Progress Report 2

Capstone Project: Gaming Addiction and Mental Health

Course Instructor: Dr. Humera Noor Minhas

Submission By: Frank Stalin Dsouza

Date: 20-05-2025

During Weeks 3 and 4 of the project, I focused on cleaning the dataset, preparing new features, and training a machine learning model that could help identify people at risk of gaming addiction based on their behavior and mental health indicators.

1. Getting the Dataset Ready

Before doing any kind of analysis or modeling, I had to make sure the dataset was in a good shape. Think of it like clearing your workspace before starting a big project.

Here's what I did:

- **Cleaned Column Names**

Some column titles had extra spaces or inconsistent formatting. I fixed those so that everything was standardized and easier to work with.

- **Removed Duplicate Entries**

I checked for and removed any repeated rows to make sure every record represented a unique individual.

- **Converted Yes/No Answers to 1s and 0s**

Survey responses like "Yes" and "No" were turned into numbers. This helps the machine learning model understand them better. For example:

"Yes" → 1

"No" → 0

- **Ordered Categories for Social Withdrawal**

People described their social withdrawal using words like "Never," "Rarely," "Sometimes," and "Often." I converted those into a scale:

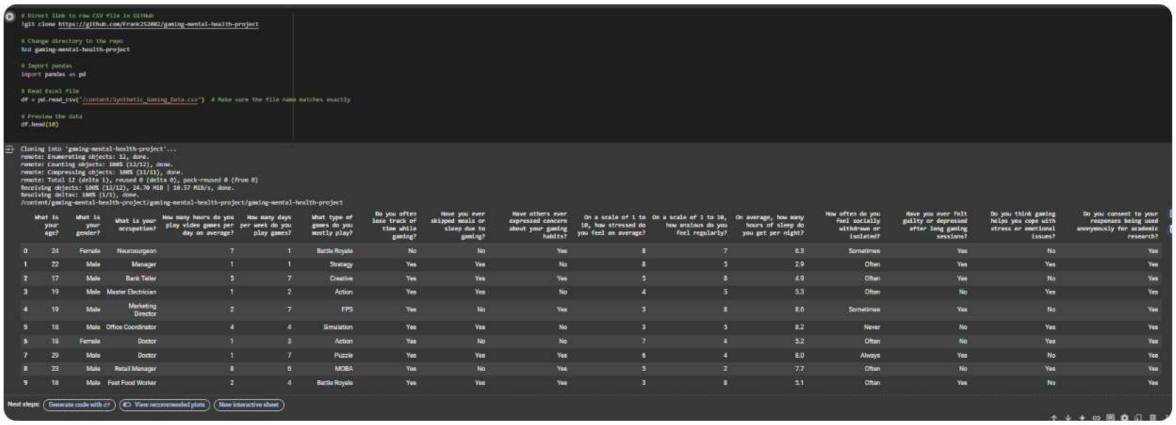
Never = 0

Rarely = 1

Sometimes = 2

Often = 3

This made it easier to treat social withdrawal as a measurable factor.



```
# Direct link to raw CSV file in GitHub
!git clone https://github.com/Frank252002/gaming-mental-health-project

# Change directory to the repo
%cd gaming-mental-health-project

# Import pandas
import pandas as pd

# Read Excel file
df = pd.read_csv('/content/Synthetic_Gaming_Data.csv')  # Make sure the file name matches exactly

# Preview the data
df.head(10)
```

```
Cloning into 'gaming-mental-health-project'...
remote: Enumerating objects: 12, done.
remote: Counting objects: 100% (12/12), done.
remote: Compressing objects: 100% (11/11), done.
remote: Total 12 (delta 1), reused 0 (delta 0), pack-reused 0 (from 0)
Receiving objects: 100% (12/12), 24.70 MiB | 10.57 MiB/s, done.
Resolving deltas: 100% (1/1), done.
/content/gaming-mental-health-project/gaming-mental-health-project/gaming-mental-health-project
```

	What is your age?	What is your gender?	What is your occupation?	How many hours do you play video games per day on average?	How many days per week do you play games?	What type of games do you mostly play?	Do you often lose track of time while gaming?	Have you ever skipped meals or sleep due to gaming?	Have others ever expressed concern about your gaming habits?	On a scale of 1 to 10, how stressed do you feel on average?	On a scale of 1 to 10, how anxious do you feel regularly?	On average, how many hours of sleep do you get per night?	How often do you feel socially withdrawn or isolated?	Have you ever felt guilty or depressed after long gaming sessions?	Do you think gaming helps you cope with stress or emotional issues?	Do you consent to your responses being used anonymously for academic research?
0	24	Female	Neurosurgeon	7	1	Battle Royale	No	No	Yes	8	7	6.3	Sometimes	Yes	No	Yes
1	22	Male	Manager	1	1	Strategy	Yes	Yes	No	8	5	2.9	Often	Yes	Yes	Yes
2	17	Male	Bank Teller	5	7	Creative	Yes	Yes	Yes	5	8	4.9	Often	Yes	No	Yes
3	19	Male	Master Electrician	1	2	Action	Yes	Yes	No	4	5	5.3	Often	No	Yes	Yes
4	19	Male	Marketing Director	2	7	FPS	Yes	No	Yes	3	8	8.6	Sometimes	Yes	No	Yes
5	18	Male	Office Coordinator	4	4	Simulation	Yes	Yes	No	3	5	8.2	Never	No	Yes	Yes
6	18	Female	Doctor	1	3	Action	Yes	No	No	7	4	5.2	Often	No	Yes	Yes
7	29	Male	Doctor	1	7	Puzzle	Yes	Yes	Yes	6	4	8.0	Always	Yes	No	Yes
8	23	Male	Retail Manager	8	6	MOBA	Yes	No	Yes	5	2	7.7	Often	No	Yes	Yes
9	18	Male	Fast Food Worker	2	4	Battle Royale	Yes	Yes	Yes	3	8	5.1	Often	Yes	No	Yes

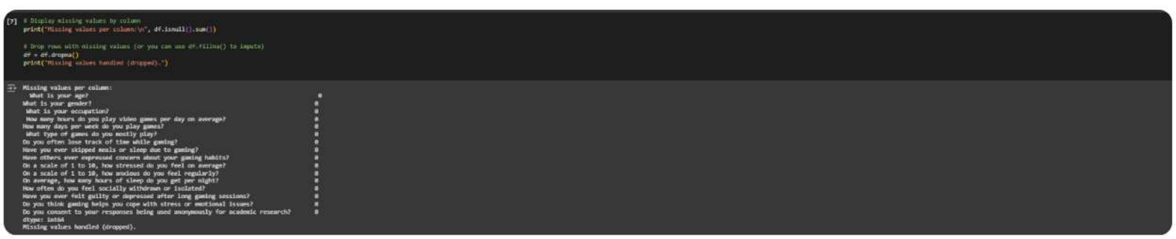
Next steps: Generate code with df | View recommended plots | New interactive sheet

```
# Count duplicates
print("Number of duplicate rows:", df.duplicated().sum())

# Drop duplicate rows if any
df = df.drop_duplicates()
print("Duplicates removed.")

df.to_csv("Synthetic_Gaming_Data.csv", index=False)
print("Cleaned dataset saved as 'Synthetic_Gaming_Data.csv'.")

df.head(10)
```

```
Number of duplicate rows: 0
Duplicates removed.
Cleaned dataset saved as 'Synthetic_Gaming_Data.csv'.
```

	What is your age?	What is your gender?	What is your occupation?	How many hours do you play video games per day on average?	How many days per week do you play games?	What type of games do you mostly play?	Do you often lose track of time while gaming?	Have you ever skipped meals or sleep due to gaming?	Have others ever expressed concern about your gaming habits?	On a scale of 1 to 10, how stressed do you feel on average?	On a scale of 1 to 10, how anxious do you feel regularly?	On average, how many hours of sleep do you get per night?	How often do you feel socially withdrawn or isolated?	Have you ever felt guilty or depressed after long gaming sessions?	Do you think gaming helps you cope with stress or emotional issues?	Do you consent to your responses being used anonymously for academic research?
0	24	Female	Neurosurgeon	7	1	Battle Royale	No	No	Yes	8	7	6.3	Sometimes	Yes	No	Yes
1	22	Male	Manager	1	1	Strategy	Yes	Yes	No	8	5	2.9	Often	Yes	Yes	Yes
2	17	Male	Bank Teller	5	7	Creative	Yes	Yes	Yes	5	8	4.9	Often	Yes	No	Yes
3	19	Male	Master Electrician	1	2	Action	Yes	Yes	No	4	5	5.3	Often	No	Yes	Yes
4	19	Male	Marketing Director	2	7	FPS	Yes	No	Yes	3	8	8.6	Sometimes	Yes	No	Yes
5	18	Male	Office Coordinator	4	4	Simulation	Yes	Yes	No	3	5	8.2	Never	No	Yes	Yes
6	18	Female	Doctor	1	3	Action	Yes	No	No	7	4	5.2	Often	No	Yes	Yes
7	29	Male	Doctor	1	7	Puzzle	Yes	Yes	Yes	6	4	8.0	Always	Yes	No	Yes
8	23	Male	Retail Manager	8	6	MOBA	Yes	No	Yes	5	2	7.7	Often	No	Yes	Yes
9	18	Male	Fast Food Worker	2	4	Battle Royale	Yes	Yes	Yes	3	8	5.1	Often	Yes	No	Yes

Next steps: Generate code with df | View recommended plots | New interactive sheet



```
[7] # Display missing values by column
print("Missing values per column:\n", df.isnull().sum())

# Drop rows with missing values (or you can use df.fillna() to impute)
df = df.dropna()
print("Missing values handled (dropped).")
```

```
Missing values per column:
  What is your age?                                                              0
What is your gender?                                                             0
 What is your occupation?                                                        0
 How many hours do you play video games per day on average?                      0
How many days per week do you play games?                                        0
 What type of games do you mostly play?                                          0
Do you often lose track of time while gaming?                                    0
Have you ever skipped meals or sleep due to gaming?                              0
Have others ever expressed concern about your gaming habits?                     0
On a scale of 1 to 10, how stressed do you feel on average?                      0
On a scale of 1 to 10, how anxious do you feel regularly?                        0
On average, how many hours of sleep do you get per night?                        0
How often do you feel socially withdrawn or isolated?                            0
Have you ever felt guilty or depressed after long gaming sessions?               0
Do you think gaming helps you cope with stress or emotional issues?              0
Do you consent to your responses being used anonymously for academic research?   0
dtype: int64
Missing values handled (dropped).
```

2. Creating New Features

We added new columns to help the model learn better patterns.

I counted how many risky behaviors a person had:

- Skipping meals or sleep due to gaming
- Losing track of time
- Feeling guilty after gaming
- Others showing concern about their habits

Each “yes” added 1 point, so the final score showed how serious the behavioral risks were.

- **Mental Burden**

This score was the sum of someone's stress and anxiety ratings. It helped measure their overall emotional load.

- **Sleep Deprivation**

If someone got less than 5 hours of sleep, I marked them as "sleep-deprived." It's a red flag that gaming might be affecting their daily life.

3. Defining the Risk Score

I created a custom scoring system to rate how "at risk" a person might be. The score went from 0 to 10, based on five behaviors. Each behavior added 2 points if it was present:

- Plays more than 5 hours a day
- Plays more than 5 days a week
- Skips sleep or meals due to gaming
- Feels guilty after gaming
- Loses track of time

This scoring system helped quantify the seriousness of someone's gaming habits.

4.Turning Scores into Risk Levels

After calculating the score, I grouped people into three categories:

- **Low Risk**: Score between 0 and 3
- **Medium Risk**: Score between 4 and 6
- **High Risk**: Score between 7 and 10

This made the results easier to understand. For example, someone with a score of 9 would clearly fall into the High Risk group.

5. Training the Model

Once everything was ready, I used a machine learning model called a Random Forest Regressor. I chose this model because:

- It works well with different kinds of data
- It's good at finding patterns
- It doesn't overfit easily (meaning it won't just memorize the data)

I trained the model to predict the risk score based on everything I'd prepared (gaming habits, stress, sleep, etc.).

6. Evaluating the Model

To check how well the model was performing, I tested it using a portion of the data it hadn't seen before. The results were very encouraging:

- **Mean Absolute Error (MAE):** The average prediction error was only 0.08 points, which is very low.
- **R^2 Score:** The model explained about 99% of the variation in the risk scores.

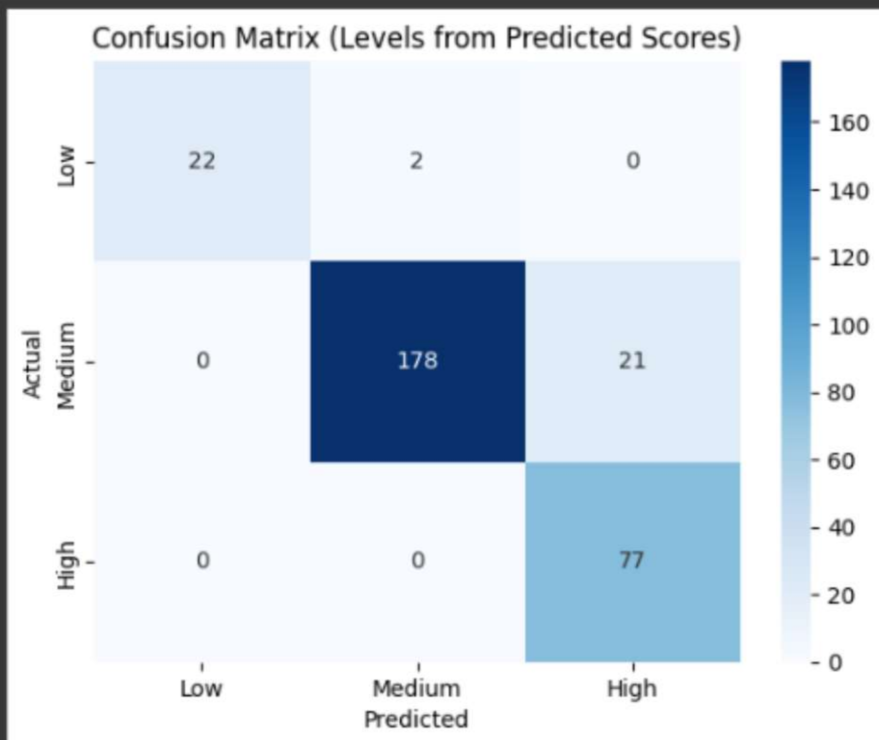
This means the model was very accurate and didn't seem to be overfitting — it worked well on both training and test data.

7. Converting Predictions Back into Risk Levels

After the model predicted a risk score for each person, I converted those scores back into Low, Medium, and High categories (using the same thresholds I created earlier).

Then I compared the predicted categories to the actual ones. The model was correct 92% of the time, and most of the mistakes were between neighboring levels like Medium and High — which is acceptable.

```
Regression Evaluation:
MAE: 0.07619999999999998
RMSE: 0.21812228985899934
R²: 0.9877456614313842

Classification Evaluation (Based on Predicted Scores):
              precision    recall  f1-score   support

        High       0.79      1.00      0.88        77
         Low       1.00      0.92      0.96        24
      Medium       0.99      0.89      0.94       199

    accuracy                           0.92       300
   macro avg       0.92      0.94      0.93       300
weighted avg       0.94      0.92      0.93       300
```




```
Overfitting Check:
Train R²: 0.999 | Test R²: 0.988
```