Online gaming anxiety data - Exploring Gamer Behaviour and Mental Health Patterns using Classification

DSC 478

Members: Nikhil Yakkala, Vishal Rajashekar, Himanshu Aneja

Dataset -
https://www.kaggle.com/datasets/divyansh22/online-gaming-anxiety-data/data

Executive Summary:

Our research leverages a global survey dataset to investigate the complex relationship between online gaming habits and mental health. Our primary objective is to uncover underlying patterns and insights into gamer behaviour, particularly focusing on anxiety levels, social phobia, and life satisfaction. This exploration is facilitated through a structured approach encompassing pre-processing, exploratory analysis, and supervised learning.

In the pre-processing phase, we meticulously handled missing values, ensured data consistency, and standardized numerical features using feature scaling techniques. Furthermore, normalization methods were employed to effectively manage outliers. Subsequently, through exploratory data analysis (EDA), we gained valuable insights into the dataset, identifying significant patterns and trends within gamer behaviour. Clustering algorithms were instrumental in grouping players based on shared characteristics, providing deeper insights into the heterogeneous gaming community.

During supervised learning, we employed various classification algorithms such as Random Forest and Logistic Regression to address classification tasks. Rigorous evaluation using metrics like accuracy, precision, and recall enabled us to gauge the performance of each algorithm in predicting gamer behaviour and mental health trends. By determining the most effective classification algorithm, we were able to draw meaningful conclusions about the mental health and behaviour of gamers, laying a foundation for future research and interventions aimed at enhancing the well-being of the gaming community.

Methods:

Data Preprocessing, Exploratory Data Analysis (EDA), Feature Engineering, K-means Clustering, Cross-Validation, Classification.

Data Analysis:

Data Schema and Size:

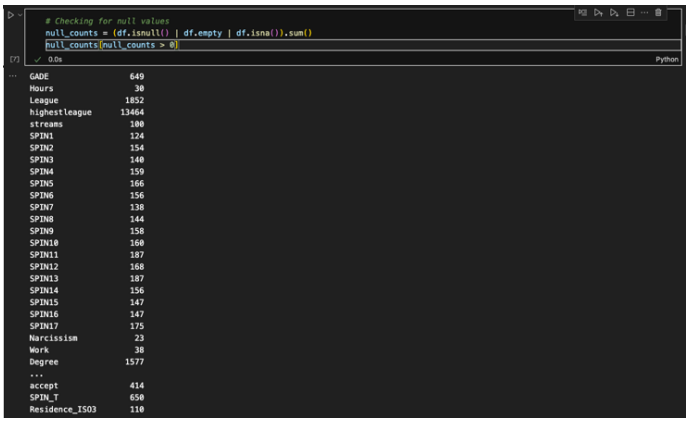
This dataset consists of 13,464 entries and 55 columns. These columns include metrics such as Generalized Anxiety Disorder (GAD) scores, Satisfaction With Life (SWL) scores, gaming-related behaviours, and demographic details.

Data Pre-processing:

Several data cleansing steps were required before moving into data exploration and machine learning tasks:

Null values:

Cleaning null values: We have replaced the null values of GADE column by its value count, whereas streams and Hours column null values by means.



```
# Checking for null values
null_counts = (df.isnull() | df.empty | df.isna()).sum()
null_counts[null_counts > 0]
```

```
GADE               649
Hours               30
League            1852
highestleague    13464
streams            100
SPIN1              124
SPIN2              154
SPIN3              140
SPIN4              159
SPIN5              166
SPIN6              156
SPIN7              138
SPIN8              144
SPIN9              158
SPIN10             160
SPIN11             187
SPIN12             168
SPIN13             187
SPIN14             156
SPIN15             147
SPIN16             147
SPIN17             175
Narcissism          23
Work                38
Degree            1577
...
accept             414
SPIN_T             650
Residence_ISO3     110
```

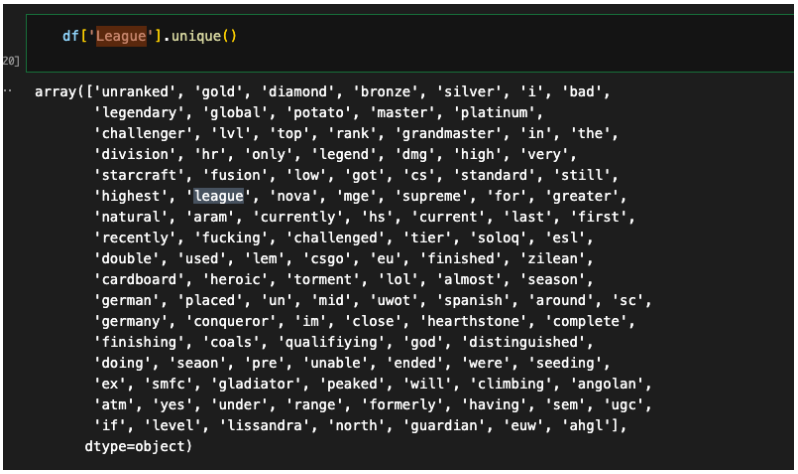
Dropping Columns:

We removed some unnecessary columns from our dataset to make it easier to analyze like 'S. No.' , 'Timestamp', Birthplace ,Birthplace_ISO3 and highestleague columns.

Standardizing the columns:

We have standardized the League column values to ensure consistency across the dataset.

Before:



```
df['League'].unique()
```

```
array(['unranked', 'gold', 'diamond', 'bronze', 'silver', 'i', 'bad',
       'legendary', 'global', 'potato', 'master', 'platinum',
       'challenger', 'lvl', 'top', 'rank', 'grandmaster', 'in', 'the',
       'division', 'hr', 'only', 'legend', 'dmg', 'high', 'very',
       'starcraft', 'fusion', 'low', 'got', 'cs', 'standard', 'still',
       'highest', 'league', 'nova', 'mge', 'supreme', 'for', 'greater',
       'natural', 'aram', 'currently', 'hs', 'current', 'last', 'first',
       'recently', 'fucking', 'challenged', 'tier', 'soloq', 'esl',
       'double', 'used', 'lem', 'csgo', 'eu', 'finished', 'zilean',
       'cardboard', 'heroic', 'torment', 'lol', 'almost', 'season',
       'german', 'placed', 'un', 'mid', 'uwot', 'spanish', 'around', 'sc',
       'germany', 'conqueror', 'im', 'close', 'hearthstone', 'complete',
       'finishing', 'coals', 'qualifiying', 'god', 'distinguished',
       'doing', 'seaon', 'pre', 'unable', 'ended', 'were', 'seeding',
       'ex', 'smfc', 'gladiator', 'peaked', 'will', 'climbing', 'angolan',
       'atm', 'yes', 'under', 'range', 'formerly', 'having', 'sem', 'ugc',
       'if', 'level', 'lissandra', 'north', 'guardian', 'euw', 'ahgl'],
      dtype=object)
```

After:

```
df['League'].unique()
```

```
array(['unranked', 'gold', 'diamond', 'bronze', 'silver', 'unspecified',
       'legendary', 'master', 'platinum', 'challenger', 'top',
       'grandmaster', 'legend', 'dmg', 'league', 'nova', 'mge', 'supreme',
       'greater', 'lem'], dtype=object)
```

Outlier Detection and Treatment

Introduction

Outliers are data points that significantly deviate from the rest of the dataset. They can adversely affect the statistical analysis and modeling process by skewing results and introducing bias. Therefore, it is essential to detect and appropriately handle outliers before proceeding with further analysis.

Methods Used

Two commonly used methods for outlier detection are the Z-score method and the Interquartile Range (IQR) method.

1. **Z-Score Method:**
   - The Z-score measures how many standard deviations a data point is from the mean of the dataset.
   - Data points with a Z-score greater than a predefined threshold (typically 3 or -3) are considered outliers.
   - Identified outliers are then flagged for further investigation or treatment.
2. **Interquartile Range (IQR) Method:**
   - The IQR measures the spread of the middle 50% of the data, calculated as the difference between the third quartile (Q3) and the first quartile (Q1).
   - Data points falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are identified as outliers.
   - These outliers are also marked for further analysis or removal.

```
Rows with outliers (using IQR method):
       GAD1  GAD2  GAD3  GAD4  GAD5  GAD6  GAD7                  GADE  SWL1  \
1         1     2     2     2     0     1     0    Somewhat difficult     3
2         0     2     2     0     0     3     1  Not difficult at all     2
3         0     0     0     0     0     0     0  Not difficult at all     2
4         2     1     2     2     2     3     2       Very difficult     2
8         2     3     2     2     0     1     2       Very difficult     2
...     ...   ...   ...   ...   ...   ...   ...                   ...   ...
13447     0     0     0     3     1     2     0  Not difficult at all     5
13451     0     0     1     0     0     1     0  Not difficult at all     5
13452     1     0     1     0     0     0     0    Somewhat difficult     2
13460     3     3     3     3     2     3     3  Extremely difficult     5
13462     3     2     1     3     0     1     3    Somewhat difficult     2
```

Outlier Detection Results

1. **Z-Score Method:**
    - The Z-score method identified a total of 1007 rows containing outliers.
    - Outliers were detected across various numerical variables, indicating potential anomalies in the data distribution.
    - Examples of variables with outliers include GAD scores, SWL scores, gaming hours, and narcissism levels.
2. **Interquartile Range (IQR) Method:**
    - The IQR method detected a larger number of outliers, totaling 4259 rows.
    - Similar to the Z-score method, outliers were found in multiple numerical features, suggesting potential data inconsistencies or extreme values.

Outlier Treatment

To ensure the integrity of subsequent analysis and modeling, the identified outliers were treated as follows:

- **Visualization:** Boxplots were created to visually inspect the distribution of numerical variables with and without outliers.
- **Data Filtering:** Outliers detected by both the Z-score and IQR methods were filtered out from the dataset.
- **Dataset Composition:** Two filtered datasets were obtained, one for each outlier detection method, ensuring robustness in subsequent analyses.

Conclusion

Outlier detection and treatment are critical steps in data preprocessing to ensure the reliability and validity of analytical results. By employing both the Z-score and IQR methods, we were able to identify and address potential outliers in the dataset. Removing outliers helps improve the accuracy of statistical analyses and machine learning models, leading to more robust insights and conclusions.

Exploratory Data Analysis (EDA)

Introduction

Exploratory Data Analysis (EDA) is an essential step in the data analysis process, aimed at understanding the underlying patterns, trends, and relationships within a dataset. In this section, we conduct an EDA on our dataset to gain insights into various aspects of the data.

Summary Statistics

Before delving into visualizations, let's first examine some summary statistics of our dataset:

- The dataset consists of 6480 observations and 36 numerical variables.

- The mean GAD (Generalized Anxiety Disorder) score is approximately 0.15, with a standard deviation of 0.13.
- Participants report moderate levels of life satisfaction (SWL) with an average score of around 0.55.
- The average SPIN (Social Phobia Inventory) score is approximately 0.21, indicating relatively low levels of social phobia on average.

	GAD1	GAD2	GAD3	GAD4	GAD5	\
count	6480.000000	6480.000000	6480.000000	6480.000000	6480.000000	
mean	0.538426	0.350000	0.646451	0.420370	0.268210	
std	0.619739	0.572859	0.742199	0.614296	0.544068	
min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	0.000000	
50%	0.000000	0.000000	1.000000	0.000000	0.000000	
75%	1.000000	1.000000	1.000000	1.000000	0.000000	
max	2.000000	2.000000	3.000000	2.000000	2.000000	

	GAD6	GAD7	SWL1	SWL2	SWL3	\
count	6480.000000	6480.000000	6480.000000	6480.000000	6480.000000	
mean	0.617593	0.304938	4.051698	4.919444	4.771451	
std	0.676188	0.559673	1.642246	1.523087	1.617952	
min	0.000000	0.000000	1.000000	1.000000	1.000000	
25%	0.000000	0.000000	3.000000	4.000000	4.000000	
50%	1.000000	0.000000	4.000000	5.000000	5.000000	
75%	1.000000	1.000000	5.000000	6.000000	6.000000	
max	2.000000	2.000000	7.000000	7.000000	7.000000	

	SWL4	SWL5	Hours	streams	SPIN1	...	\
count	6480.000000	6480.000000	6480.000000	6480.000000	6480.000000	...	
mean	4.075617	3.686111	0.002401	0.001020	0.637809	...	
std	1.737476	1.863764	0.001275	0.000809	0.706827	...	
min	1.000000	1.000000	0.000000	0.000000	0.000000	...	
25%	3.000000	2.000000	0.001250	0.000444	0.000000	...	
50%	4.000000	3.000000	0.002500	0.000889	1.000000	...	
75%	6.000000	5.000000	0.003125	0.001555	1.000000	...	
max	7.000000	7.000000	0.006500	0.003333	2.000000	...	

	SPIN8	SPIN9	SPIN10	SPIN11	SPIN12
count	6480.000000	6480.000000	6480.000000	6480.000000	6480.000000
mean	1.054012	1.198765	0.504784	1.467438	0.465432
std	1.200343	1.217762	0.676126	1.375997	0.679815
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000
50%	1.000000	1.000000	0.000000	1.000000	0.000000
75%	2.000000	2.000000	1.000000	3.000000	1.000000
max	4.000000	4.000000	2.000000	4.000000	2.000000

	SPIN13	SPIN14	SPIN15	SPIN16	SPIN17
count	6480.000000	6480.000000	6480.000000	6480.000000	6480.000000
mean	0.243981	0.827778	0.980093	0.308488	0.602160
std	0.540570	0.920836	1.093445	0.568822	0.902943
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	1.000000	1.000000	0.000000	0.000000
75%	0.000000	1.000000	2.000000	1.000000	1.000000
max	2.000000	4.000000	4.000000	2.000000	4.000000

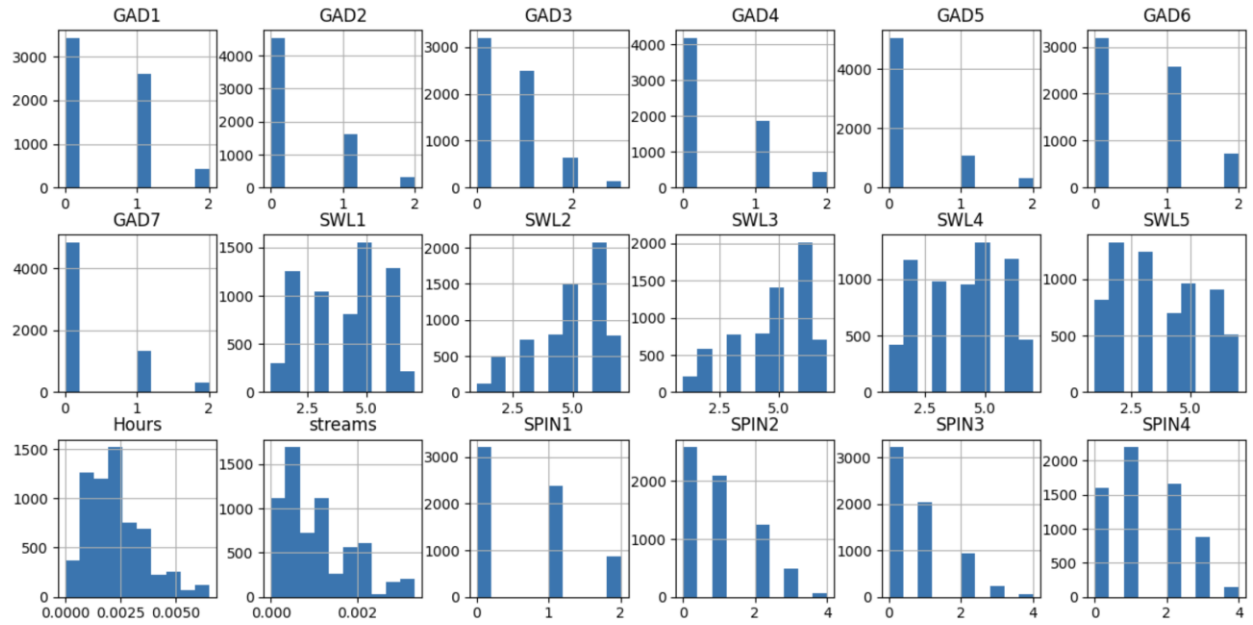
	Narcissism	Age	GAD_T	SWL_T	SPIN_T
count	6480.000000	6480.000000	6480.000000	6480.000000	6480.000000
mean	1.968981	0.079406	0.149809	0.550144	0.206663
std	1.001986	0.071123	0.130404	0.221685	0.127787
min	1.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	0.026316	0.047619	0.400000	0.102941
50%	2.000000	0.052632	0.142857	0.566667	0.191176
75%	3.000000	0.131579	0.238095	0.733333	0.279412
max	5.000000	0.289474	0.666667	1.000000	0.764706

[8 rows x 36 columns]

Histograms

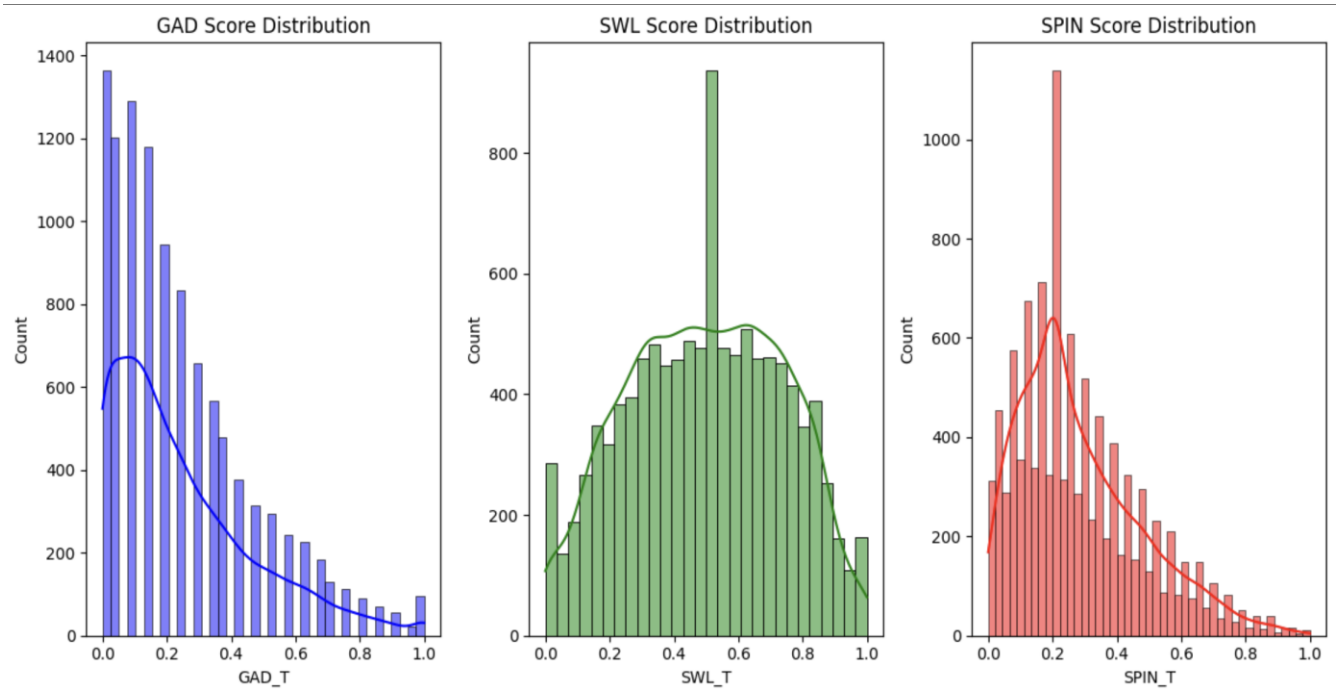
Histograms provide a visual representation of the distribution of numerical variables in the dataset. Here are some key insights from the histograms:

- The distribution of GAD scores appears to be positively skewed, with a majority of participants reporting low anxiety levels.
- Life satisfaction scores (SWL) exhibit a more symmetric distribution, with peaks at both ends of the scale, indicating varying levels of satisfaction among participants.
- Social phobia scores (SPIN) also show a right-skewed distribution, suggesting that most participants report low levels of social phobia.



Distribution of Scores

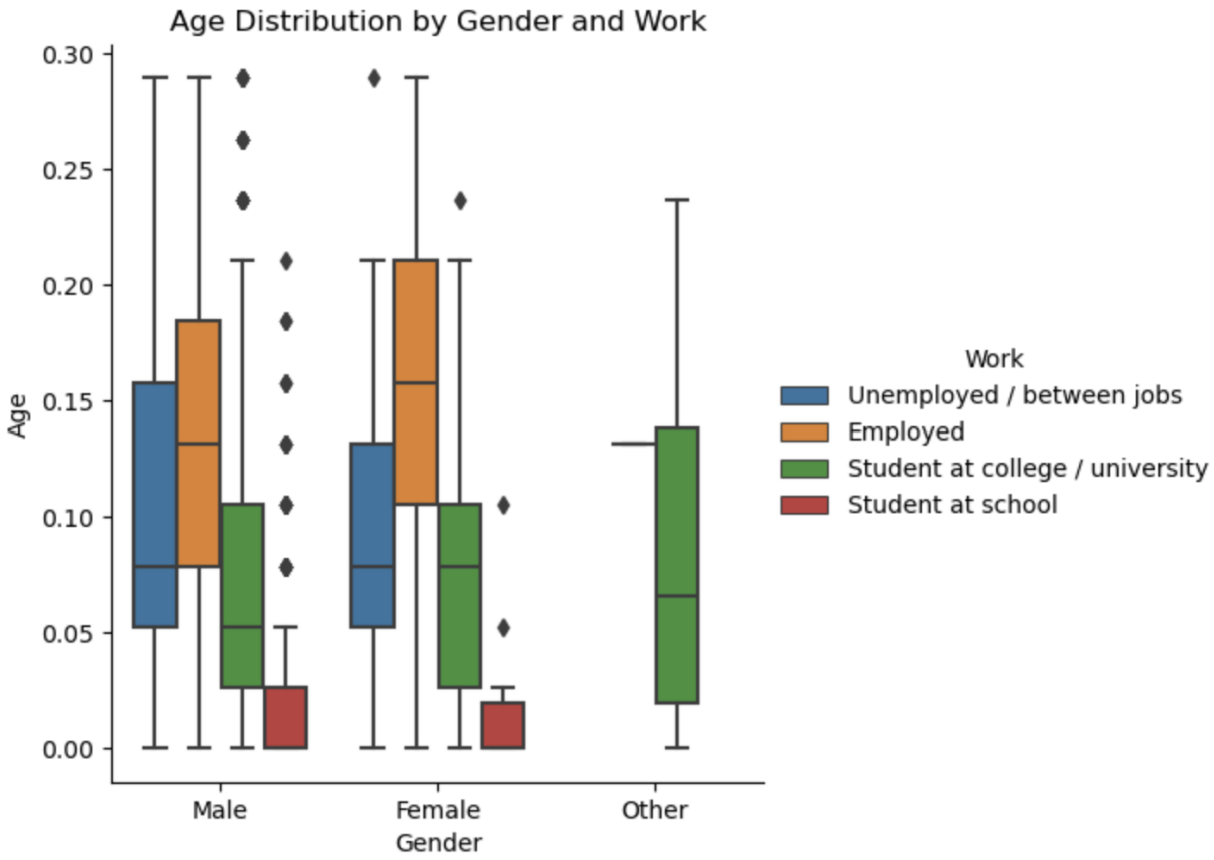
- The GAD score distribution exhibited a slightly positively skewed pattern, indicating varying levels of generalized anxiety disorder among participants.
- Conversely, the distribution of SWL scores appeared relatively symmetric, suggesting a diverse range of life satisfaction levels among the surveyed individuals.
- Similarly, the SPIN score distribution displayed a slight positive skewness, indicating varying degrees of social phobia reported by participants.



Age Distribution by Gender and Work

Examining the age distribution by gender and work status reveals:

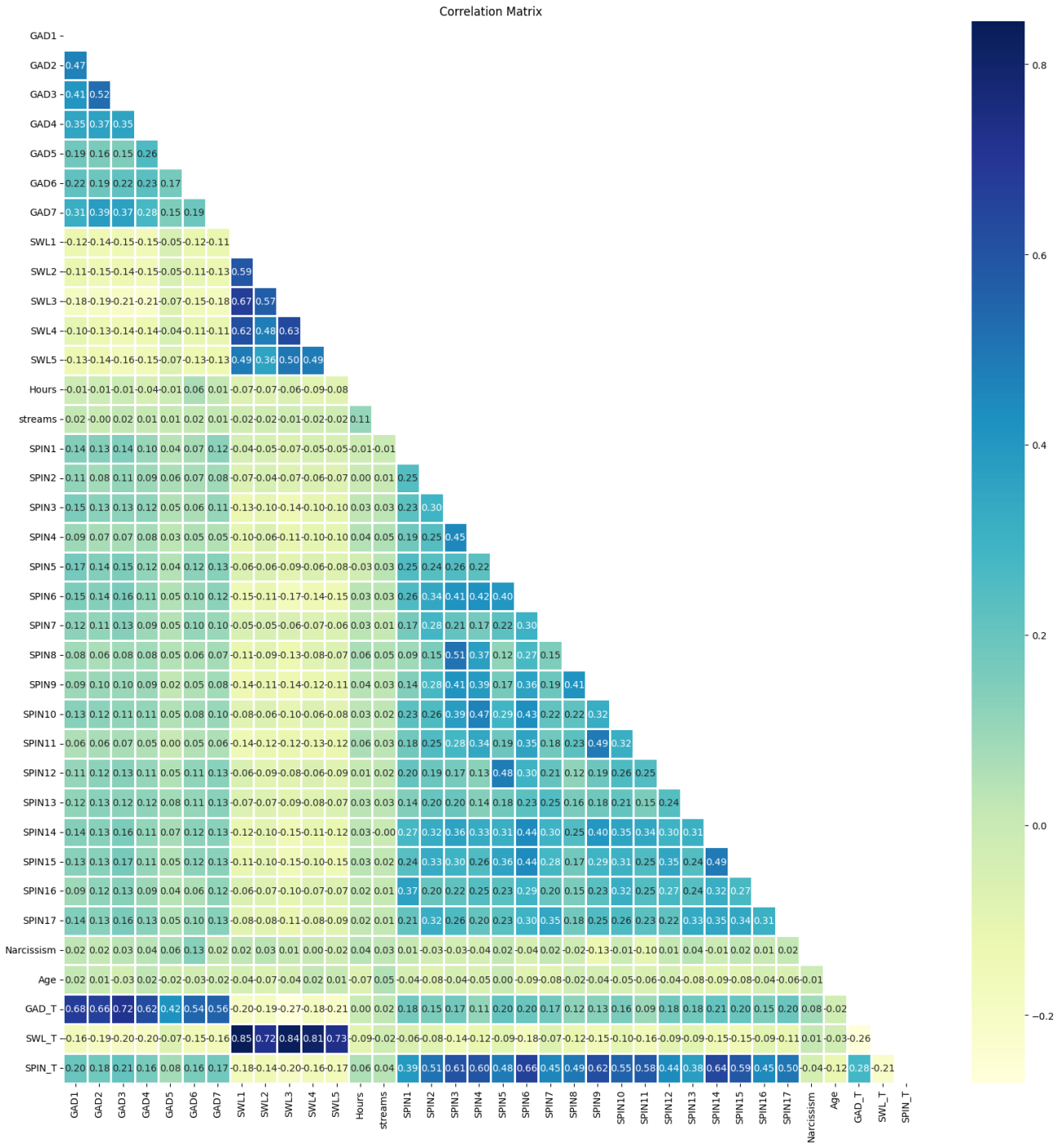
- The age distribution varies across different gender and work categories.
- Employed individuals tend to be slightly older on average compared to unemployed or student participants.
- Gender differences in age distribution are also evident, with male participants generally tending to be slightly older than female participants.



Correlation Heatmap

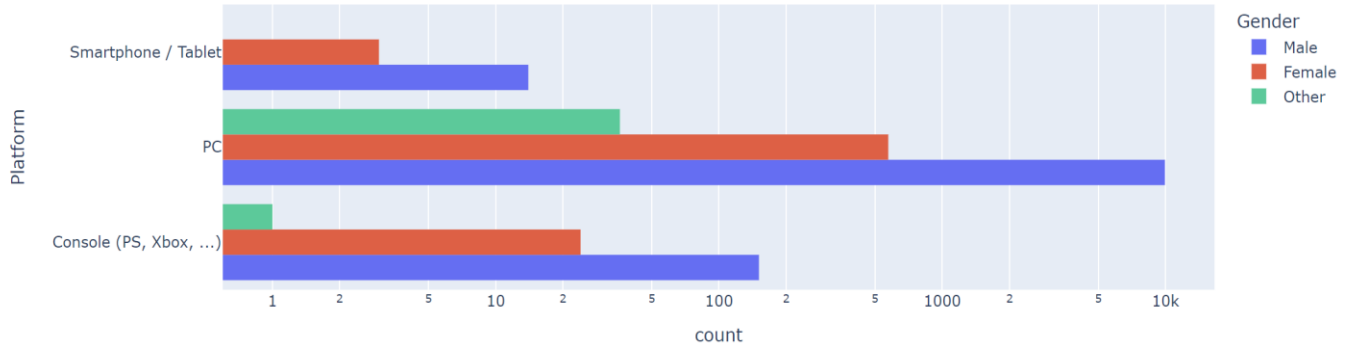
Understanding the correlation between numerical variables is crucial for identifying potential relationships. Insights from the correlation heatmap include:

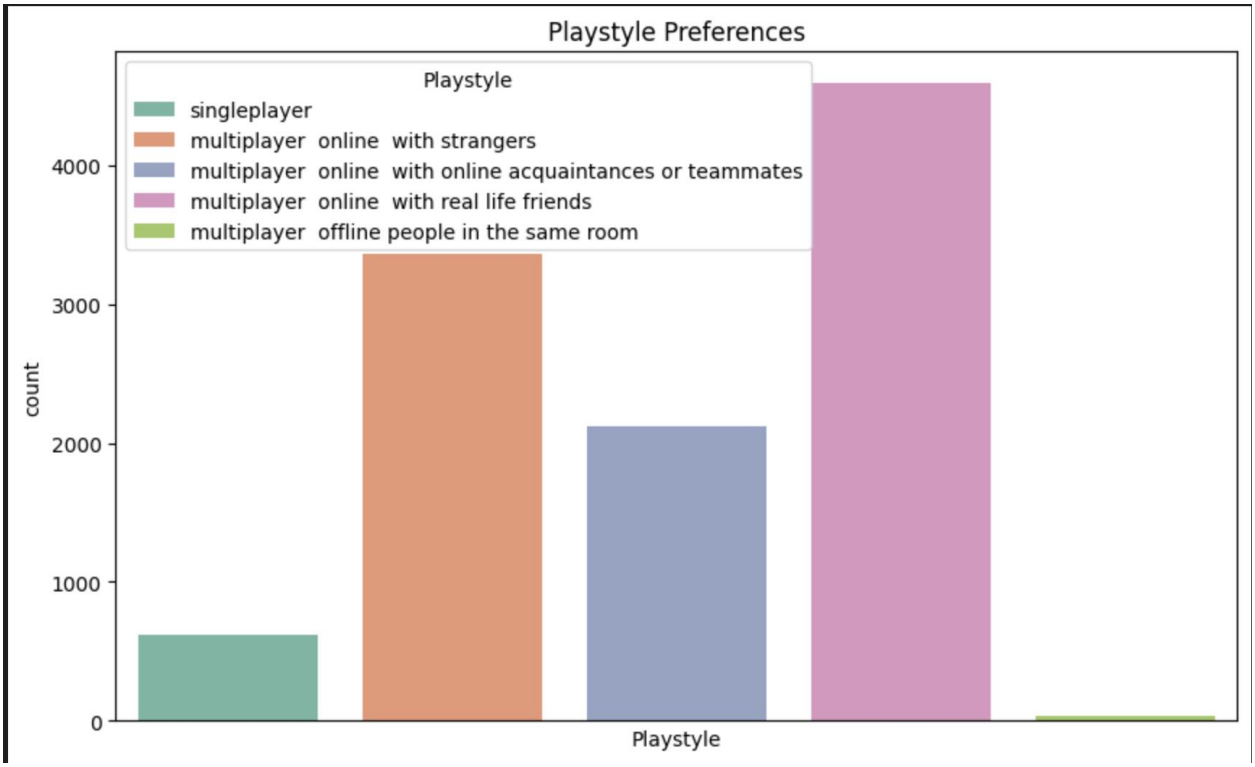
- There is a positive correlation between certain variables, such as GAD scores and SPIN scores, indicating a potential association between anxiety and social phobia.
- Other variables, such as age and narcissism levels, show weak or no correlation with anxiety, satisfaction with life, or social phobia.



Additional Insightful Plots

In addition to the distribution of scores, several other plots offer valuable insights into the dataset. These plots provide further understanding of the demographic and geographical characteristics of the surveyed individuals, shedding light on potential correlations and trends within the data.





Insights and Conclusion

The exploratory data analysis provides valuable insights into various aspects of the dataset, including the distribution of key variables, demographic characteristics, and potential relationships between variables. These insights lay the foundation for further analysis and modeling, guiding the development of hypotheses and research questions. Overall, the EDA process aids in understanding the underlying structure of the data and informs subsequent analytical decisions.

Model Training and Evaluation

Feature Engineering:

We used MinMaxScaler to normalize the Numerical features.

We have used Label encoding method to convert Categorical Columns to Numerical columns.

Before Label Encoding:

GADE	SWL1	SWL2	SWL3	SWL4	SWL5	Game	Platform	...	SPIN16	SPIN17	Narcissism	Gender	Age	Work	Degree
Not ficult at all	3	5	5	5	5	Skyrim	Console (PS, Xbox, ...)	...	1.0	0.0	1.0	Male	0.184211	Unemployed / between jobs	Bachelor (or equivalent)
Not ficult at all	3	5	3	3	3	Other	Console (PS, Xbox, ...)	...	0.0	0.0	2.0	Male	0.157895	Employed	Bachelor (or equivalent)
Not ficult at all	3	4	4	3	2	Other	PC	...	1.0	1.0	2.0	Male	0.289474	Employed	High school diploma (or equivalent)
newhat difficult	3	6	4	3	7	Other	Console (PS, Xbox, ...)	...	0.0	0.0	5.0	Female	0.131579	Employed	Bachelor (or equivalent)
newhat difficult	3	3	3	2	2	World of Warcraft	PC	...	0.0	2.0	1.0	Female	0.236842	Employed	High school diploma (or equivalent)

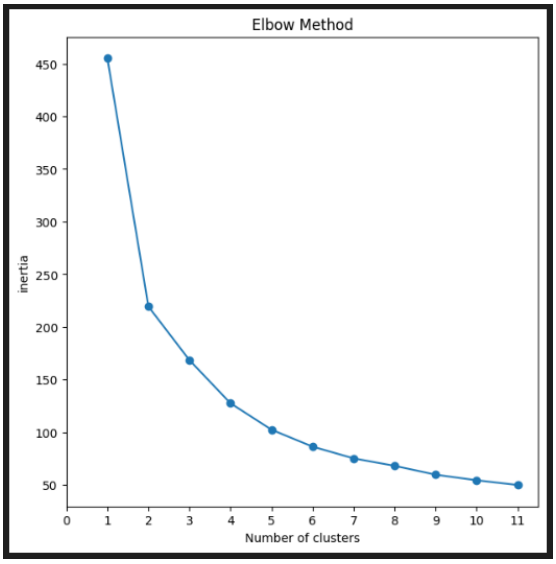
After Label Encoding:

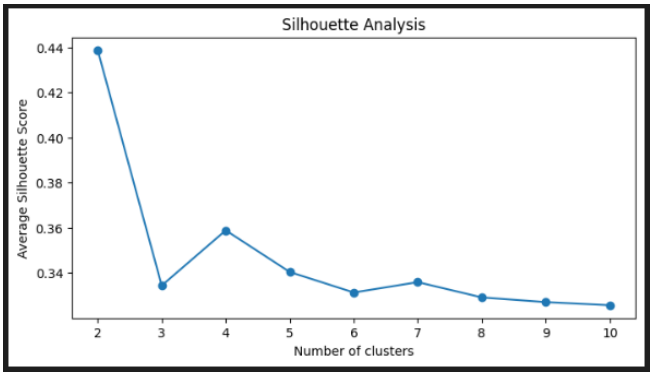
GADE	SWL1	SWL2	SWL3	SWL4	SWL5	Game	Platform	...	SPIN16	SPIN17	Narcissism	Gender	Age	Work	Degree	Residence
1	3	5	5	5	5	8	0	...	1.0	0.0	1.0	1	0.184211	3	0	88
1	3	5	3	3	3	7	0	...	0.0	0.0	2.0	1	0.157895	0	0	88
1	3	4	4	3	2	7	1	...	1.0	1.0	2.0	1	0.289474	0	1	88
2	3	6	4	3	7	7	0	...	0.0	0.0	5.0	0	0.131579	0	0	41
2	3	3	3	2	2	10	1	...	0.0	2.0	1.0	0	0.236842	0	1	25

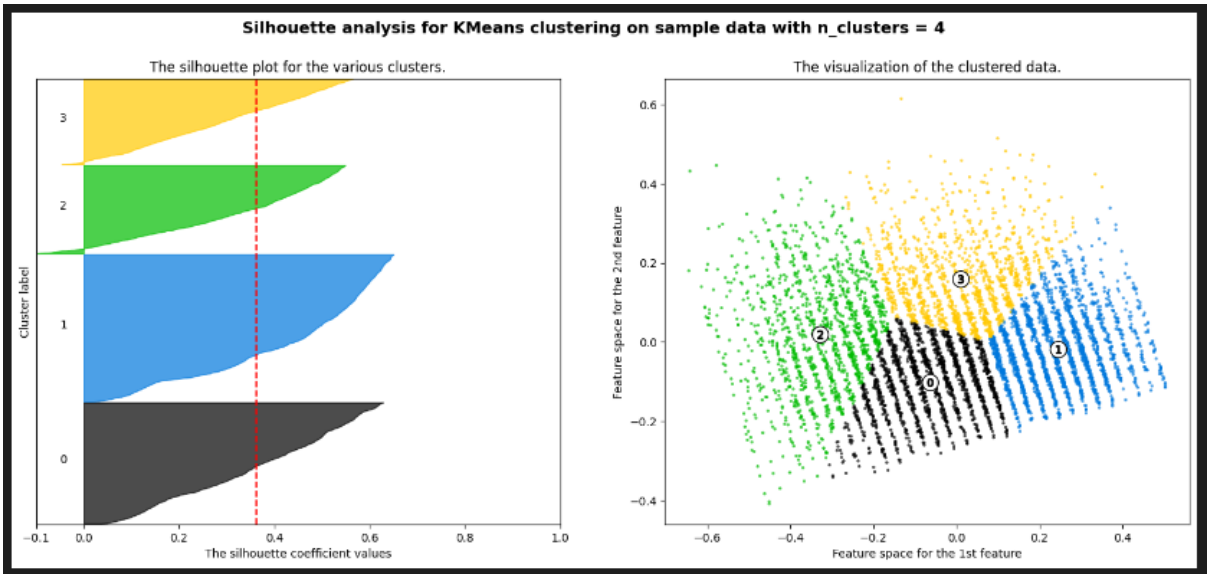
We performed Principal Component Analysis on the final features to analyse the data easily.

K-means Clustering:

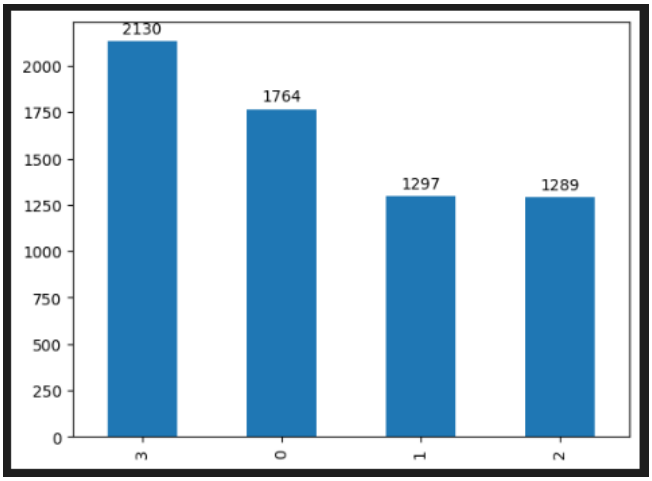
In the data that we have we do not have Labels. So, we performed Clustering to make clusters and use them as Labels. To find the how many clusters we need we used Elbow method and Silhouette Analysis.







Upon analysing the Inertia and Silhouette Score, it's evident that the ideal cluster count (K) stands at four. As the number of clusters increases, Inertia declines, underscoring the need for a balanced approach. The Silhouette Score, pivotal for assessing cluster distinctness, peaks at K=4, signifying well-defined cluster boundaries. Hence, we strongly advocate employing K=4 in the K-Means algorithm for this dataset.



Classification:

We then used Classification Algorithms to check the Model Performance. We have used 3 different classification algorithms (Random Forest, Logistic Regression and XgBoost).

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.936963	0.915966	0.926346	357.000000
1	0.811847	0.943320	0.872659	247.000000
2	0.964844	0.953668	0.959223	259.000000
3	0.992574	0.926097	0.958184	433.000000
accuracy	0.932099	0.932099	0.932099	0.932099
macro avg	0.926557	0.934763	0.929103	1296.000000
weighted avg	0.937269	0.932099	0.933322	1296.000000

Logistic Regression Classification Report:

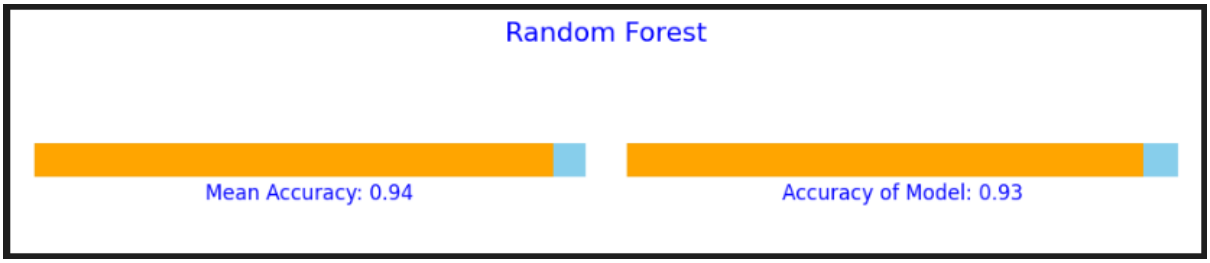
	precision	recall	f1-score	support
0	0.979943	0.985591	0.982759	347.000000
1	0.951220	0.978495	0.964664	279.000000
2	0.988281	0.976834	0.982524	259.000000
3	0.992574	0.975669	0.984049	411.000000
accuracy	0.979167	0.979167	0.979167	0.979167
macro avg	0.978004	0.979147	0.978499	1296.000000
weighted avg	0.979431	0.979167	0.979226	1296.000000

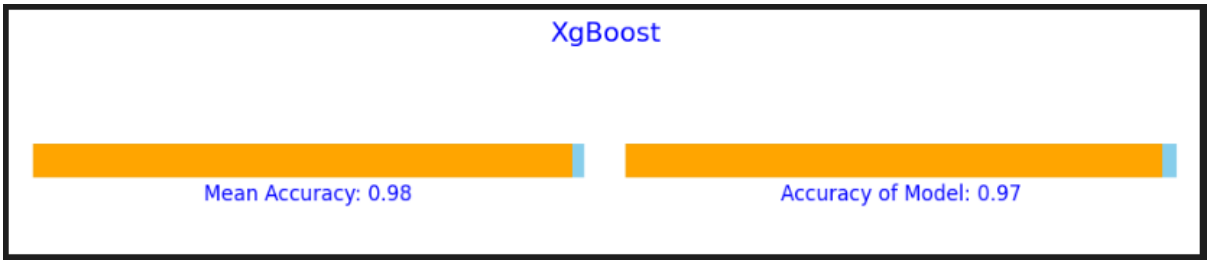
XGBoost Classification Report:

	precision	recall	f1-score	support
0	0.959885	0.976676	0.968208	343.000000
1	0.947735	0.957746	0.952715	284.000000
2	0.992188	0.973180	0.982592	261.000000
3	0.987624	0.977941	0.982759	408.000000
accuracy	0.972222	0.972222	0.972222	0.972222
macro avg	0.971858	0.971386	0.971568	1296.000000
weighted avg	0.972461	0.972222	0.972290	1296.000000

Cross Validation:

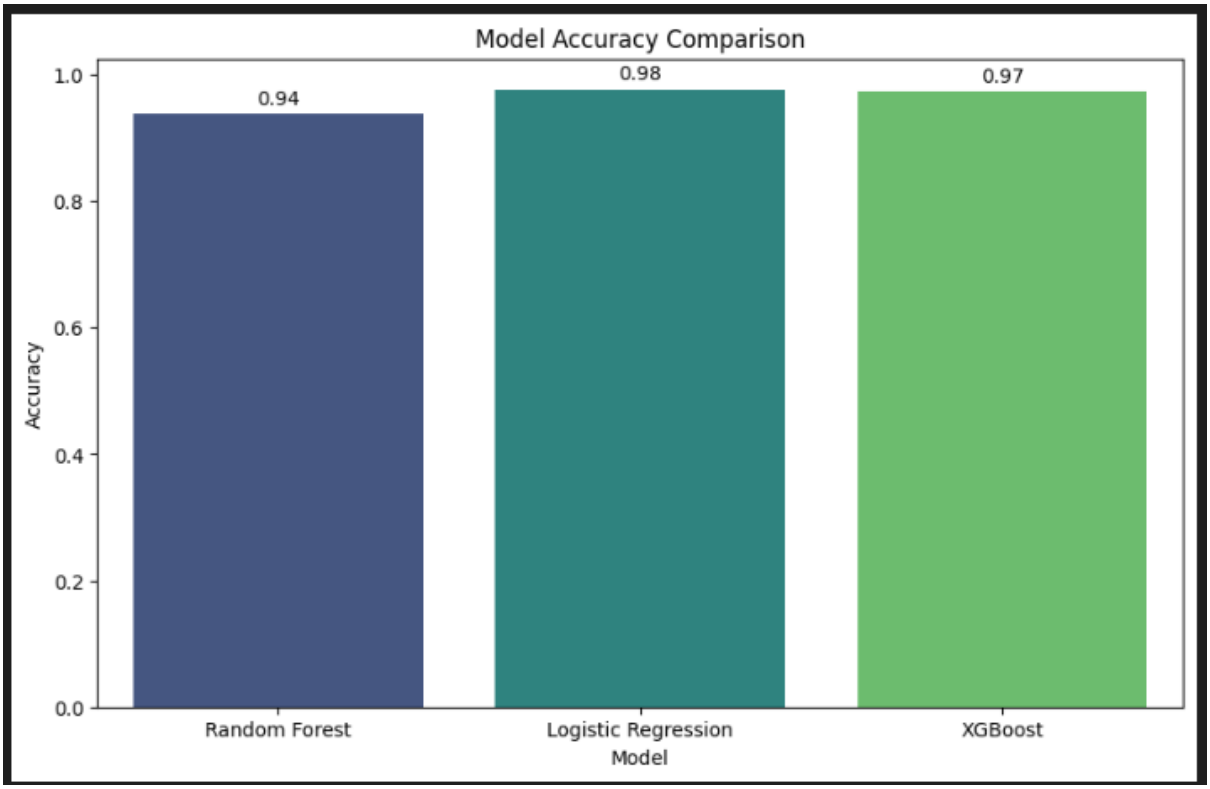
We employed K-Fold Cross-Validation on every model (with k=5) to ensure robust model evaluation. We plotted the mean accuracy scores from the CV to the Accuracy of the Models.





Conclusion:

By comparing the Accuracies and F1-scores we can say that Logistic regression algorithm performed well compared to all the models with a high accuracy of 98%.



Contribution Summary:

In our project, the work was distributed among three team members, each contributing to distinct phases of the project:

1. **Data Preprocessing and Outlier Detection:**
    - **Himanshu Aneja** was responsible for preparing the dataset for analysis. This included cleaning the data, handling missing values, and detecting and managing outliers to ensure the data's quality and reliability.
2. **Exploratory Data Analysis (EDA):**
    - **Vishal Rajashekar** conducted the Exploratory Data Analysis. This involved summarizing the main characteristics of the dataset, visualizing data distributions, and identifying patterns and correlations that provided valuable insights to guide the subsequent modelling phase.
3. **Model Training and Evaluation:**
    - **Nikhil Yakkala** focused on the model training and evaluation stage. This included developing and validating various machine learning models, such as clustering and classification algorithms. Additionally, this member performed cross-validation to assess model performance and ensure robust and reliable results.

Each member's contributions were integral to the successful completion of the project, ensuring a comprehensive approach from data preprocessing to model evaluation.