NORTH SOUTH UNIVERSITY

Identifying Internet Gaming Disorder using Machine Learning Techniques

SUBMITTED TO
DR. SIFAT MOMEN
OF NORTH SOUTH UNIVERSITY
IN THE PARTIAL FULFILMENT OT THE REQUIREMENTS
FOR THE DEGREE OF BACHELOR OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING IN
DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Date
06th May 2022, Friday

Declaration

It is hereby acknowledged that:

- No illegitimate procedure has been practiced during the preparation of this document.
- This document does not contain any previously published material without proper citation.
- This document represents our own accomplishment while being Undergraduate Students in the **North South University**

Sincerely,

Kazi Mosaddequr

Student 1: Kazi Mosaddequr
1831543042, Signature

Samya

Student 2: Samya Sunibir Das
1911563642, Signature

Student 3: Emon Emtiyaz
1813128642, Signature

Abstract

Internet Gaming Disorder (also known as IGD in short) is a psychological disorder common to younger adult generation of the world. In Bangladesh with the increase in internet user, many user may have IGD but remain without diagnosis. The danger of IGD includes negative mental health impact, addiction to gambling and many others. Machine learning model for identifying Internet Gaming Disorder is proposed this paper. The data was received from Mendeley as well as collected from the students of Bangladesh. The Data set then underwent processing to make it suitable for machine learning. Exploratory data analysis was done to see the nature of the data set. Feature extraction was done to find the important features. Finally, model was created and accuracy was measured to find the best model for the proposed problem.

Contents

List of Figures

1 Introduction

Internet Gaming Disorder (also known as IGD in short) is a psychological disorder common to younger adult generation of the world. With the improvement in availability of internet and games, it has now became the emerging diagnosis[1]. In this paper, we are proposing a machine learning model to predict the IGD status of the internet users (mainly the university going students of Bangladesh).

In Bangladesh with the increase in internet user, many user may have IGD but remain without diagnosis. The danger of IGD includes negative mental health impact, addiction to gambling and many others[2]. Therefore, it is important to identify if a person has an IGD or not.

Our goal is to create an efficient machine learning model to predict if a person has IGD or not by using some common available feature such as the user's age, gender, sleeping hours, etc.

We have taken a data set from a similar work done in Colombo, Sri Lanka from Mendeley [3]. Exploratory data analysis was done on the data set along with feature relevance. Then relevant feature was put into a survey and spread among University students to collect data. Around 50 responses were received. Model was trained based on Sri Lankan data set and its performance was evaluated on Bangladeshi data base. The hyper-parameters of the model was tuned to obtain optimal results.

We have searched in Google scholars and other places and could not find this kind of work done for Bangladesh. For comparison of the model, we have implemented Zero-R classification, One-R Classification, Decision Tree, K Nearest Neighbour Classification, Random forest and Logistic Regression.

Rest of the paper is organized as follows. Section 2 discusses relevant works. Section 3 sheds light on the process and steps taken to collect data and create model. We discuss our results in Section 4 and state how our model has met the expectation. We conclude in Section 5 and provide some future directions.

2 Literature Review

A study on Internet Gaming Disorder (IGD) among Sri Lankan school students was conducted by M Manchayanke, [3] there was a database with 395 records formed from the survey. The database primarily assessed IGD among the survey participants using the IGDS9-SF scale. [4] Binary Logistic Regression was then applied in this database and predicted IGD of individuals with an accuracy score of 94.7%. The same database was used to train our model.

Song, Kun-Ru [5] and his associates applied a modified machine learning algorithm, which is Connectome-based Predictive Modeling (CPM) combined with a Support Vector Machine (SVM), on a dataset consisting resting-state fMRI data of 72 individuals with IGD and 41 healthy individuals. Their analysis proved that the brain's default mode network (DMN) was the most informative network in predicting IGD in patients both in terms of classification and regression. The accuracy score for classification was 78.76% while for regression correspondence between predicted and actual psychometric scale score: $r = 0.44$, $P < 0.001$.

Another similar research work was conducted by Zi-Liang Wang [6] and his associates where they used a machine-learning method called multi-voxel pattern analysis (MVPA) to see if neural characteristics could be used to predict IGD status and treatment result for IGD. Their dataset consisted of cue-reactivity fMRI-task data where 40 male subjects had IGD and 19 male subjects were healthy. They classified IGD and healthy participants with 92.37% accuracy.

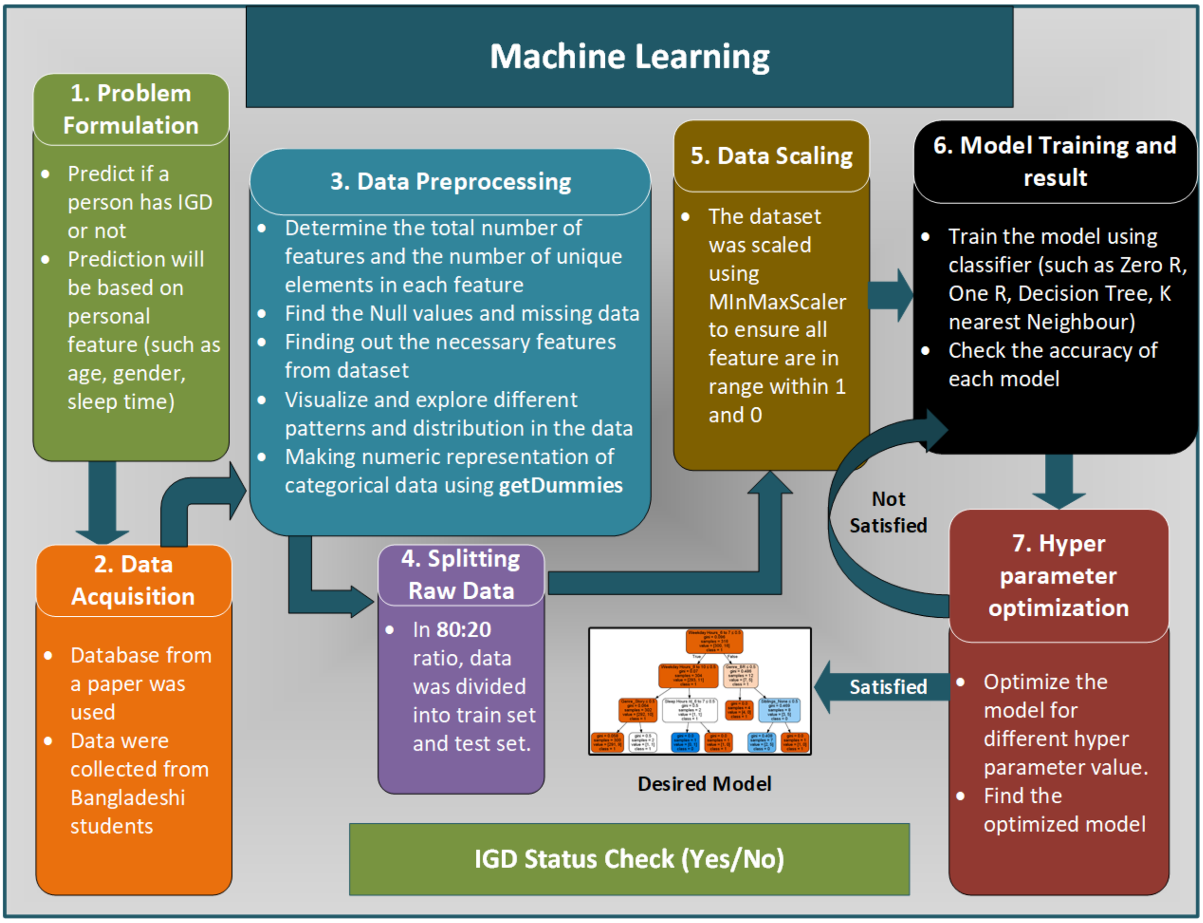
An accuracy score of 93.18% was achieved in predicting IGD in paper [7], where they used supervised learning algorithms. The specific algorithm that helped them get their highest accuracy was Logistic Regression. In that paper, the method for predicting was based around if a player of PlayerUnknown's Battlegrounds (PUBG, a MOBA game) suffers from IGD and other related psychological disorders using game and player information as well as a self-esteem measure. One of their key findings was game statistics of players showed a strong positive correlation with the disorders.

Xu Han, and colleagues [8] carried a research work involved using a radiomics-based machine-learning approach, they discovered differences in brain morphology between IGD participants and healthy candidates, which might aid in the uncovering of underlying IGD-related neurobiological mechanisms.They constructed random forest classifiers and evaluated based on the identified features. The mean classification accuracy score that they obtained was 73%.

Jiheyon Ha et al.[9] classified gamers using multiple physiological signals to contribute to the treatment and prevention of IGD. Participants were divided into three different groups. They performed the classification task using a 2-layers feedforward neural network. They fused the physiological signals and got their best results. Their highest accuracy score was 0.90 and highest F1 score was 0.93.

3 Methodology

The approach adopted in this work is outlined in Figure 3.1:



1. Problem Formulation

- Predict if a person has IGD or not
- Prediction will be based on personal feature (such as age, gender, sleep time)

2. Data Acquisition

- Database from a paper was used
- Data were collected from Bangladeshi students

3. Data Preprocessing

- Determine the total number of features and the number of unique elements in each feature
- Find the Null values and missing data
- Finding out the necessary features from dataset
- Visualize and explore different patterns and distribution in the data
- Making numeric representation of categorical data using **getDummies**

4. Splitting Raw Data

- In **80:20** ratio, data was divided into train set and test set.

5. Data Scaling

- The dataset was scaled using MInMaxScaler to ensure all feature are in range within 1 and 0

6. Model Training and result

- Train the model using classifier (such as Zero R, One R, Decision Tree, K nearest Neighbour)
- Check the accuracy of each model

7. Hyper parameter optimization

- Optimize the model for different hyper parameter value.
- Find the optimized model

Figure 3.1: Flow Chart of Research Methodology

3.1 Data Acquisition

We collected data from a similar work done in Colombo, Sri Lanka from Mendeley [3]. The data set consists of 92 features of students and used them to predict if a student has IGD or not. Similar data set for Bangladesh was searched but it was not available. So we decided to collect them from University going students through surveys.

The data set [3] consisted of 92 features and it was difficult to put all the 92 features in the survey questions. The relevant and important features among those 92 features were analysed. The features were narrowed down to 17 features (shown in Table 3.1). We wanted to include income in the survey, However, in the initial phase when we included income, many subjects were not interested to fill it out and avoided our survey. Once we removed it. we were able to obtain around 50 responces.

Table 3.1: Description of selected features

Variable Name	Description
Age	The age of the subject
Medium	The medium of education the subject received
Sex	The gender of the subject
Siblings	The number of siblings of the subject
Ethnicity	The ethnicity of the subject
Sports	What kind of sportsman person is the subject
Sleep Hours /d	How long does subject sleep each day in average
SQ Category	How well subject sleeps
Appearance	How the subject feels about his/her appearance
Self-esteem	How the subject feels about his/her self-esteem or confidence
Game Type	what type of game does the subject play
Weekday Hours	The weekday hours of gaming
Weekend Hours	The weekend hours of gamin
Genre	The genre of game played by the subject
Device	The device used by the subject to play games
Start Age	The age in which subject started gaming
Friends Cat.	How many friends do the subject have?
IGD Status	The internet gaming disorder status of the subject (Yes or No)

3.2 Data Pre-processing

In data Pre-processing, we remove the null values descibes in Section 3.2.1. Then we do exploratory data analysis in Section 3.2.2. The data in data set is then encoded discussed in Section 3.2.3.

3.2.1 Removing Null Values

In the data Pre-processing stage, we remove the null values present in the featuers. Two approaches were taken. On the one hand, we removed the column with null value. On the other hand, we remove the records containing null vlalues. We chose the data set with the second approach, Although we lost some data, we were left with some important feature that were needed for the model. An instance of cleaned data set is shown in Table3.2.

Table 3.2: First five row of the cleaned data set

	Age	Medium	Sex	Siblings	Ethnicity	Sports	Sleep Hours /d	SQ Category	Appearance	Self-esteem	Game Type	Weekday Hours	Weekend Hours	Genre	Device	Start Age	Friends Cat.	IGD Status
0	18	Sinhala	M	Two	Sinhala	Moderate	4 to 5	Average	Average	4	Online Multiplayer	6 to 7	4 to 5	BR	PC	8 to 10	Many	Y
1	18	Sinhala	M	Two	Sinhala	Moderate	4 to 5	Average	Average	3	Online Multiplayer	1 or less	2 to 3	BR	Mobile	14 to 16	Many	N
2	18	Sinhala	M	Two	Sinhala	Minor	6 to 7	Good	Good	4	Online Multiplayer	4 to 5	4 to 5	Non BR Shooter	PC	11 to 13	Moderate	N
3	18	Sinhala	M	One	Sinhala	Minor	6 to 7	Good	Good	3	Online Multiplayer	2 to 3	2 to 3	BR	Mobile	14 to 16	Many	N
4	18	Sinhala	M	None	Sinhala	Moderate	4 to 5	Average	Average	5	Offline Single player	1 or less	1 or less	BR	Mobile	14 to 16	Few	N

3.2.2 Exploratory Data Analysis

By examining the pre-processed data, several trends in Sri lankan's Internet gaming disorder and their life styles are found. Some of the count plots for the categorical data are given below:

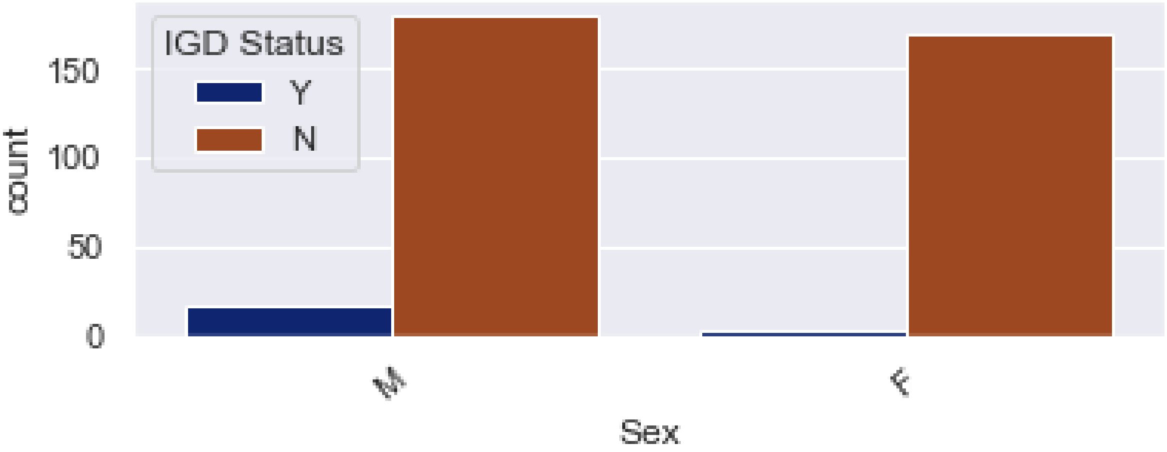


Figure 3.2: Count plot for Sex

Male show more sign of Internet Gaming Disorder than female.

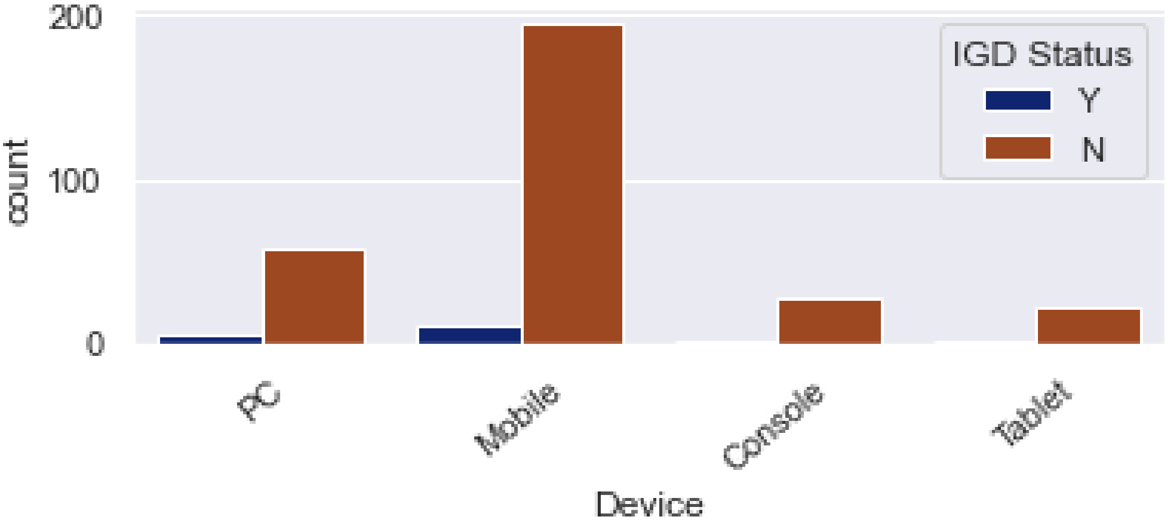


Figure 3.3: Count plot for Device

Subjects who play games on mobile device has higher chance of IGD.

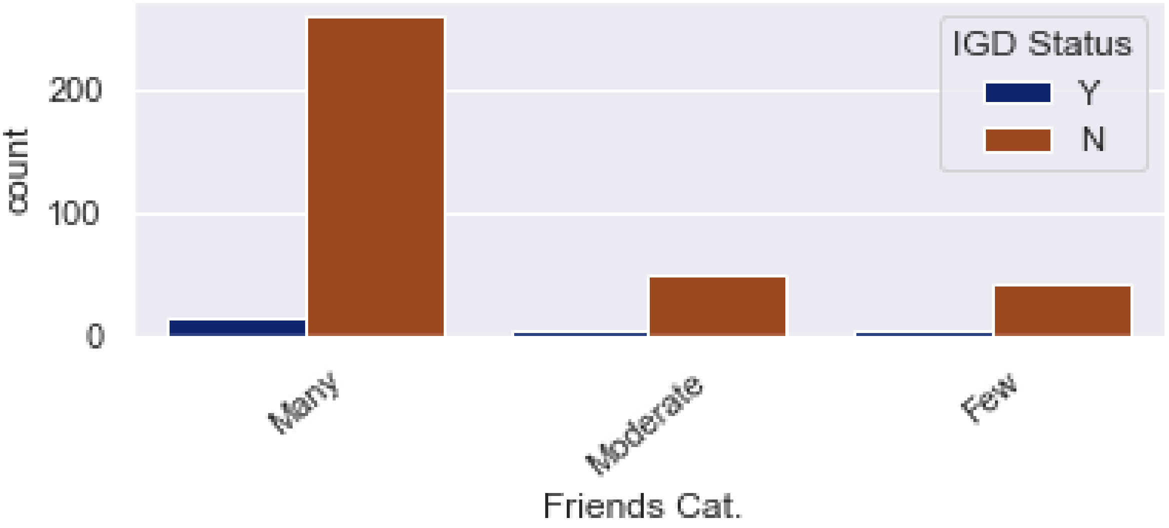


Figure 3.4: Count plot for Friends

We also found out that people with large number of friends tends to have IGD.

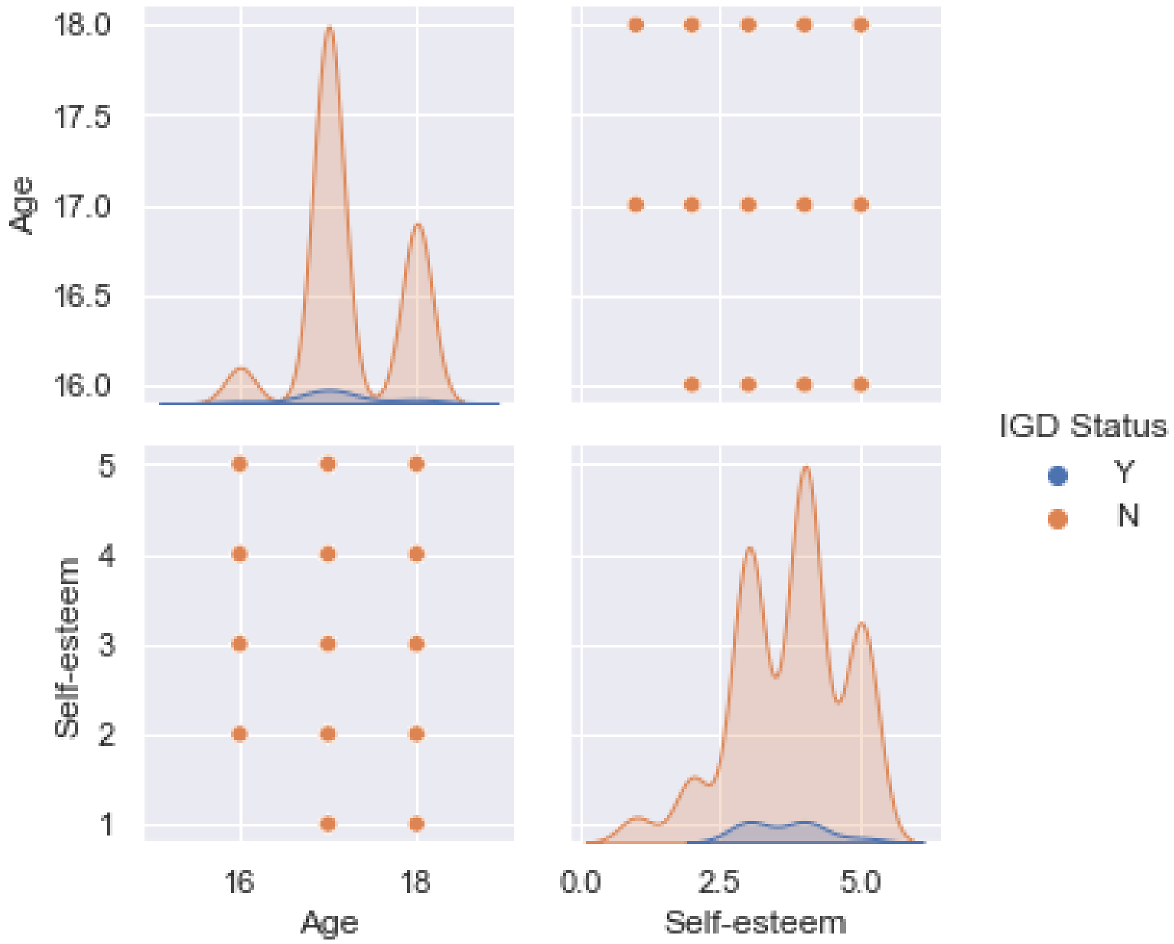


Figure 3.5: Pair plot for continuous data

By analyzing the pair plots, one can determine pair-wise correlation among the features. From the pair plot of age and Self-esteem we can see that they are not related that much. Lower age and self esteem results in Internet Gaming Disorder.

3.2.3 Data Encoding

The data inside our data set has categorical value. Some classifier requires the data to be numerical in order to create a model. To be able to train our model, we created a numerical representation of these categorical values using getDummies function.

3.3 Splitting the data

The Data is then split ted in to 80:20 ratio. Where 80% of the data is used as train set and 20% of the data is used as test set. Here we also find the correlation of features and drop if the correlation of any feature is greater than 90%. The code used for correlation dropping is given below:

```python
def correlation(dataset, threshold):
col_corr = set() # Set of all the names of the redundant columns
corr_matrix = dataset.corr()
for i in range(len(corr_matrix.columns)):
    for j in range(i):
        if(abs(corr_matrix.iloc[i, j])) > threshold:
            colname = corr_matrix.columns[i]
            col_corr.add(colname)
return col_corr
```

Listing 3.1: Python example

For visualization part, heat-map plotting was used to see which feature had greater correlation (as shown in Figure 3.6).

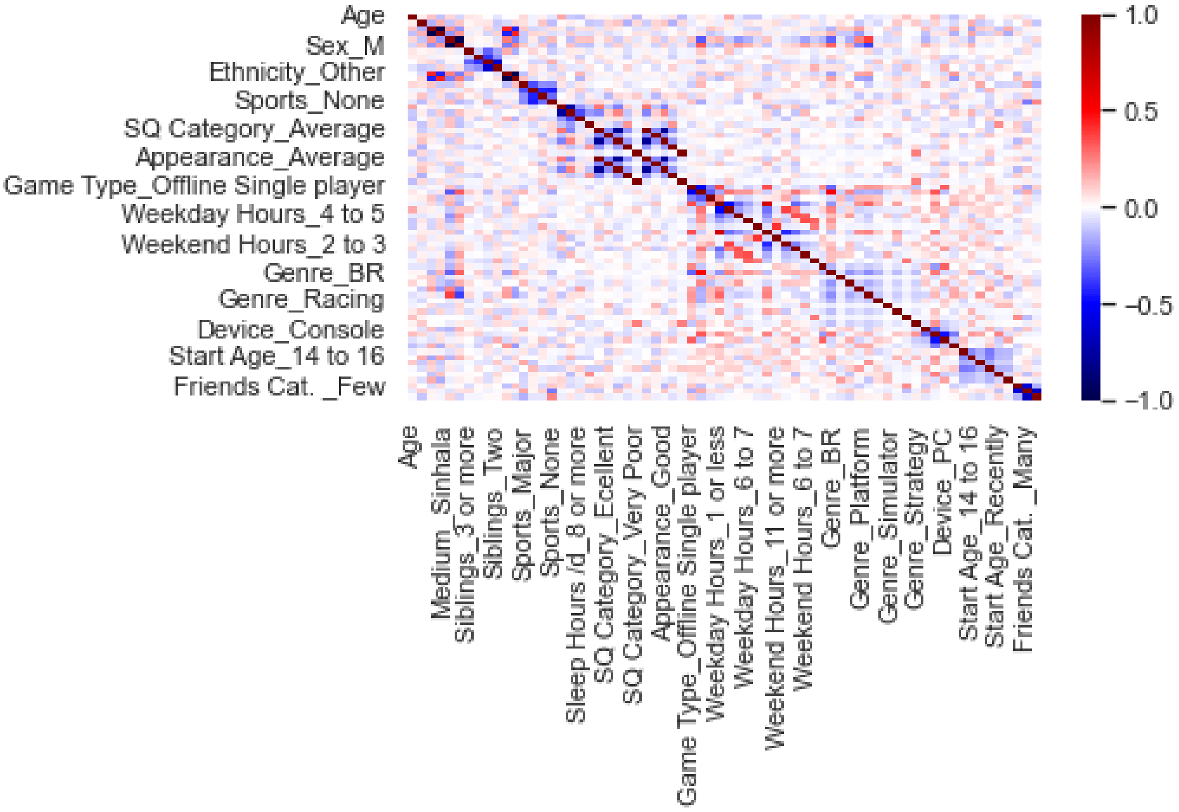


Figure 3.6: heat-map of correlation

3.4 Data Scaling

Some of the classifier such as K Nearest Neighbour uses Euclidean distance to measure the distance between two points. If the data are not scaled, then some feature may dominate other features. In order to avoid this situation, it is important to scale our dataset. We used the MinMaxScaler function from Scikit-learn library [10] to scale our dataset. By doing so, all features were transformed into the range [0, 1] meaning that the minimum and maximum value of a feature is going to be 0 and 1, respectively. The formula 3.1 for scaling is as follows:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.1}$$

3.5 Model Training

The data set was then trained using classifiers. The classifier used were

1. Zero-R classifier
2. One-R classifier
3. Decision Tree
4. K Nearest Neighbour
5. Random Forest
6. Logistic Regression

Models were trained and their accuracy score was checked. Then we did hyper parameter training for some classifier discussed in Section 3.6

3.6 Hyper parameter optimization

For Decision tree and K Nearest Neighbour we have done hyper parameter training. we ran a loop and plotted the accuracy for different hyper parameter values.

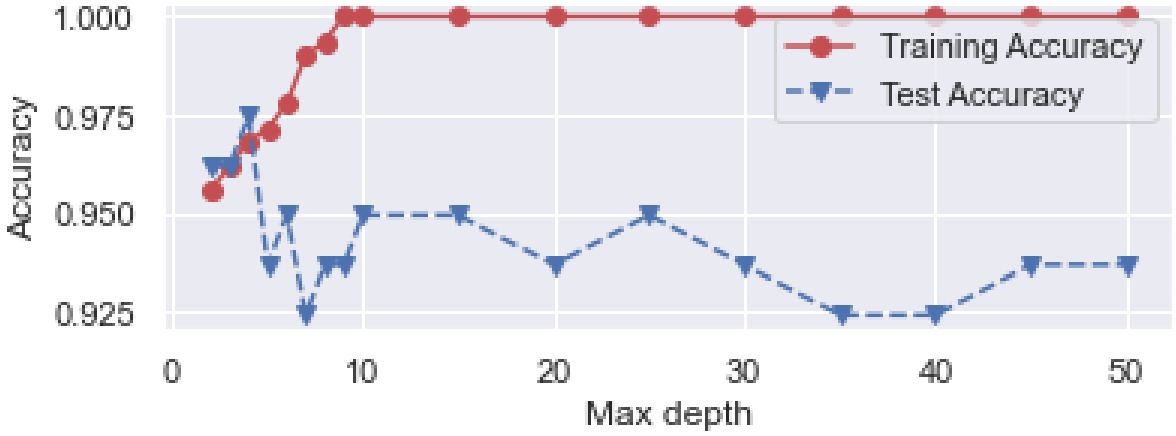


Figure 3.7: hyper parameter analysis Decision Tree

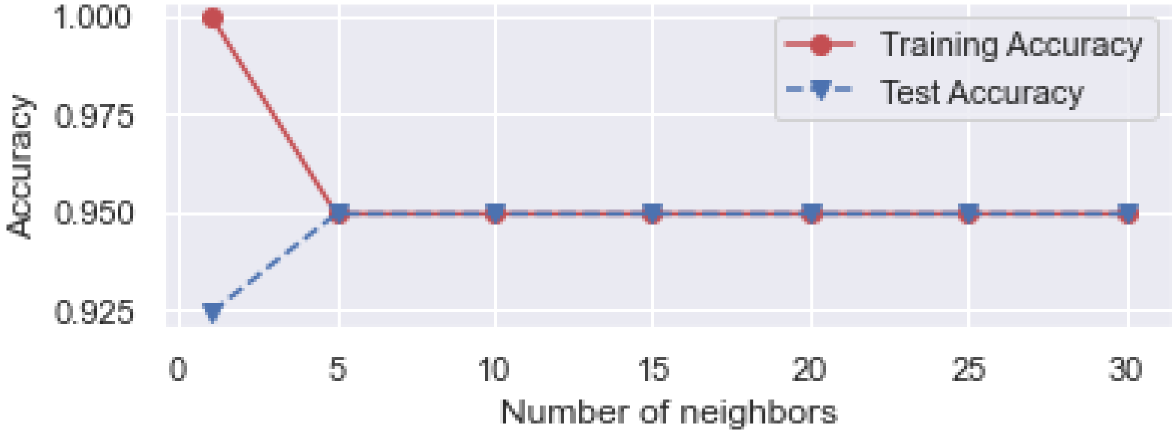


Figure 3.8: hyper parameter analysis KNN

From Figure 3.7 we choose 20 as maximum depth of Decision Tree and from Figure 3.8 we chose the value 5 as the value of k.

The model is then recreated by using above mentioned values.

4 Results and Discussions

The data set was trained with the aforementioned six classifiers. The result was plotted to confusion matrix (Figure 4.1) for easy understanding. To determine which classifiers perform the best, they were assessed against two performance indicators. The two metrics are accuracy (Table 4.1) and f1 score (Table 4.2).

Table 4.1: Accuracy on training and test data set

Classifier	Train Accuracy	Test Accuracy
Zero-R	93.54%	93.26%
One-R	93.54%	93.26%
Decision Tree	96.35%	93.25%
KNN	93.54%	93.25%
Better Decision Tree	100.00%	87.64%
Better KNN	93.53%	93.25%
Random Forest	100.00%	93.25%
Logistic Regression	95.22%	91.01%

From the Table 4.1 we can see that all the models were able to achieve accuracy greater than 94%. Decision Tree and Better Decision Tree give best accuracy.

Table 4.2: F1 score on training and test data set

	F1 score	
Classifier	Train	Test
Zero-R	0.00	0.00
One-R	0.00	0.00
Decision Tree	0.62	0.25
KNN	0.00	0.00
Better Decision Tree	1.00	0.26
Better KNN	0.00	0.00

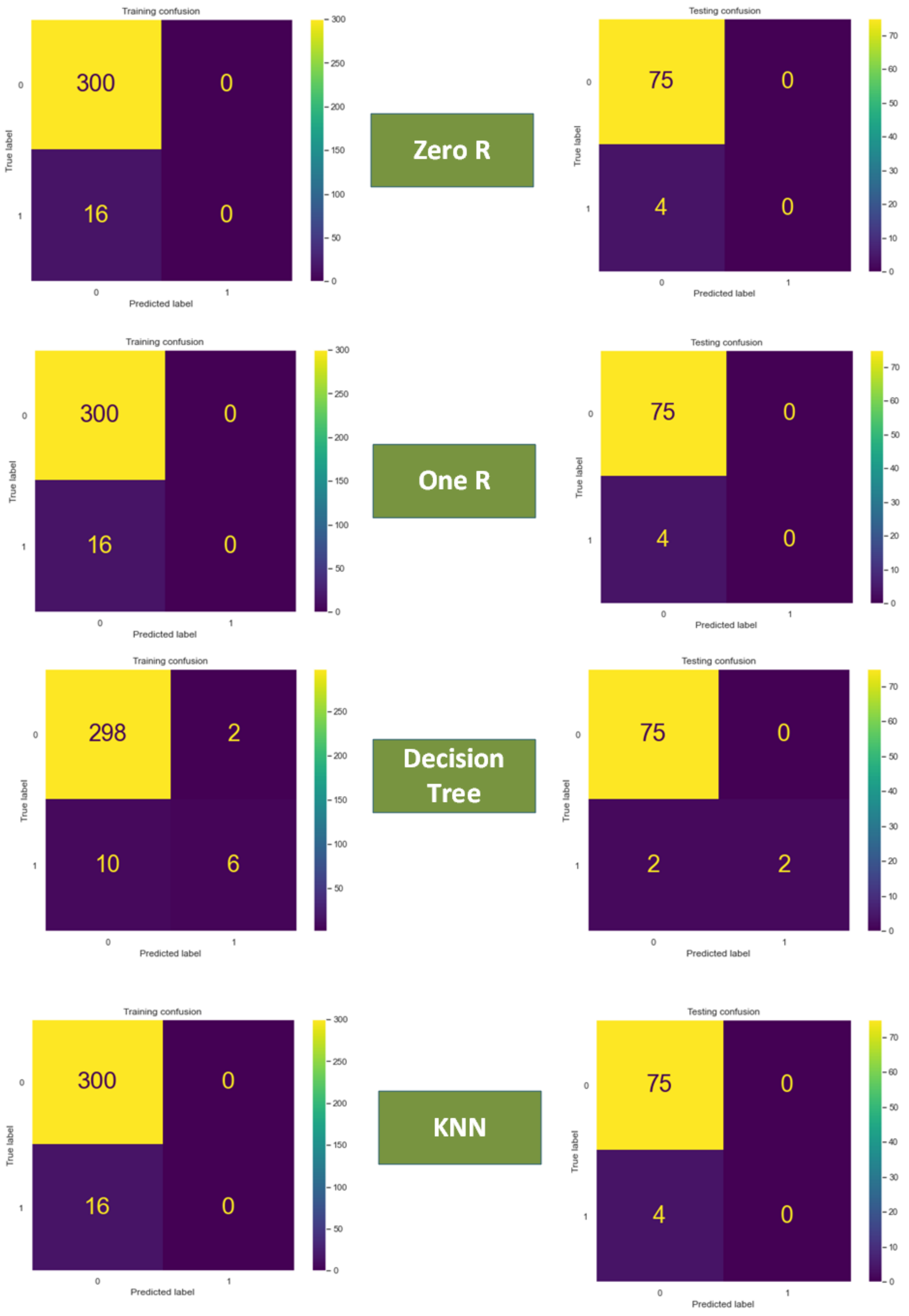


Figure 4.1: confusion matrix

Finally, comparison was done with all the classifier and was plotted in the graph (Figure 4.2).

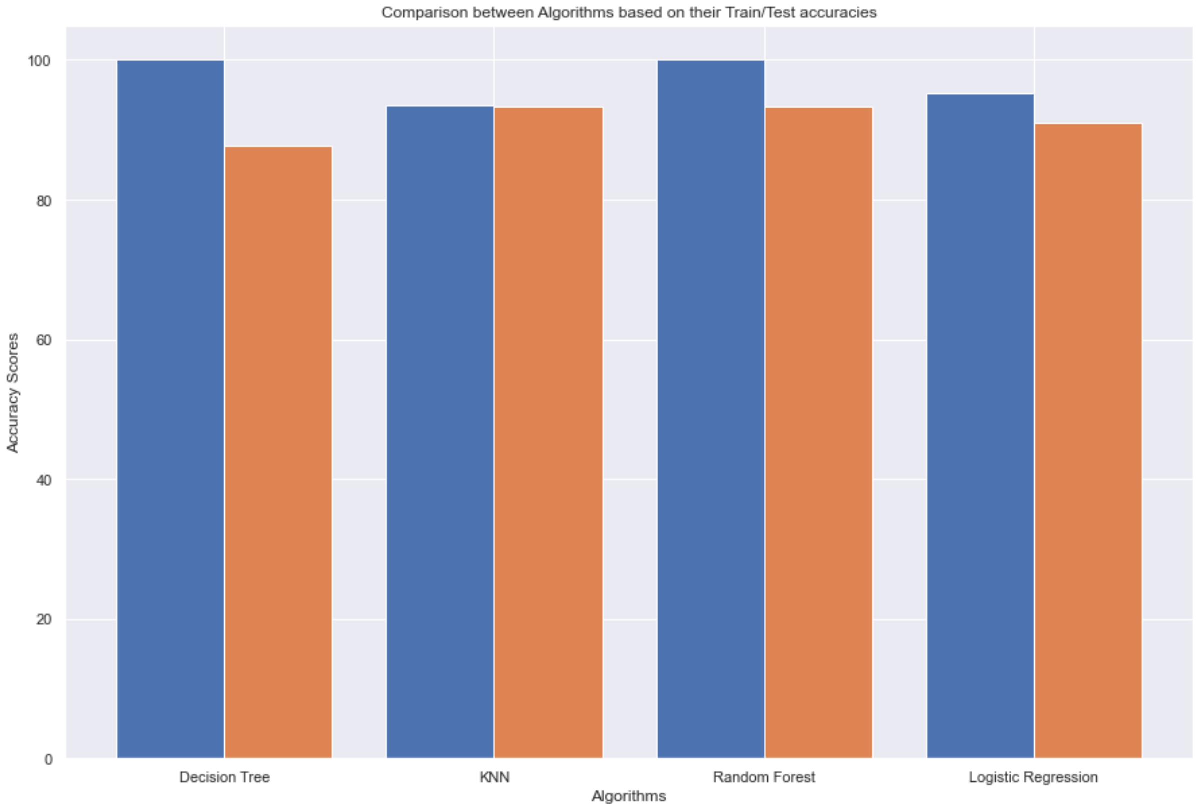


Figure 4.2: Comparison between Algorithms based on their Train/Test accuracies

From the figure above 4.2, we can see that Random Forest Classifier has the best accuracy. We chose and propose Random Forest Classifier to be used as model for the evaluation of the Internet Gaming Disorder.

5 Conclusion and Future Work

This article discusses a machine learning approach for finding Internet Gaming Disorder Status in context of Bangladesh. In the beginning we started with 92 features and finally found 14 features to be useful. Since the data acquired was voluntary, some of the feature needed to be discarded. However, if we could give the subjects some incentives and get those data then we would get better model. In future, we would like to work on making the model more better. We would also like to deploy the model on web application so that user can check themselves without going and spending money on doctors.

References

[1] MD Alison Yarp. Internet gaming disorder: Symptoms, diagnosis, treatment, Oct 2021.

[2] Daniel L. King and Paul H. Delfabbro. The concept of "harm" in internet gaming disorder. *Journal of Behavioral Addictions*, 7(3):562 – 564, 2018.

[3] M Manchanayake. Data for: Internet gaming disorder: An emerging public health concern among an advanced level student population from colombo, sri lanka. *Mendeley Data*, 2021.

[4] Halley M Pontes and Mark D Griffiths. Measuring dsm-5 internet gaming disorder: Development and validation of a short psychometric scale. *Computers in human behavior*, 45:137–143, 2015.

[5] Kun-Ru Song, Marc N Potenza, Xiao-Yi Fang, Gao-Lang Gong, Yuan-Wei Yao, Zi-Liang Wang, Lu Liu, Shan-Shan Ma, Cui-Cui Xia, Jing Lan, et al. Resting-state connectome-based support-vector-machine predictive modeling of internet gaming disorder. *Addiction Biology*, 26(4):e12969, 2021.

[6] Zi-Liang Wang, Marc N. Potenza, Kun-Ru Song, Xiao-Yi Fang, Lu Liu, Shan-Shan Ma, Cui-Cui Xia, Jing Lan, Yuan-Wei Yao, and Jin-Tao Zhang. Neural classification of internet gaming disorder and prediction of treatment response using a cue-reactivity fmri task in young men. *Journal of Psychiatric Research*, 145:309–316, 2022.

[7] Swati Aggarwal, Shivin Saluja, Varshika Gambhir, Shubhi Gupta, and Simrat Pal Singh Satia. Predicting likelihood of psychological disorders in playerunknown's battlegrounds (pubg) players from asian countries using supervised machine learning. *Addictive Behaviors*, 101:106132, 2020.

[8] Xu Han, Lei Wei, Yawen Sun, Ying Hu, Yao Wang, Weina Ding, Zhe Wang, Wenqing Jiang, He Wang, and Yan Zhou. Mri-based radiomic machine-learning model may accurately distinguish between subjects with internet gaming disorder and healthy

controls. *Brain Sciences*, 12(1), 2022.

[9] Jihyeon Ha, Sangin Park, Chang-Hwan Im, and Laehyun Kim. Classification of gamers using multiple physiological signals: Distinguishing features of internet gaming disorder. *Frontiers in Psychology*, 12, 2021.

[10] Ekaba Bisong. Introduction to scikit-learn. In *Building machine learning and deep learning models on Google cloud platform*, pages 215–229. Springer, 2019.