THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Generalized Anxiety Disorder

Presented by:

Aasish Kumar

Greeshmanjali

Mahtab Barkhordarian

Sagar Tripathi

Today's Discussion

- Introduction
- Dataset Description
- Data Preprocessing
- Model Development
- Performance Measure
- Results

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Introduction

- The focus of this project is to predict, and screen generalized anxiety disorder in people who play online games. The reason for selecting this dataset is most of the youth, educated, not educated as many are playing games so much time and it is becoming like their hobby. The cost is not only the money, but also their most precious time and health.

- It will be demonstrated using three machine learning algorithms: Decision Tree Classifier, Random forest Classifier, Support Vector Classifier.

- Developed a GUI based application to display the end-to-end modelling.

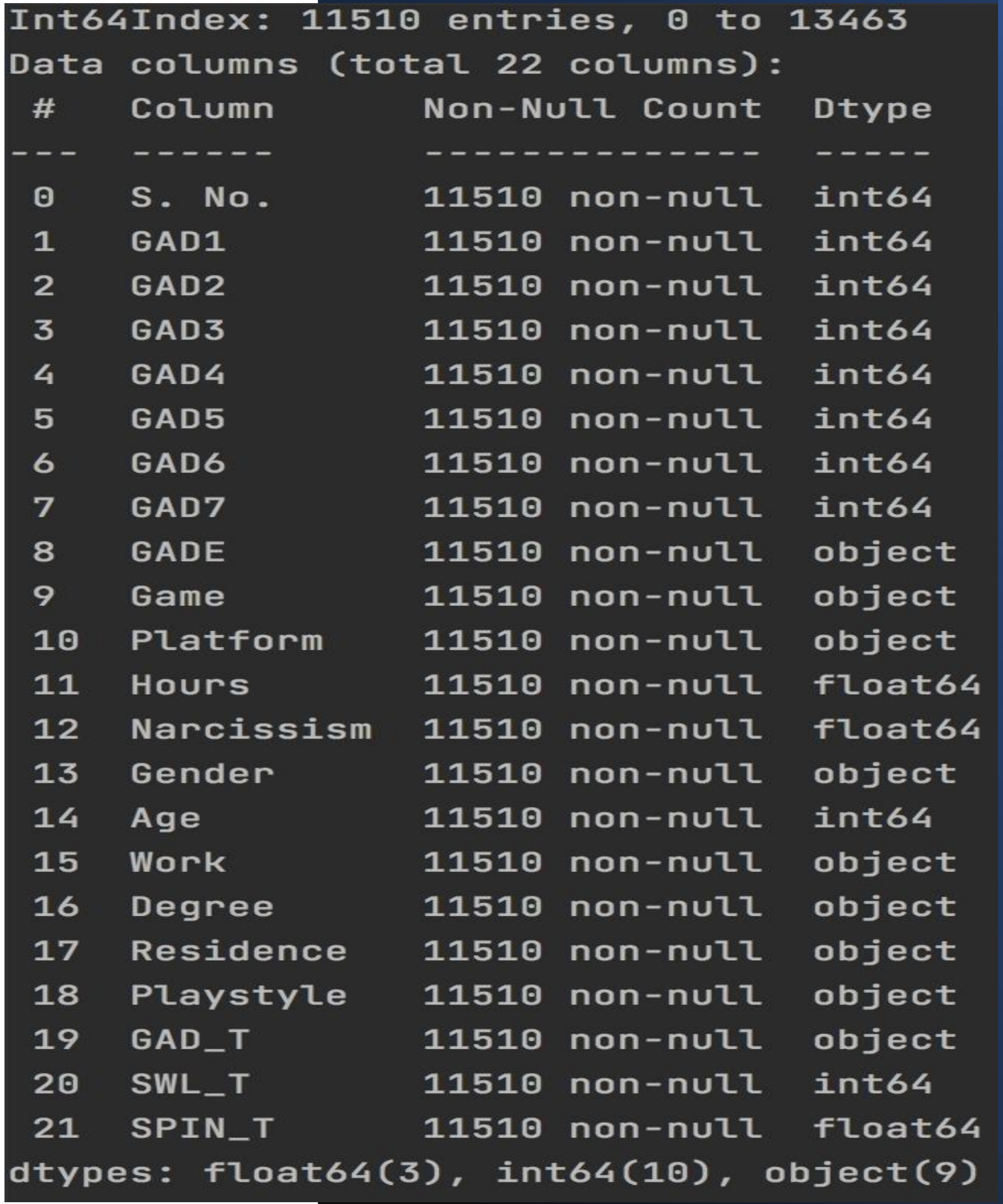
Dataset Description

- The dataset used is sourced from Kaggle .

  Source: https://www.kaggle.com/divyansh22/online-gaming-anxiety-data
- The dataset has 14250 observations and 55 features, most features are categorical (Nominal, Ordinal, Binary) and some with high cardinality.
- Packages used: matpoltlib-NumPy-sklearn-pandas-seaborn-math-PyQt5.
- We will measure the effect of gaming on different platforms with regards to Generalized Anxiety Disorder screener.(GAD_T)
- The performance of the models will be calculated using metrics like Accuracy score, Confusion matrix, Classification report, recall score, f1 score.

Data Preprocessing

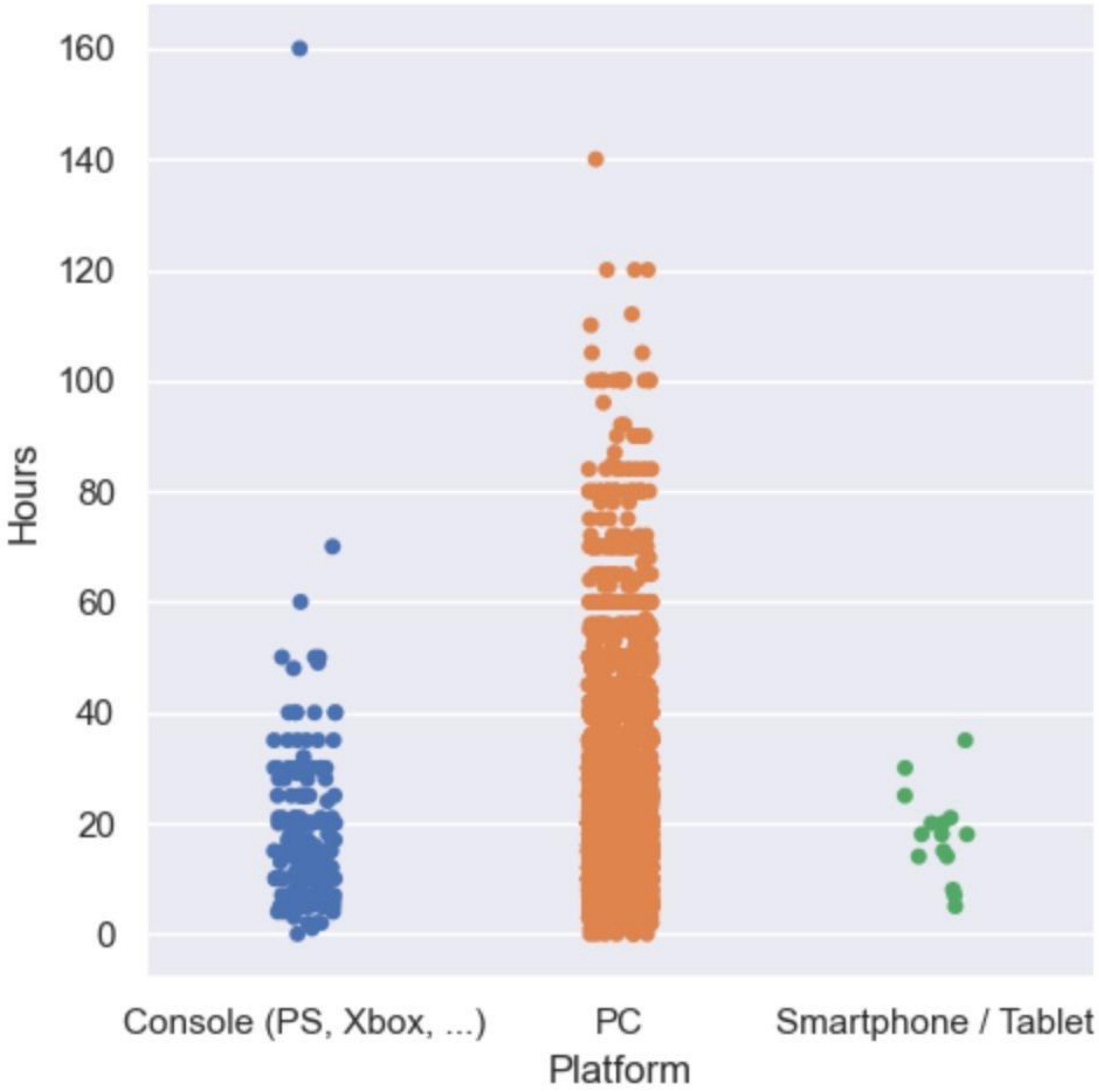
- Features with null values are updated with maximum value count of their respective columns.
- Standardization of data using z-score to remove outliers.
- Encoding is done to decide in a better way on how these labels must be operated and labels are converted into numeric format.
- Label encoder is applied to the target variable as our use case being the classification problem.
- Label binarize is applied to the target variable as it is a multi-class.



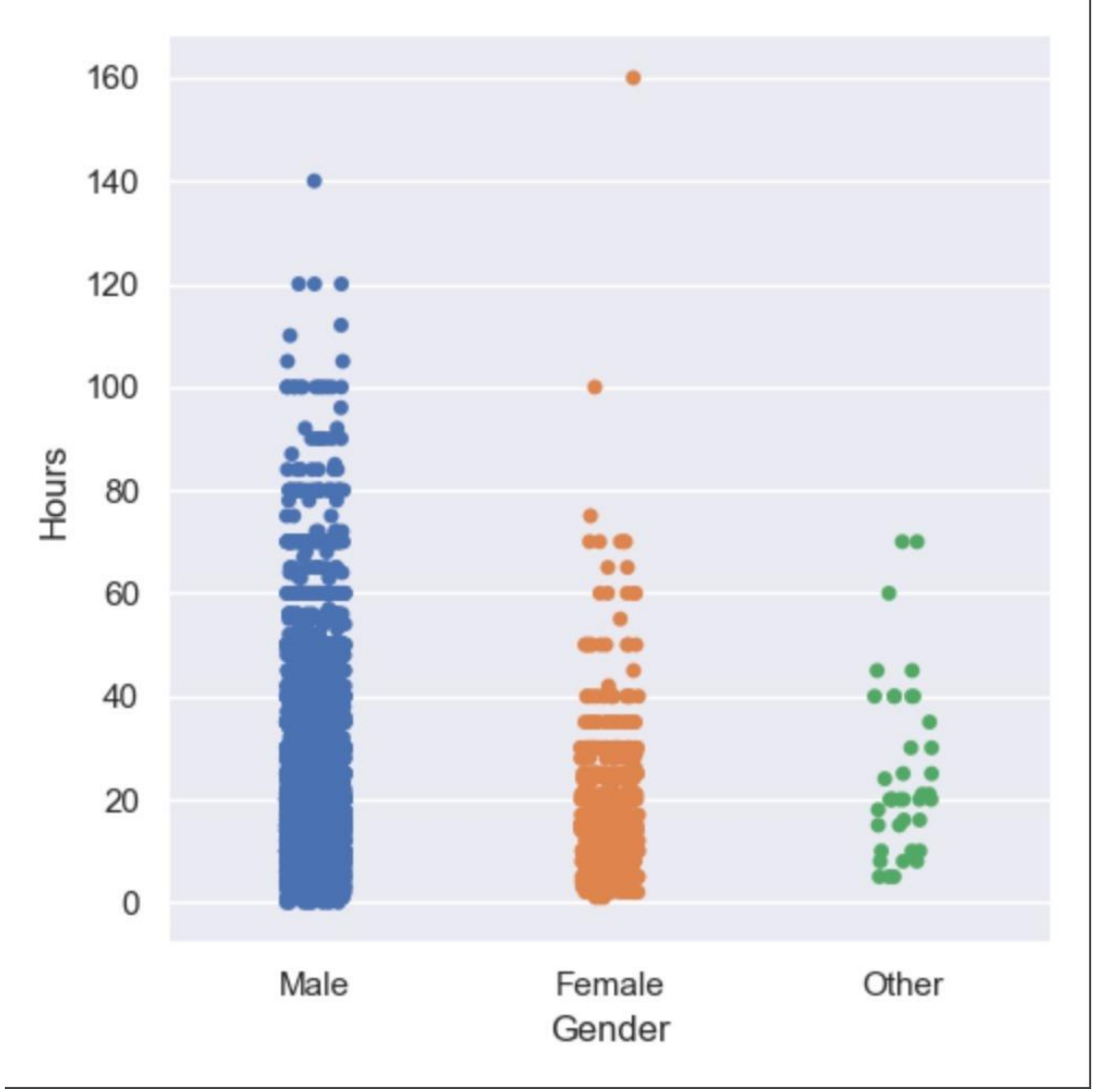
```
Int64Index: 11510 entries, 0 to 13463
Data columns (total 22 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   S. No.       11510 non-null  int64
 1   GAD1         11510 non-null  int64
 2   GAD2         11510 non-null  int64
 3   GAD3         11510 non-null  int64
 4   GAD4         11510 non-null  int64
 5   GAD5         11510 non-null  int64
 6   GAD6         11510 non-null  int64
 7   GAD7         11510 non-null  int64
 8   GADE         11510 non-null  object
 9   Game         11510 non-null  object
 10  Platform     11510 non-null  object
 11  Hours        11510 non-null  float64
 12  Narcissism   11510 non-null  float64
 13  Gender       11510 non-null  object
 14  Age          11510 non-null  int64
 15  Work         11510 non-null  object
 16  Degree       11510 non-null  object
 17  Residence    11510 non-null  object
 18  Playstyle    11510 non-null  object
 19  GAD_T        11510 non-null  object
 20  SWL_T        11510 non-null  int64
 21  SPIN_T       11510 non-null  float64
dtypes: float64(3), int64(10), object(9)
```

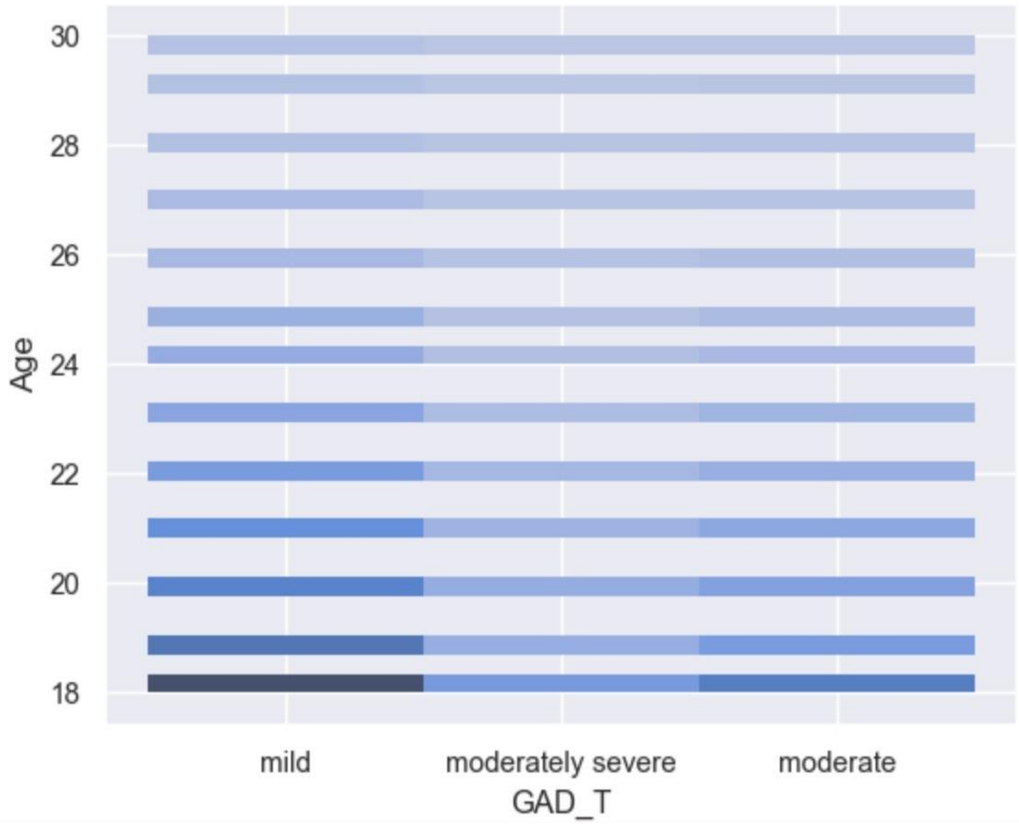
EDA Analysis

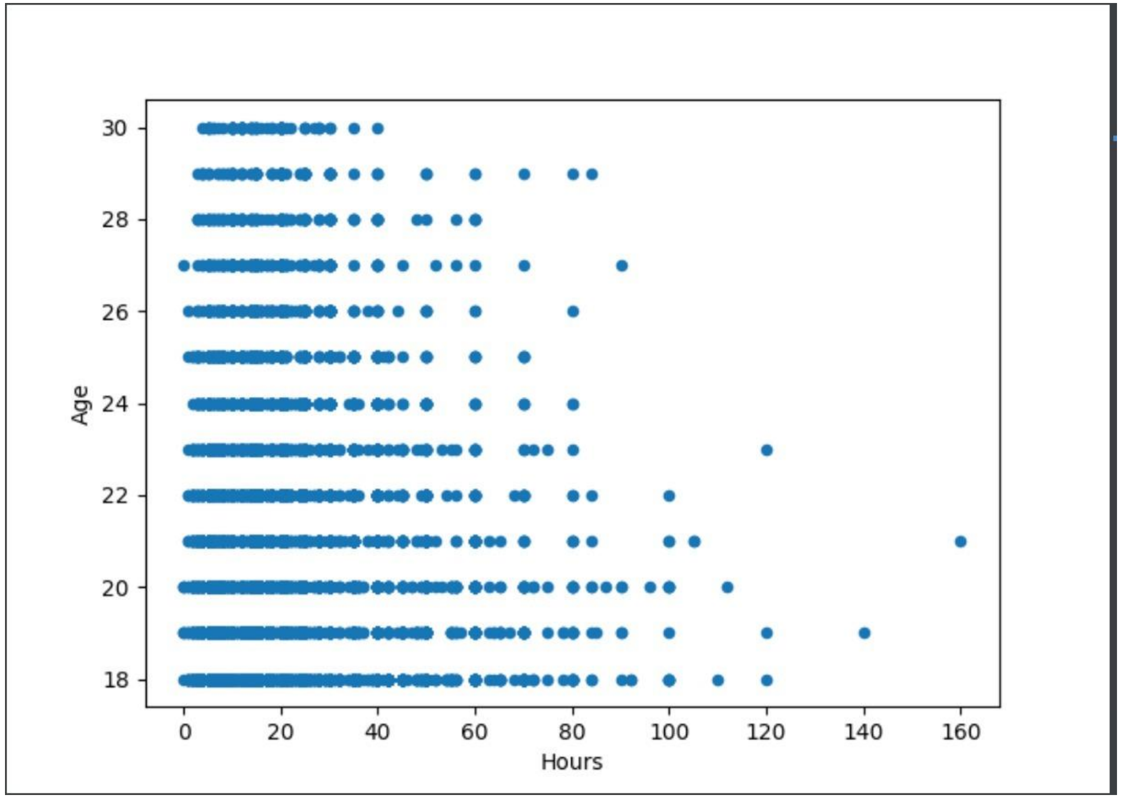
- This catplot shows the hours people are spending time on different platforms.

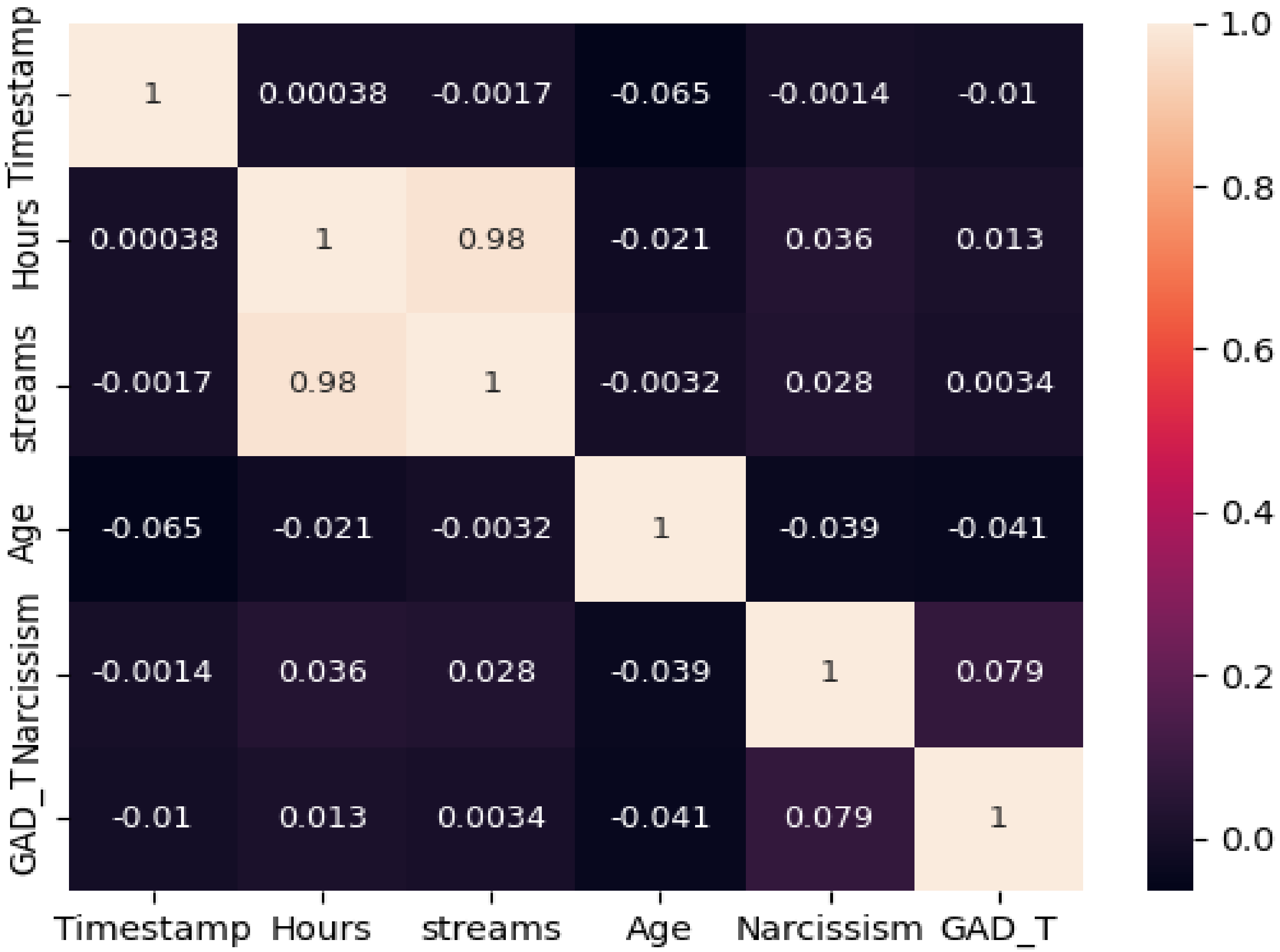


- This catplot shows the time different genders spend on online gaming.

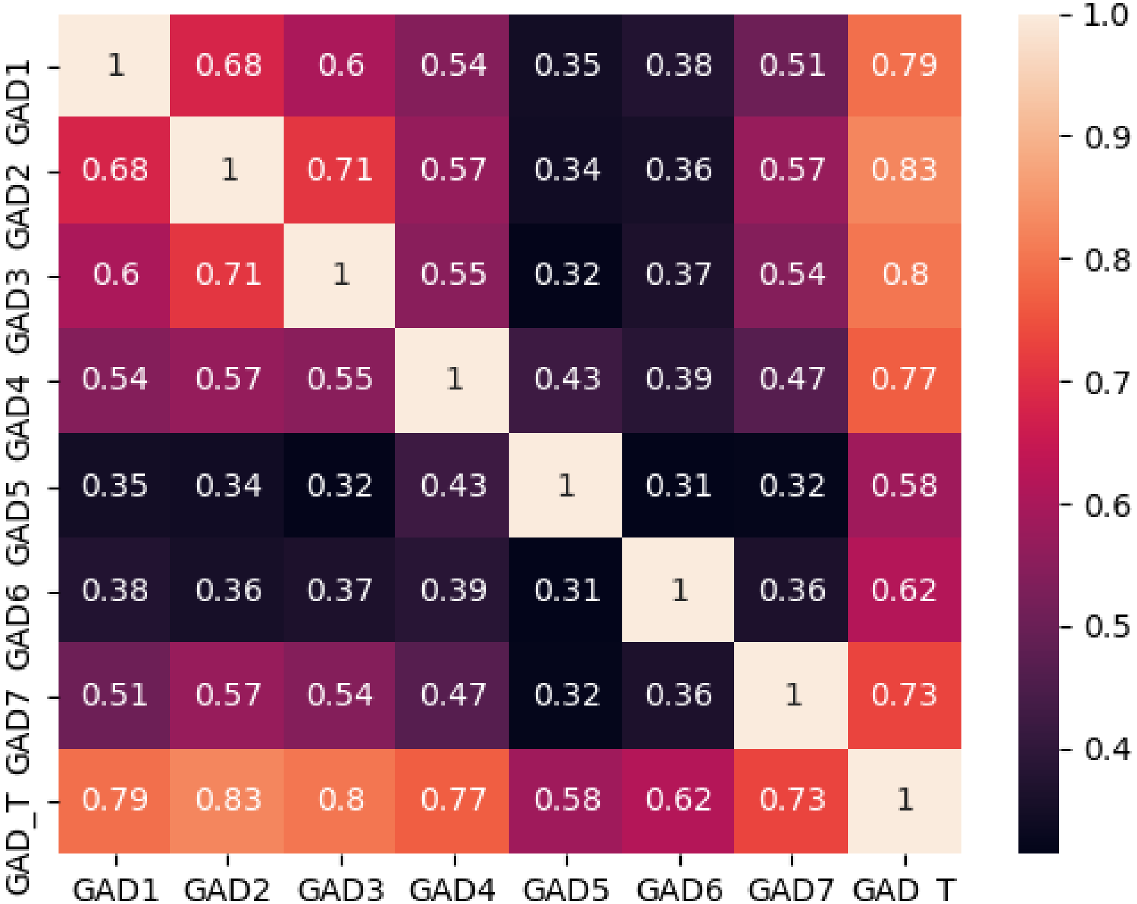



30
28
26
24
22
20
18
Age
mild
moderately severe
moderate
GAD_T

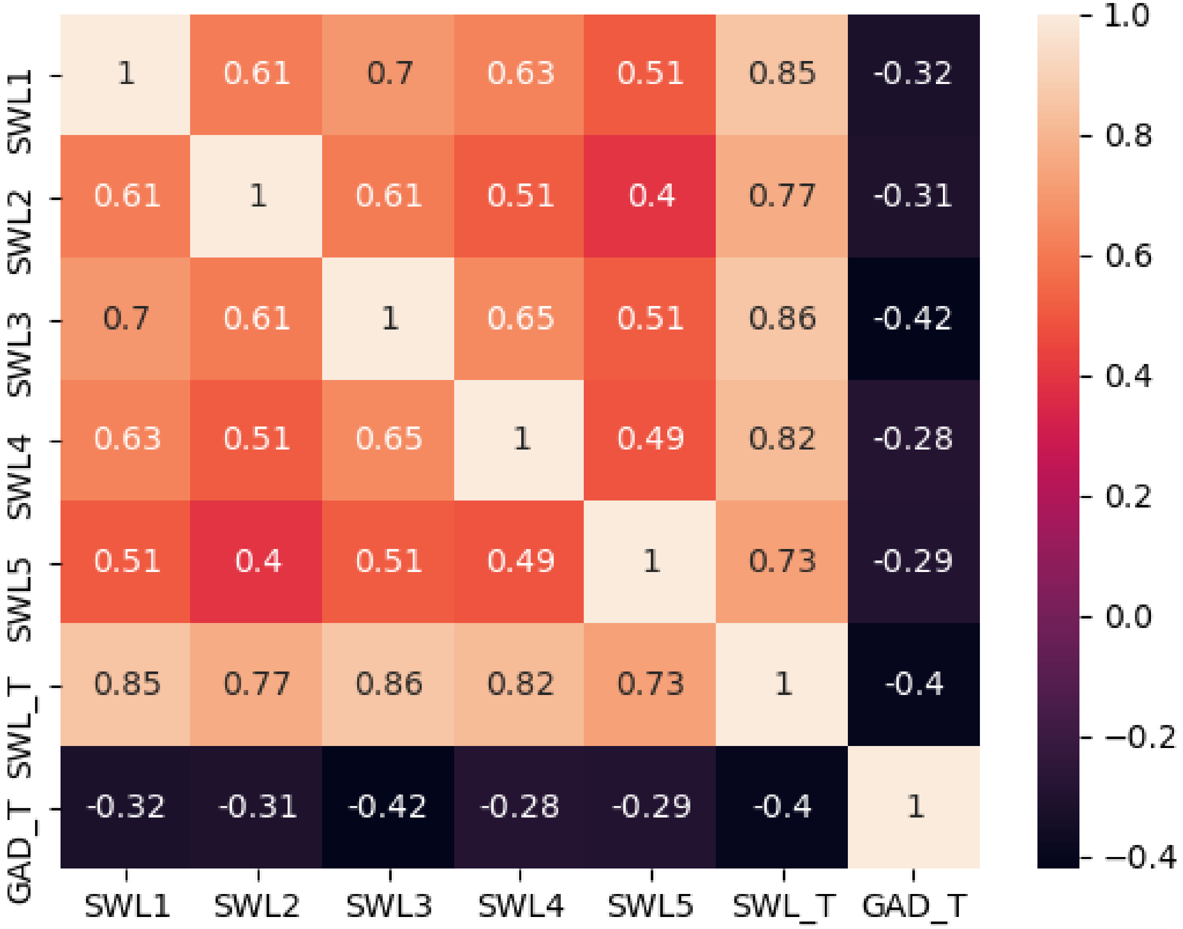

30
28
26
24
22
20
18
Age
0
20
40
60
80
100
120
140
160
Hours



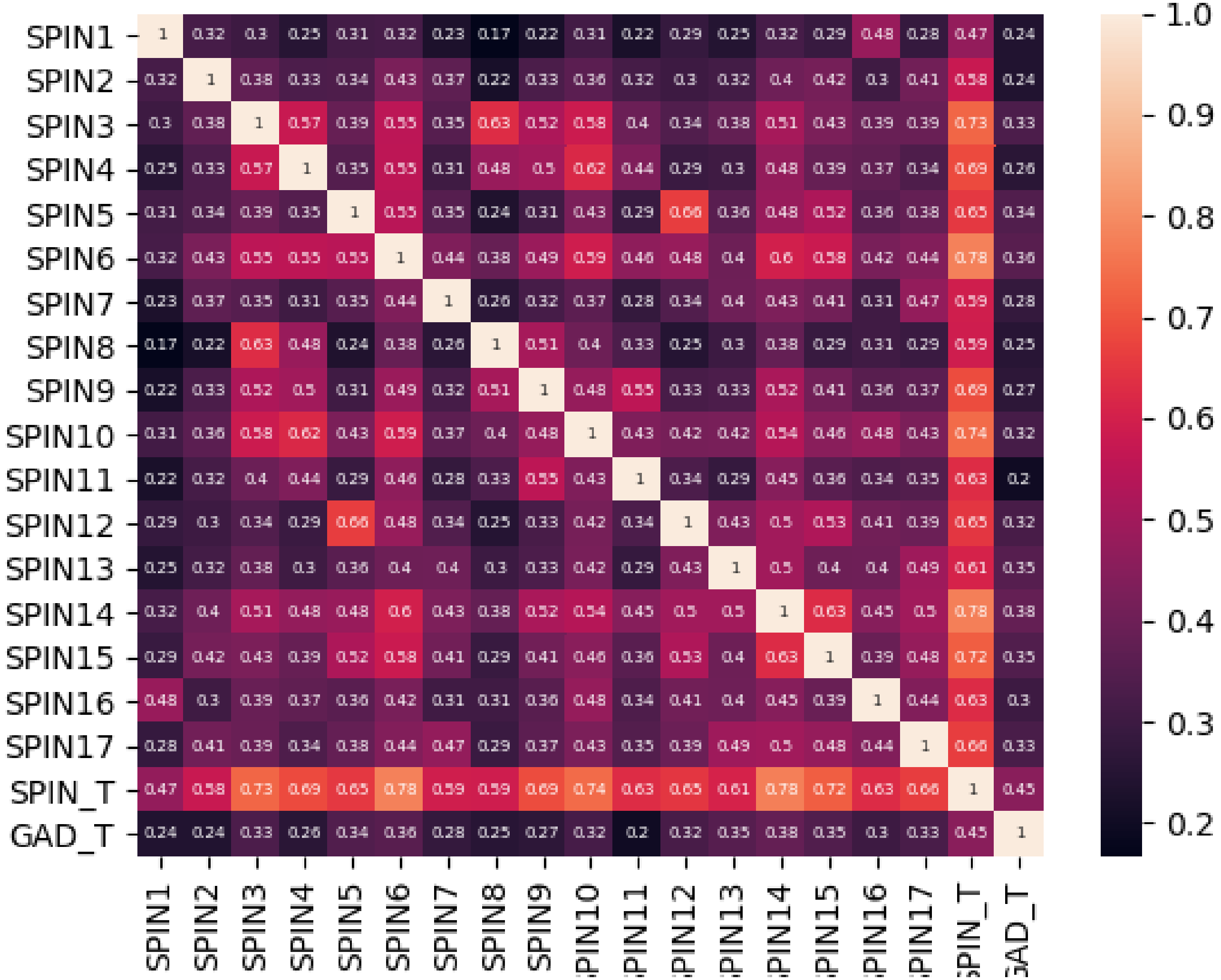
	Timestamp	Hours	streams	Age	Narcissism	GAD_T
Timestamp	1	0.00038	-0.0017	-0.065	-0.0014	-0.01
Hours	0.00038	1	0.98	-0.021	0.036	0.013
streams	-0.0017	0.98	1	-0.0032	0.028	0.0034
Age	-0.065	-0.021	-0.0032	1	-0.039	-0.041
Narcissism	-0.0014	0.036	0.028	-0.039	1	0.079
GAD_T	-0.01	0.013	0.0034	-0.041	0.079	1



	GAD1	GAD2	GAD3	GAD4	GAD5	GAD6	GAD7	GAD_T
GAD1	1	0.68	0.6	0.54	0.35	0.38	0.51	0.79
GAD2	0.68	1	0.71	0.57	0.34	0.36	0.57	0.83
GAD3	0.6	0.71	1	0.55	0.32	0.37	0.54	0.8
GAD4	0.54	0.57	0.55	1	0.43	0.39	0.47	0.77
GAD5	0.35	0.34	0.32	0.43	1	0.31	0.32	0.58
GAD6	0.38	0.36	0.37	0.39	0.31	1	0.36	0.62
GAD7	0.51	0.57	0.54	0.47	0.32	0.36	1	0.73
GAD_T	0.79	0.83	0.8	0.77	0.58	0.62	0.73	1



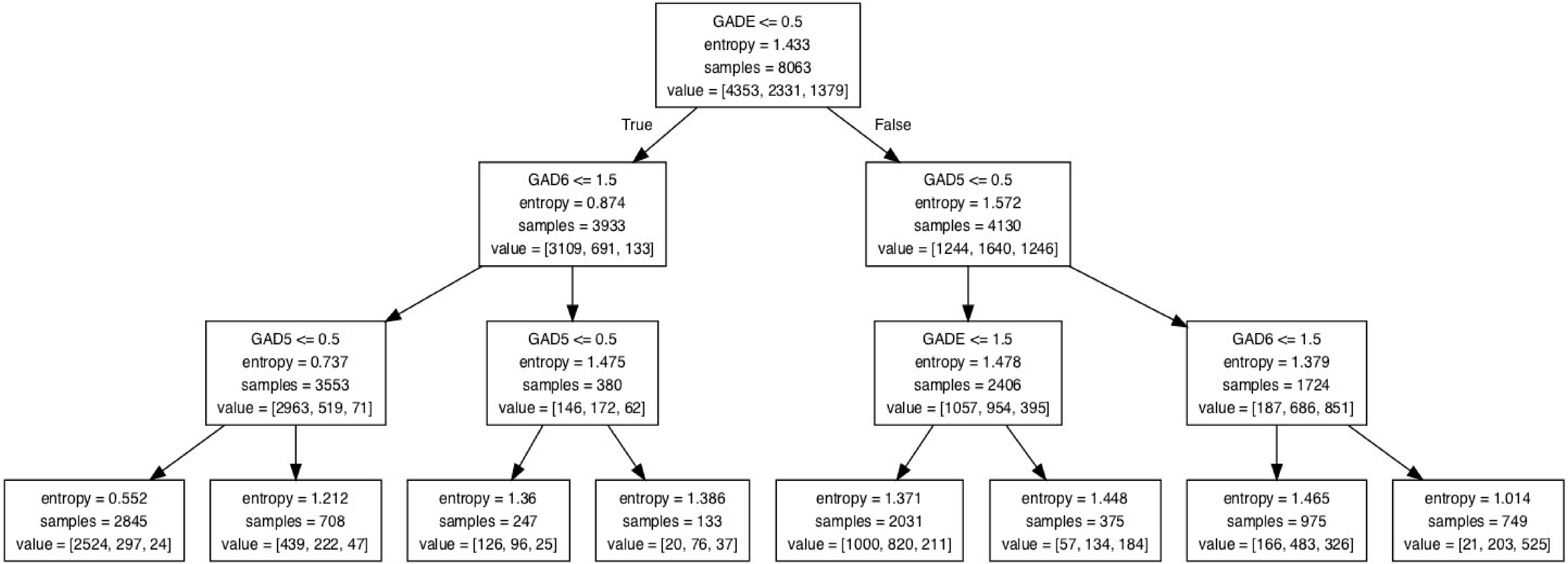
	SWL1	SWL2	SWL3	SWL4	SWL5	SWL_T	GAD_T
SWL1	1	0.61	0.7	0.63	0.51	0.85	-0.32
SWL2	0.61	1	0.61	0.51	0.4	0.77	-0.31
SWL3	0.7	0.61	1	0.65	0.51	0.86	-0.42
SWL4	0.63	0.51	0.65	1	0.49	0.82	-0.28
SWL5	0.51	0.4	0.51	0.49	1	0.73	-0.29
SWL_T	0.85	0.77	0.86	0.82	0.73	1	-0.4
GAD_T	-0.32	-0.31	-0.42	-0.28	-0.29	-0.4	1



	SPIN1	SPIN2	SPIN3	SPIN4	SPIN5	SPIN6	SPIN7	SPIN8	SPIN9	SPIN10	SPIN11	SPIN12	SPIN13	SPIN14	SPIN15	SPIN16	SPIN17	SPIN_T	GAD_T
SPIN1	1	0.32	0.3	0.25	0.31	0.32	0.23	0.17	0.22	0.31	0.22	0.29	0.25	0.32	0.29	0.48	0.28	0.47	0.24
SPIN2	0.32	1	0.38	0.33	0.34	0.43	0.37	0.22	0.33	0.36	0.32	0.3	0.32	0.4	0.42	0.3	0.41	0.58	0.24
SPIN3	0.3	0.38	1	0.57	0.39	0.55	0.35	0.63	0.52	0.58	0.4	0.34	0.38	0.51	0.43	0.39	0.39	0.73	0.33
SPIN4	0.25	0.33	0.57	1	0.35	0.55	0.31	0.48	0.5	0.62	0.44	0.29	0.3	0.48	0.39	0.37	0.34	0.69	0.26
SPIN5	0.31	0.34	0.39	0.35	1	0.55	0.35	0.24	0.31	0.43	0.29	0.66	0.36	0.48	0.52	0.36	0.38	0.65	0.34
SPIN6	0.32	0.43	0.55	0.55	0.55	1	0.44	0.38	0.49	0.59	0.46	0.48	0.4	0.6	0.58	0.42	0.44	0.78	0.36
SPIN7	0.23	0.37	0.35	0.31	0.35	0.44	1	0.26	0.32	0.37	0.28	0.34	0.4	0.43	0.41	0.31	0.47	0.59	0.28
SPIN8	0.17	0.22	0.63	0.48	0.24	0.38	0.26	1	0.51	0.4	0.33	0.25	0.3	0.38	0.29	0.31	0.29	0.59	0.25
SPIN9	0.22	0.33	0.52	0.5	0.31	0.49	0.32	0.51	1	0.48	0.55	0.33	0.33	0.52	0.41	0.36	0.37	0.69	0.27
SPIN10	0.31	0.36	0.58	0.62	0.43	0.59	0.37	0.4	0.48	1	0.43	0.42	0.42	0.54	0.46	0.48	0.43	0.74	0.32
SPIN11	0.22	0.32	0.4	0.44	0.29	0.46	0.28	0.33	0.55	0.43	1	0.34	0.29	0.45	0.36	0.34	0.35	0.63	0.2
SPIN12	0.29	0.3	0.34	0.29	0.66	0.48	0.34	0.25	0.33	0.42	0.34	1	0.43	0.5	0.53	0.41	0.39	0.65	0.32
SPIN13	0.25	0.32	0.38	0.3	0.36	0.4	0.4	0.3	0.33	0.42	0.29	0.43	1	0.5	0.4	0.4	0.49	0.61	0.35
SPIN14	0.32	0.4	0.51	0.48	0.48	0.6	0.43	0.38	0.52	0.54	0.45	0.5	0.5	1	0.63	0.45	0.5	0.78	0.38
SPIN15	0.29	0.42	0.43	0.39	0.52	0.58	0.41	0.29	0.41	0.46	0.36	0.53	0.4	0.63	1	0.39	0.48	0.72	0.35
SPIN16	0.48	0.3	0.39	0.37	0.36	0.42	0.31	0.31	0.36	0.48	0.34	0.41	0.4	0.45	0.39	1	0.44	0.63	0.3
SPIN17	0.28	0.41	0.39	0.34	0.38	0.44	0.47	0.29	0.37	0.43	0.35	0.39	0.49	0.5	0.48	0.44	1	0.66	0.33
SPIN_T	0.47	0.58	0.73	0.69	0.65	0.78	0.59	0.59	0.69	0.74	0.63	0.65	0.61	0.78	0.72	0.63	0.66	1	0.45
GAD_T	0.24	0.24	0.33	0.26	0.34	0.36	0.28	0.25	0.27	0.32	0.2	0.32	0.35	0.38	0.35	0.3	0.33	0.45	1

Model Development

Decision Tree:



GADE <= 0.5
entropy = 1.433
samples = 8063
value = [4353, 2331, 1379]

True

False

GAD6 <= 1.5
entropy = 0.874
samples = 3933
value = [3109, 691, 133]

GAD5 <= 0.5
entropy = 1.572
samples = 4130
value = [1244, 1640, 1246]

GAD5 <= 0.5
entropy = 0.737
samples = 3553
value = [2963, 519, 71]

GAD5 <= 0.5
entropy = 1.475
samples = 380
value = [146, 172, 62]

GADE <= 1.5
entropy = 1.478
samples = 2406
value = [1057, 954, 395]

GAD6 <= 1.5
entropy = 1.379
samples = 1724
value = [187, 686, 851]

entropy = 0.552
samples = 2845
value = [2524, 297, 24]

entropy = 1.212
samples = 708
value = [439, 222, 47]

entropy = 1.36
samples = 247
value = [126, 96, 25]

entropy = 1.386
samples = 133
value = [20, 76, 37]

entropy = 1.371
samples = 2031
value = [1000, 820, 211]

entropy = 1.448
samples = 375
value = [57, 134, 184]

entropy = 1.465
samples = 975
value = [166, 483, 326]

entropy = 1.014
samples = 749
value = [21, 203, 525]

Random Forest:

- The Random forest classifier creates a set of decision trees from a randomly selected subset of the training set.

- It is basically a set of decision trees (DT) from a randomly selected subset of the training set and then It collects the votes from different decision trees to decide the final prediction.

- The dashboard incorporates the feature selection and plotting the roc-auc graphs.

- The random forest classifier algorithm is being built using the Scikit Learn Package.

Support Vector Machine:

- Support vector Machines(SVM) is a widely used supervised learning method and it can be used for regression, classification, anomaly detection problems.

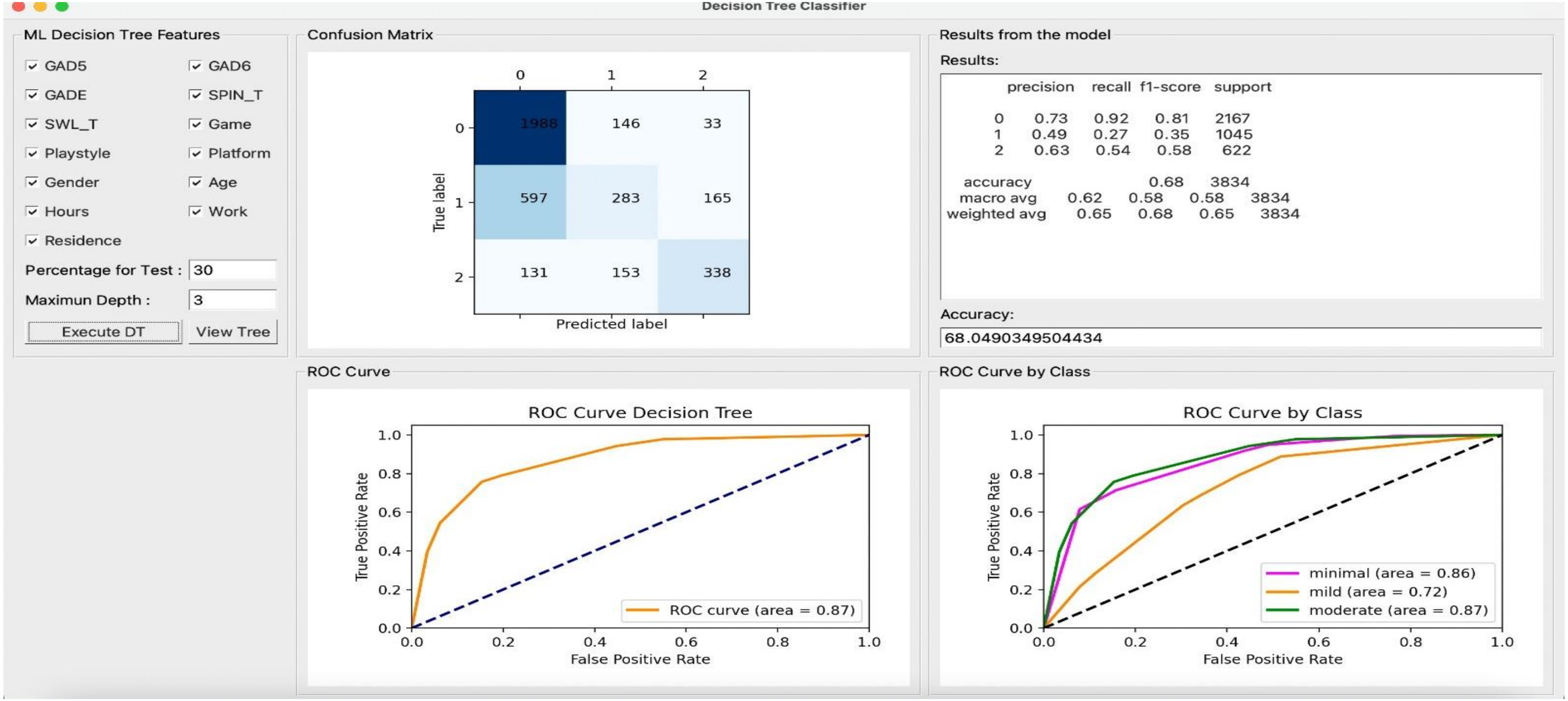
- The SVM based classifier is called the SVC(Support Vector Classifier) and which can be used it in classification problems.

- The kernel has been set to **linear** for our data by default.

Performance Measure

- **Confusion Matrix:**
  - ✓ The number of correct and incorrect predictions are summarized with count values and broken down by each class.
- **Classification report:**
  - ✓ Classification Report shows the main classification metrics precision, recall and f1-score on per class basis.
- Accuracy is the fraction of predictions our model got right.
- ROC is the probability curve.

Decision Tree:



Decision Tree Classifier

ML Decision Tree Features

- ☑ GAD5
- ☑ GADE
- ☑ SWL_T
- ☑ Playstyle
- ☑ Gender
- ☑ Hours
- ☑ Residence
- ☑ GAD6
- ☑ SPIN_T
- ☑ Game
- ☑ Platform
- ☑ Age
- ☑ Work

Percentage for Test : 30

Maximun Depth : 3

Execute DT   View Tree

Confusion Matrix

True label \ Predicted label	0	1	2
0	1988	146	33
1	597	283	165
2	131	153	338

Results from the model

Results:

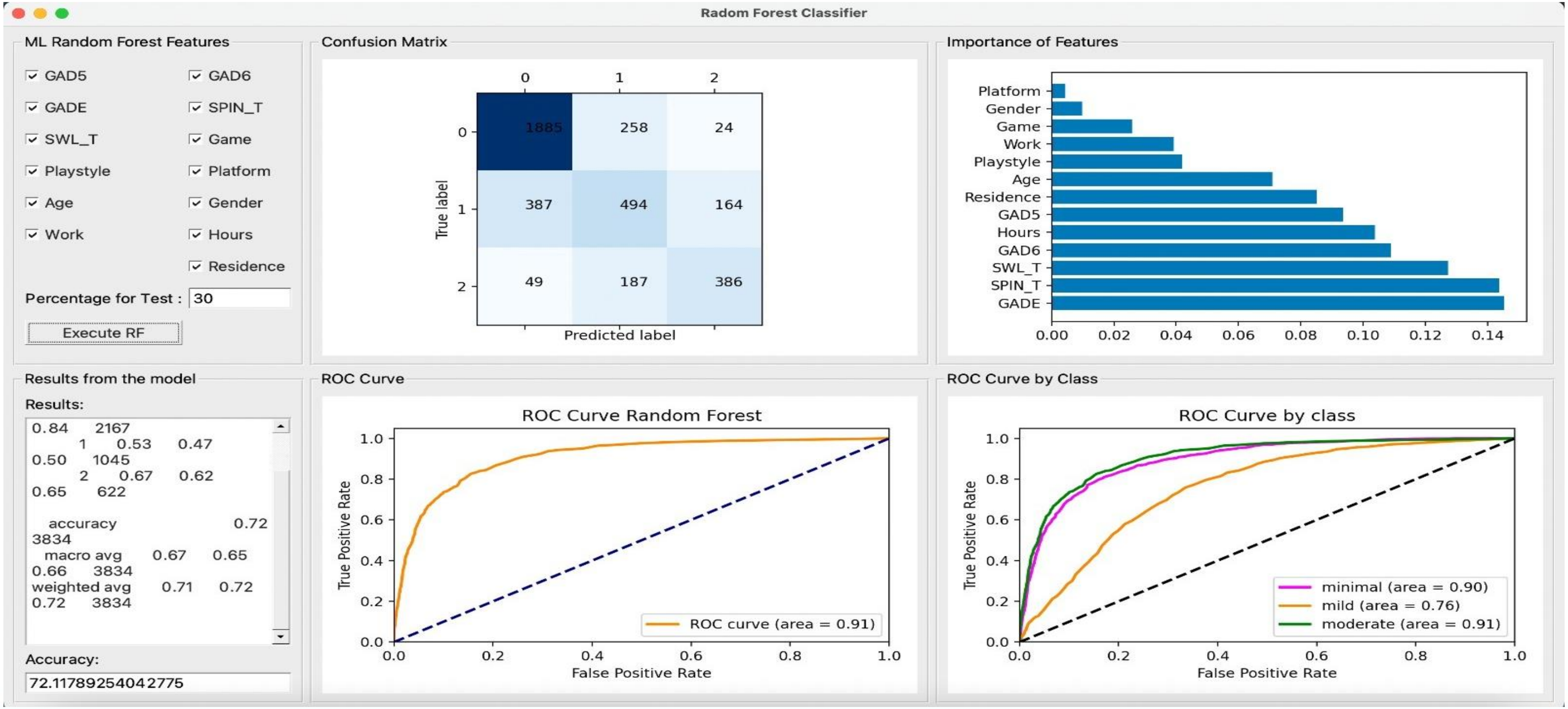
```
              precision    recall  f1-score   support

           0       0.73      0.92      0.81      2167
           1       0.49      0.27      0.35      1045
           2       0.63      0.54      0.58       622

    accuracy                           0.68      3834
   macro avg       0.62      0.58      0.58      3834
weighted avg       0.65      0.68      0.65      3834
```

Accuracy: 68.0490349504434

ROC Curve

ROC Curve by Class

Random Forest:



ML Random Forest Features

☑ GAD5 ☑ GAD6
☑ GADE ☑ SPIN_T
☑ SWL_T ☑ Game
☑ Playstyle ☑ Platform
☑ Age ☑ Gender
☑ Work ☑ Hours
☑ Residence

Percentage for Test : 30

Confusion Matrix

	0	1	2
0	1885	258	24
1	387	494	164
2	49	187	386

Results from the model

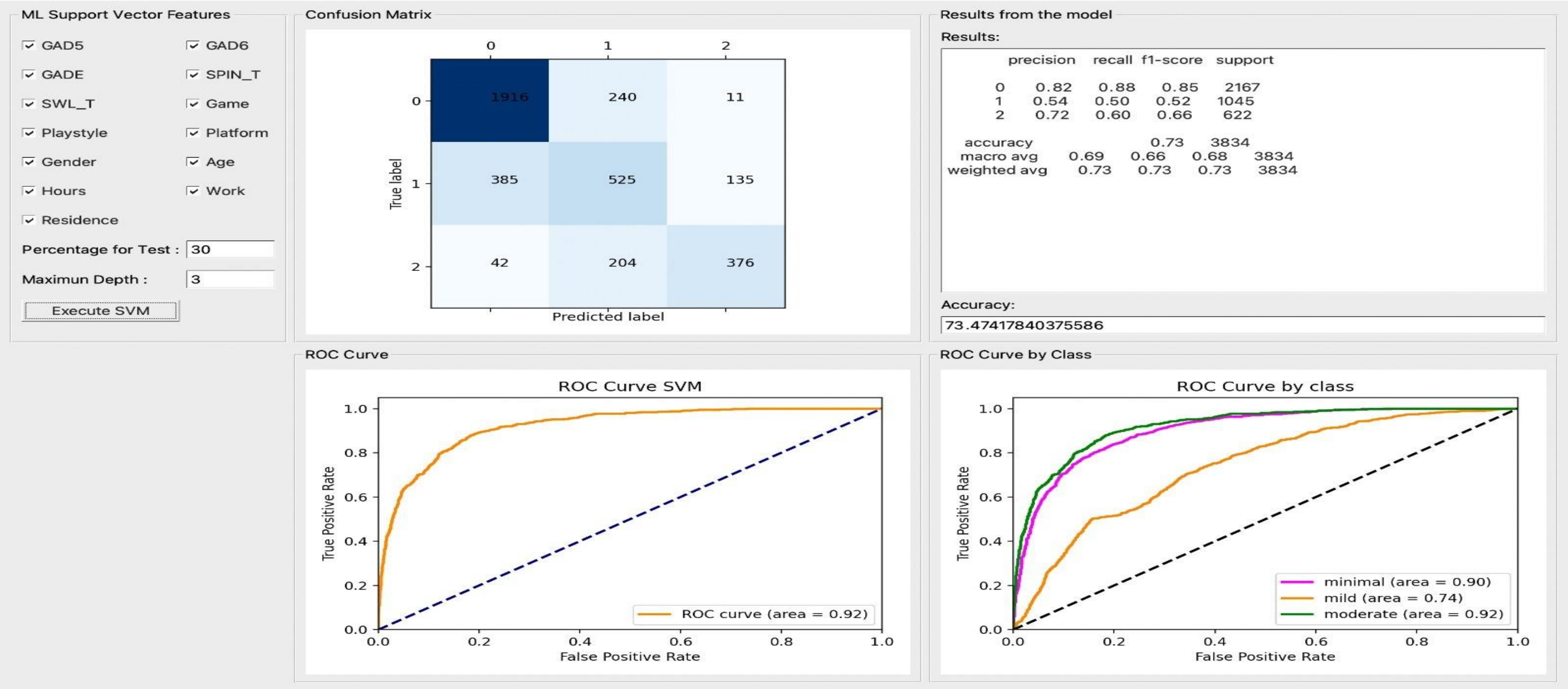
Results:

```
0.84     2167
        1     0.53     0.47
0.50     1045
        2     0.67     0.62
0.65     622

   accuracy                          0.72
3834
   macro avg      0.67     0.65
0.66     3834
weighted avg      0.71     0.72
0.72     3834
```

Accuracy:

72.11789254042775

Support Vector Machine:



ML Support Vector Features

☑ GAD5
☑ GADE
☑ SWL_T
☑ Playstyle
☑ Gender
☑ Hours
☑ Residence

☑ GAD6
☑ SPIN_T
☑ Game
☑ Platform
☑ Age
☑ Work

Percentage for Test : 30

Maximun Depth : 3

Execute SVM

Confusion Matrix

True label \ Predicted label	0	1	2
0	1916	240	11
1	385	525	135
2	42	204	376

Results from the model

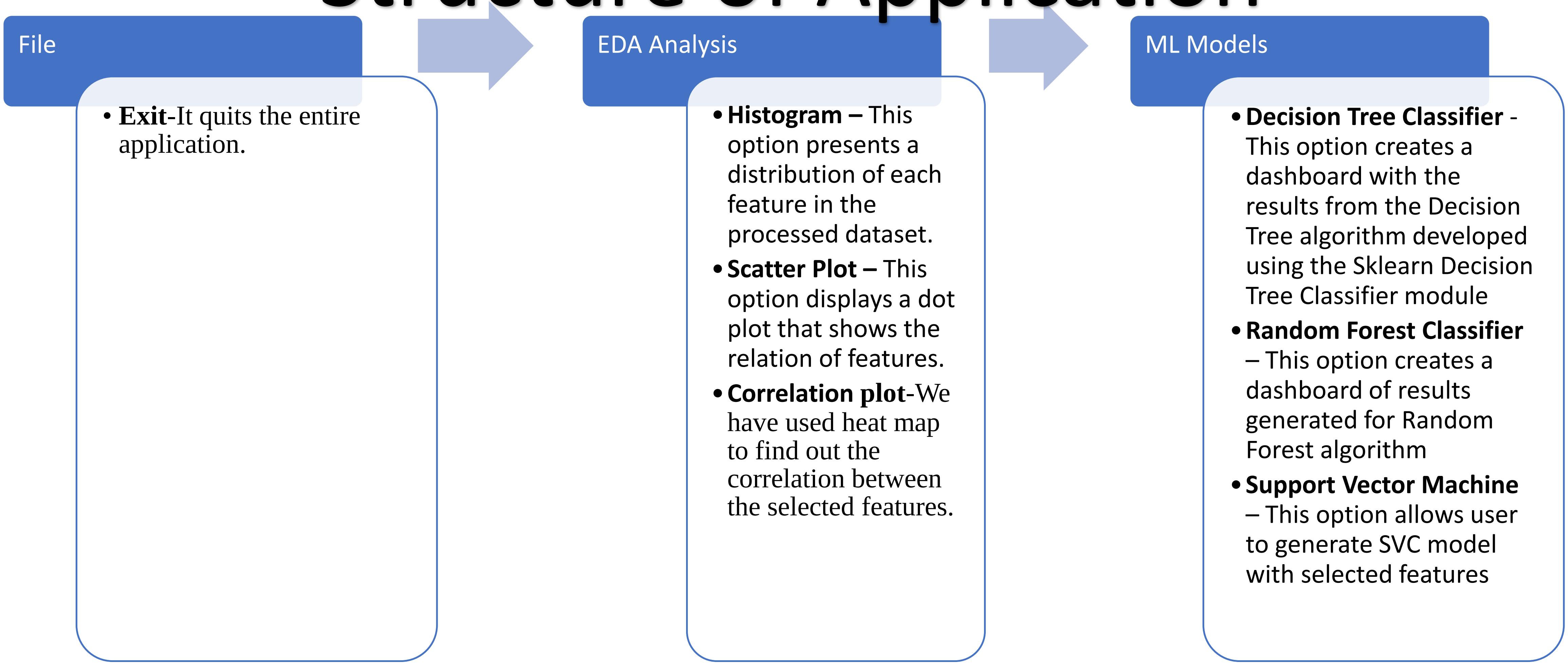
Results:

```
              precision    recall  f1-score   support

           0       0.82      0.88      0.85      2167
           1       0.54      0.50      0.52      1045
           2       0.72      0.60      0.66       622

    accuracy                           0.73      3834
   macro avg       0.69      0.66      0.68      3834
weighted avg       0.73      0.73      0.73      3834
```

Accuracy:

73.47417840375586

Structure of Application



File

- **Exit**-It quits the entire application.

EDA Analysis

- **Histogram –** This option presents a distribution of each feature in the processed dataset.
- **Scatter Plot –** This option displays a dot plot that shows the relation of features.
- **Correlation plot**-We have used heat map to find out the correlation between the selected features.

ML Models

- **Decision Tree Classifier** - This option creates a dashboard with the results from the Decision Tree algorithm developed using the Sklearn Decision Tree Classifier module
- **Random Forest Classifier** – This option creates a dashboard of results generated for Random Forest algorithm
- **Support Vector Machine** – This option allows user to generate SVC model with selected features

Results

Decision Tree Classifier:

Accuracy of model = 68.04%(Test size=30%)
Precision of 0’s=0.73 Precision of 1’s=0.49 Precision of 2’s = 0.63

Random Forest Classifier:
Accuracy of model = 72.11% (Test size=30%)

Precision of 0’s=0.84 Precision of 1’s=0.50 Precision of 2’s= 0.65

Support Vector Classifier:
Accuracy of model = 73.47% (Test size= 30%)
Precision of 0’s=0.82 Precision of 1’s=0.54 Precision of 2’s = 0.72

Conclusion

- Comparing the results of models, almost all the three models has accuracy value in between 68% -74%.

- Support Vector Machine tops the list by having the highest accuracy of 73.47%

- The model in future enhancement needs to be tuned to predict the effect of online gaming on General Anxiety Disorder(GAD).

DEMO

Video link:

https://drive.google.com/file/d/1Qah2_25ts8iJ44lg-YvKWEP0jkiQgzse/view

Thank You

GitHub Link: https://github.com/stripathi50/Final-Project-Group1-