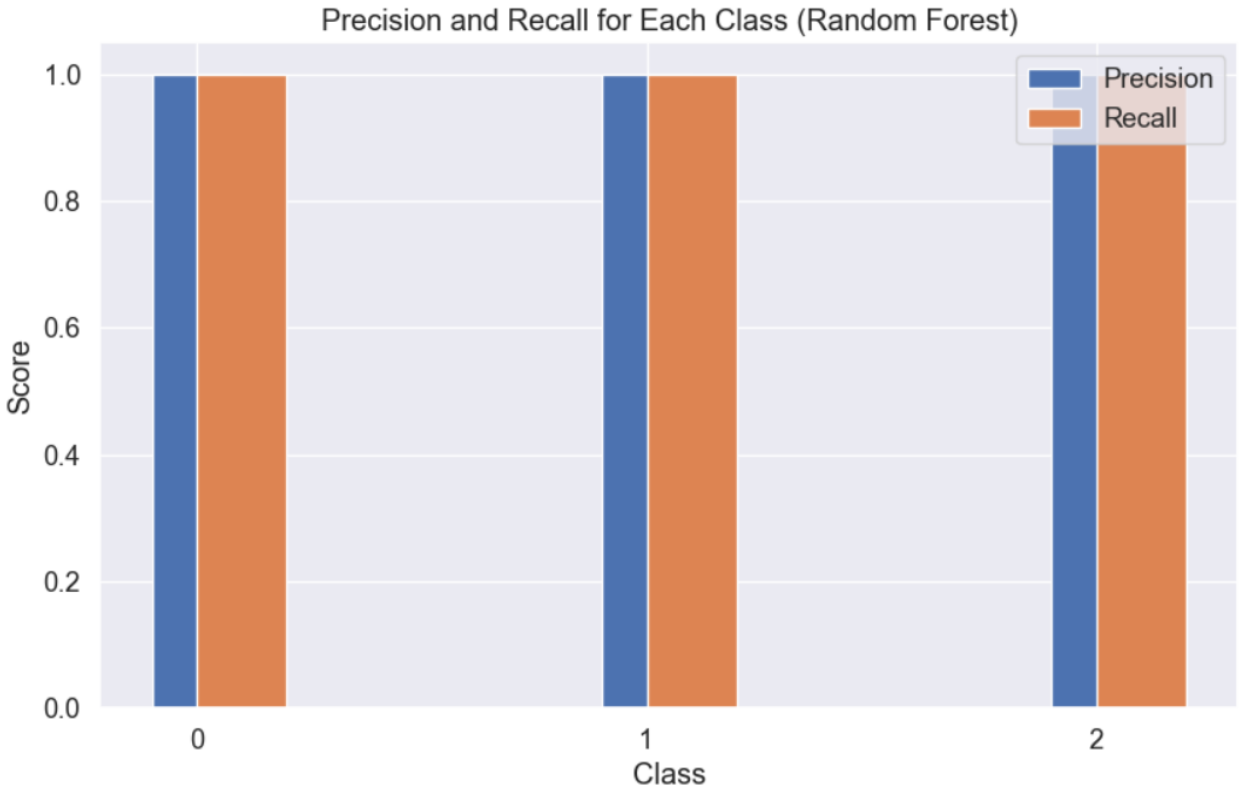


Figure 16: Random Forest Precision and Recall

Naïve Bayes

Accuracy of Naive Bayes : 0.944

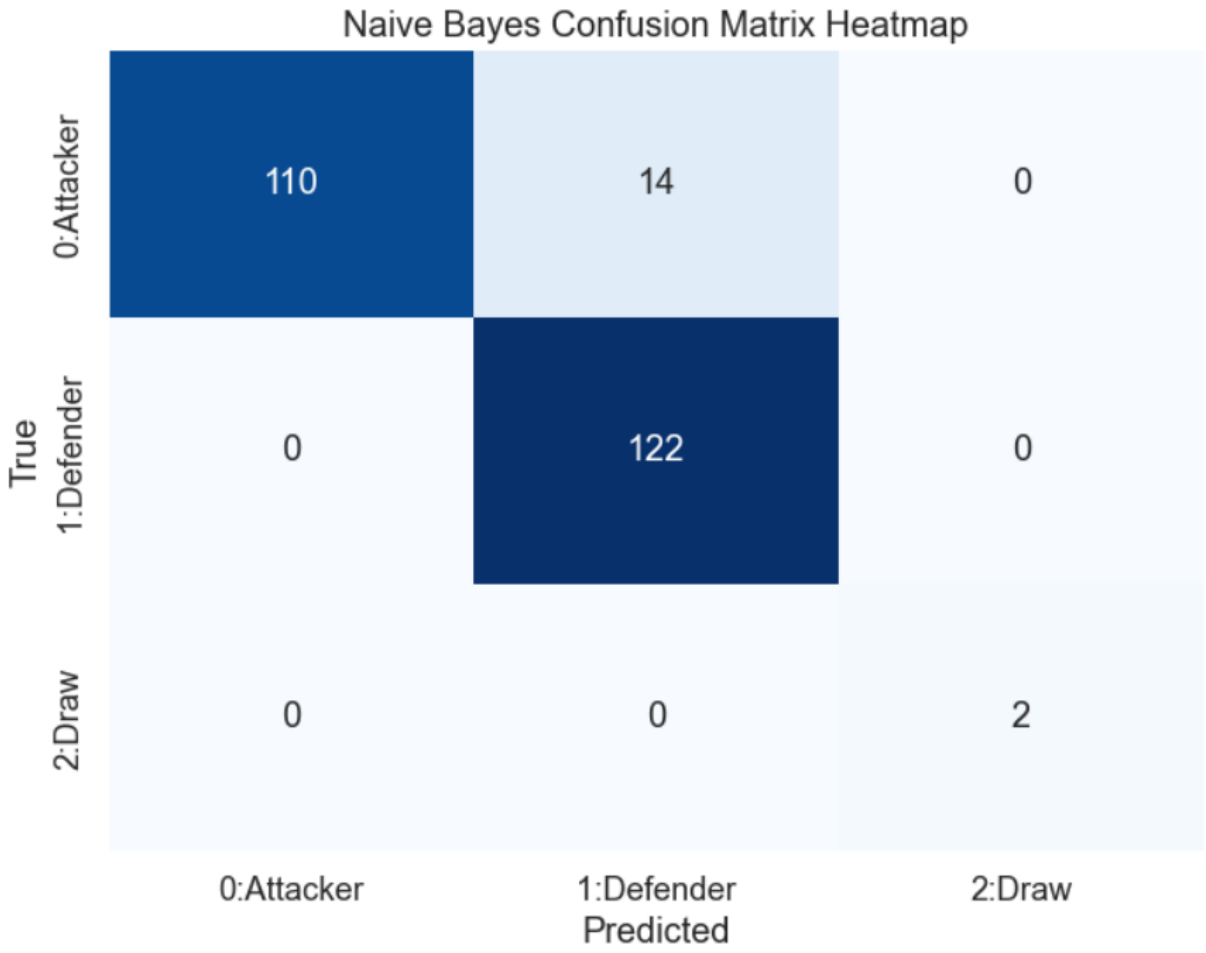


Figure 17: Naive Bayes Confusion Matrix Heatmap

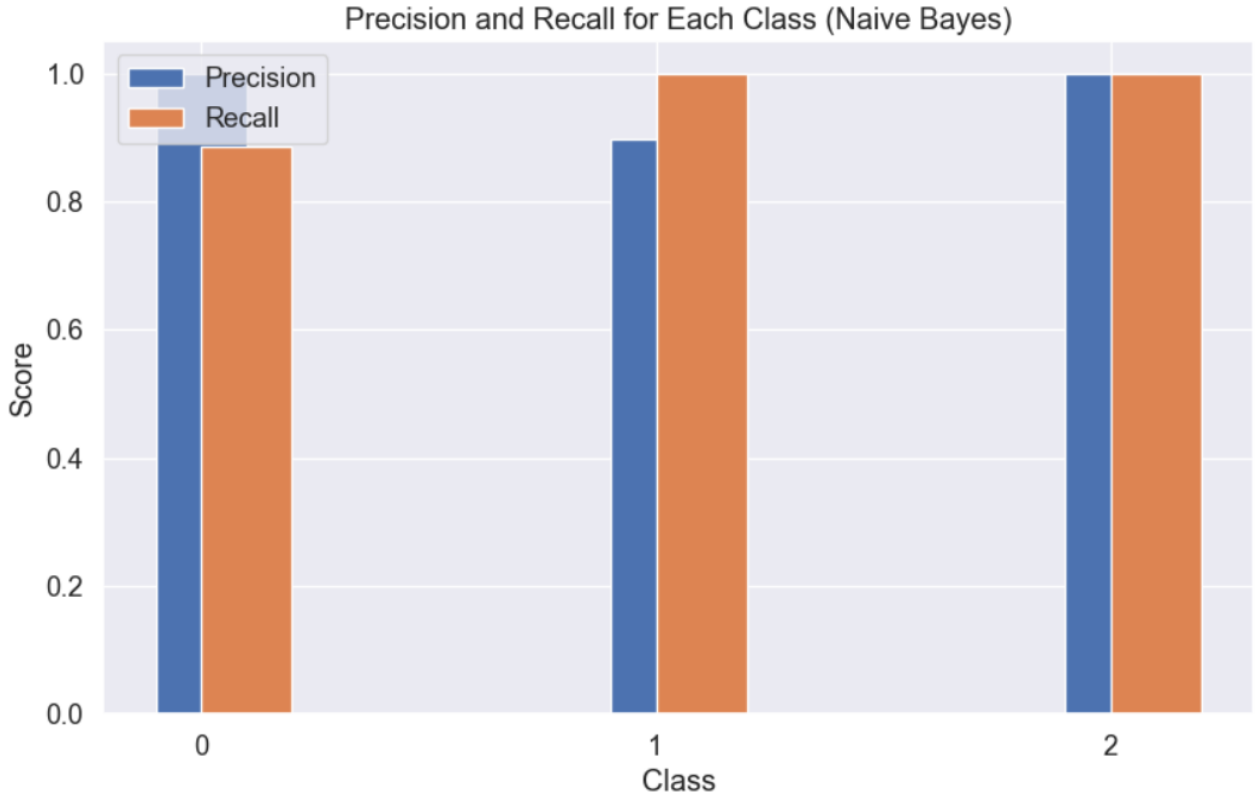
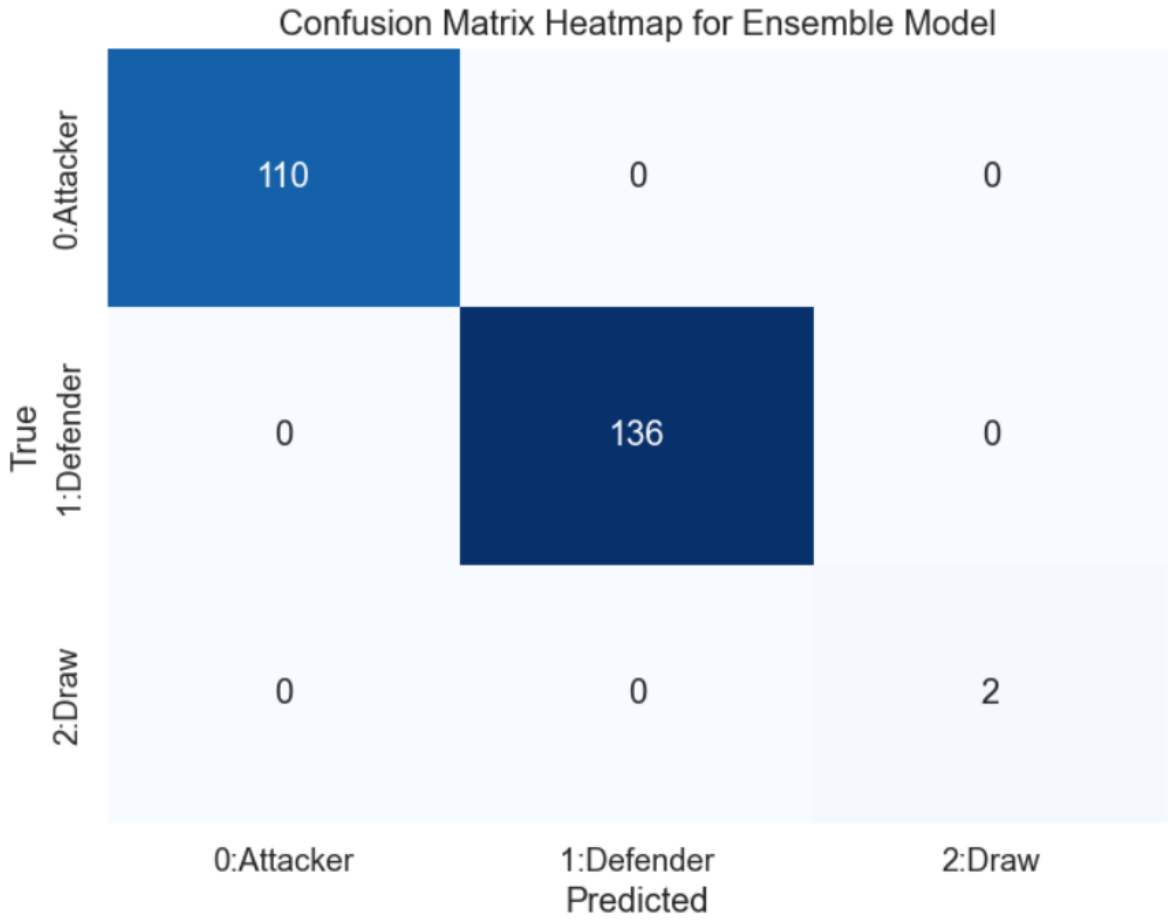


Figure 18: Naive Bayes Precision and Recall

Ensemble model

Accuracy of Ensemble Model : 1.0



ROC (Receiver Operating Characteristic)

Some modifications had to be made for the ROC model as it only handles binary classifications. Due to the fact that there were three classes it did not work for the traditional ROC therefore a multiclass ROC model had to be developed

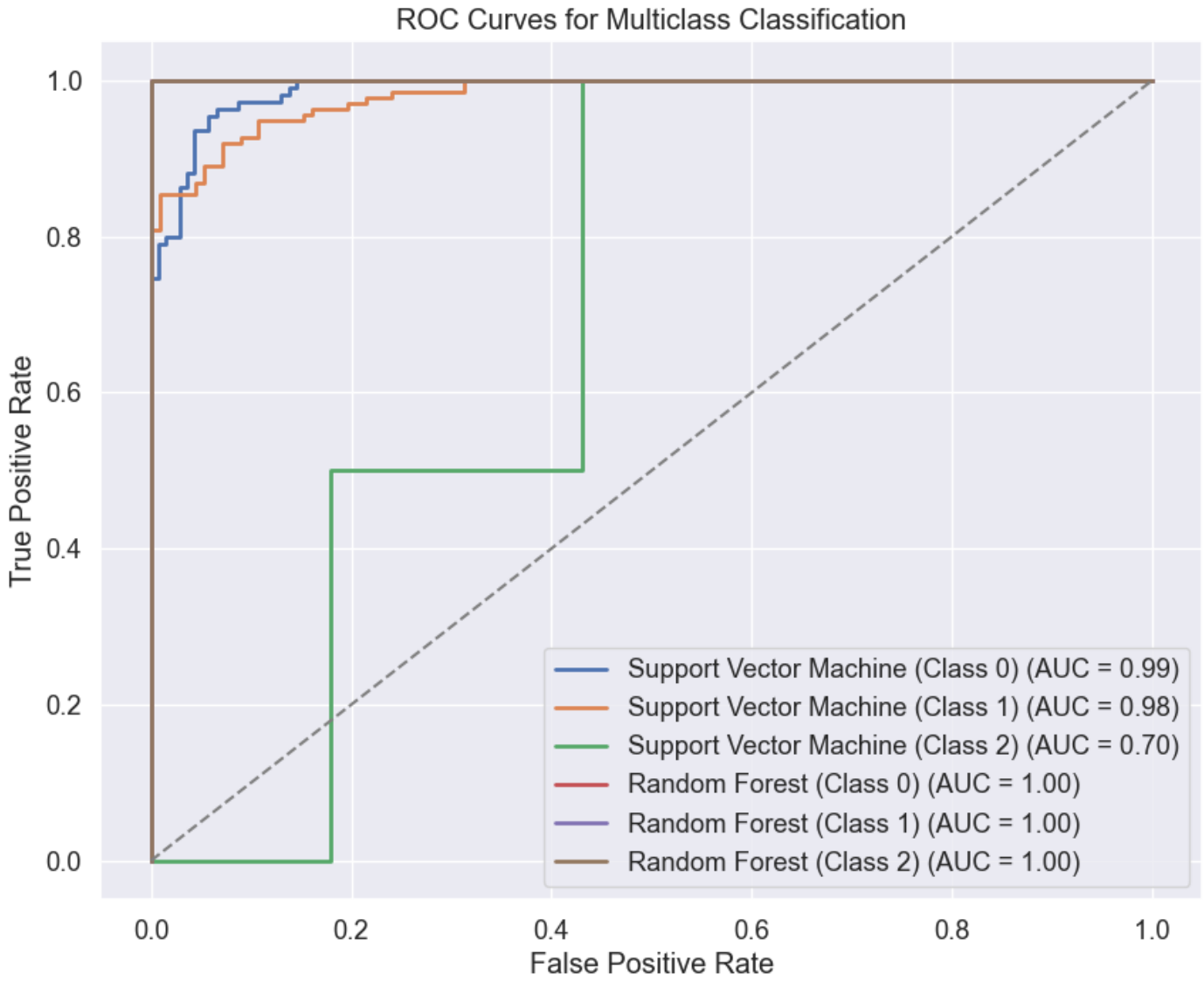


Figure 19: ROC model

The ROC curves show that the Random Forest model outperforms the Support Vector Machine (SVM) model across all classes. For Classes 0 and 1, the SVM does a great job, with AUC scores close to 1, meaning it's really good at telling these classes apart. However, it struggles with Class 2, scoring just 0.70, indicating it's not as effective in distinguishing this class. On the other hand, the Random Forest model is spot-on for every class, scoring a perfect 1.0 across the board. Its curves hug the top left corner of the graph, showing it has almost no false positives and catches all the true positives, making it the clear winner in this comparison. The Naïve Bayes did not appear as it was not able to produce probability estimates.

Lazy Predict

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
AdaBoostClassifier	1.00	1.00	None	1.00	0.18
BaggingClassifier	1.00	1.00	None	1.00	0.03
BernoulliNB	1.00	1.00	None	1.00	0.01
DecisionTreeClassifier	1.00	1.00	None	1.00	0.01
ExtraTreeClassifier	1.00	1.00	None	1.00	0.01
ExtraTreesClassifier	1.00	1.00	None	1.00	0.10
RandomForestClassifier	1.00	1.00	None	1.00	0.13
GaussianNB	0.93	0.96	None	0.93	0.01
NearestCentroid	0.81	0.87	None	0.89	0.01
XGBClassifier	1.00	0.83	None	1.00	0.14
SGDClassifier	1.00	0.83	None	1.00	0.02
LGBMClassifier	1.00	0.83	None	1.00	0.24
PassiveAggressiveClassifier	0.99	0.67	None	0.99	0.02
Perceptron	0.99	0.67	None	0.99	0.01
CalibratedClassifierCV	0.99	0.67	None	0.99	0.10
LinearSVC	0.99	0.67	None	0.99	0.02
LogisticRegression	0.99	0.66	None	0.98	0.02
KNeighborsClassifier	0.99	0.66	None	0.98	0.03
SVC	0.99	0.66	None	0.98	0.03
LinearDiscriminantAnalysis	0.96	0.64	None	0.95	0.03
RidgeClassifier	0.96	0.64	None	0.95	0.02
RidgeClassifierCV	0.96	0.64	None	0.95	0.01
DummyClassifier	0.55	0.33	None	0.39	0.01

Of all the models, the Random Forest took the longest time to run, followed by the SVM and Naïve Bayes. This output indicates that the accuracy of the models in order of most to least accurate is Random Forest, Naïve Bayes and SVM.

Discussion

From the beginning with the distribution graphs, there was some skewness that could be observed in the Attacker Score, Defender Score and Time data, this was reduced to be able to fit the models and was done through transformations and normalising the data to get a more normal distribution.

The results show that based on the time and scores as well as the level of the players, the winner outcome can be easily determined. The best performing model was the random forest. It is know as one of the best machine learning algorithms so this comes as no surprise. The other models are not far behind as they are all high scoring in the precision and recall statistics.

There was also not a very significant difference between the defender and attacker scores as well as the winner outcome between the defender and attacker as seen in Figure 9: Winner Count, which means that a large number of players do not have a lot of cyber security awareness as was discussed (Alotaibi *et al.*, 2016).

The game balance is largely fair though as majority leans towards the Defender winning however the Attacker wins are not far behind.

There are several limitations in the study, such as the fact that nicknames are not set and players can change their nickname with every game they play, meaning that there is no consistency in checking the progression of players because they may be counted as a separate person. The Naïve Bayes model assumes independence of the features which in this case does not hold true at all. The attacker and defender scores determine the winner outcome meaning there is dependence, and the time as well as the outcome determines the level given.

Conclusion

The winner outcomes were easy to predict due to the simplicity of the game and the dataset given as well as the high correlation values from the correlation matrix.

There is more work to be done in the cybergaming space as well as in the cyber security awareness space. With respect to the cybergaming space, there needs to be more research done to determine the correlation between user statistics and the kind of gamers the users are in order to personalise the gaming experience. Although there has been some research done as shown by ... there has not been very extensive work done to confirm the conclusions of the study. In the cyber security awareness games, although it is an educational tool, it is very important to try and make the games as interesting as possible when it comes to the visual aspect as well as making it challenging enough to enjoy but not too challenging that it feels like too much work and the fun aspect is taken away from it.

The suggestions for this game are to ensure that users have an account to keep track of their data to be certain of the results produced by that account. Having an account can also help track more variables such as timestamps for the time of day and demographic information which can help to personalise the game through game mechanics and user experience. More data means that more analytics can be determined and more improvements can be made. Other areas to improve the game include adding more levels so that players have more of a challenge and something to look forward to as the game can be easily accomplished as there is only one level and the attacks can be predicted. Due to the fact that the purpose of this “CyberVigilance” game is to increase cyber security awareness within the greater population, it would be better for the game to be transferred to a mobile application to make it more accessible so that it reaches more people.

References

Alotaibi, F., Furnell, S., Stengel, I. & Papadaki, M. 2016. A review of using gaming technology for cyber-security awareness. *Int. J. Inf. Secur. Res.(IJISR),* 6(2):660-666.

Bakkes, S.C.J., Spronck, P.H.M. & van Lankveld, G. 2012. Player behavioural modelling for video games. *Entertainment Computing,* 3(3):71-79.

Busch, M., Mattheiss, E., Hochleitner, W., Hochleitner, C., Lankes, M., Fröhlich, P., Orji, R. & Tscheligi, M. 2016. Using Player Type Models for Personalized Game Design – An Empirical Investigation. *Interaction Design and Architecture(s),* 28:145-163.

Gosztonyi, M. 2023. Who are the gamers? Profiling adult gamers using machine learning approaches. *Telematics and Informatics Reports,* 11:100074.

Maurice Hendrix, Ali Al-Sherbaz & Victoria Bloom 2016. Game Based Cyber Security Training: are Serious Games suitable for cyber security training? *International Journal of Serious Games,* 3(1).

Wa Nkongolo, M. 2024. Infusing Morabaraba game design to develop a cybersecurity awareness game (CyberMoraba). Johannesburg.