GROUP MEMBERS:

Priyanka Singh - 201774818
Sathyasri Sudhakar - 201749908
Omkar Pawar - 201694894
Akshay Deshmukh - 201773407

INTRODUCTION:

The purpose of this report is to create a more nuanced understanding of the interactions between socio-demographic characteristics, in-game behaviours, and global-scale environmental consciousness. The dataset used contains a wide range of information, including socio-demographic details, responses to world events, environmental viewpoints, gaming habits, in-game activities, and player emotional experiences. This dataset is a significant resource for investigating the global convergence of virtual gaming, environmental awareness, and human actions. Our methodology is divided into three stages: data quality, data characterization, and detailed analysis objectives. The aim of data quality assessment is to ensure the accuracy, consistency, and reliability of the dataset by identifying and rectifying discrepancies. In parallel, the data characterization phase seeks to provide a comprehensive overview of key characteristics, patterns, and trends within the dataset. The detailed analysis is for conducting an in-depth examination of the dataset, extracting valuable insights and conclusions.

DATA QUALITY:

1. Inconsistencies in Spelling and Representation:
   - There are discrepancies in spelling and representation of nationality names in column A1_1, exemplified by variations such as 'American,' 'American (USA),' and 'American US.'
   - Typos are present in nationality names, evident in instances like 'Vieynam' instead of the correct 'Vietnam.'
   - The usage of abbreviations like 'uk' instead of 'UK' introduces inconsistency in the dataset.
   - Various representations for the same nationality contribute to redundancy and potential confusion.

2. Contextually Irrelevant Entries:
   - Column A1_1 contains values like ‘29’ and ‘friendly’ that do not align with the context of nationality, including entries that are either invalid or unexpected.
   - Column D7 features values like '??' and 'M,' which could serve as placeholders or indicate erroneous entries.

3. Inconsistencies in Age Representation:
   - Age inconsistencies are observed in column A5, with representations such as '30s' and 'sub 28'.(Mention that age is numeric data, but mentioned as object type)

4. Incorrect Datatypes:
   - The data type for A5(Age) and п..O1 was incorrect, both being of object data type. A5 should be of numeric data type, and п..O1 should be of datetime data type.

5. Missing Values:
   - Columns D1, D2, D3, and D7 exhibit missing values:
   - D1 has 6 missing values.
   - D2 has 5 missing values.
   - D3 has 1 missing value.
   - D7 has 6 missing values.

6. Timestamp Column Encoding:

- The timestamp column, named 'ï..O1,' is not UTF-8 encoded and is instead in Latin-1, potentially causing complications during the CSV file import process.

7. Data Organization Concerns:
   - The index row is unsorted, posing potential challenges for data retrieval and analysis. Sorting is recommended to enhance organization and coherence.

DETAILED ANALYSIS:

A. Exploratory Data Analysis

1. Age distribution of the players:

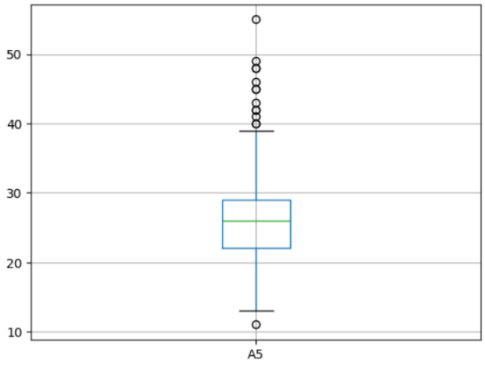


Figure 1: Boxplot for Age distribution

The box plot illustrates the distribution of player ages, ranging from a minimum of 11 to a maximum of 55. Outliers are evident showcasing values beyond the typical range. The first quartile, representing the lower 25% of the data, is at 22, while the median (50th percentile) falls at 25.5. The third quartile, encapsulating the lower 75% of the data, is situated at 29.

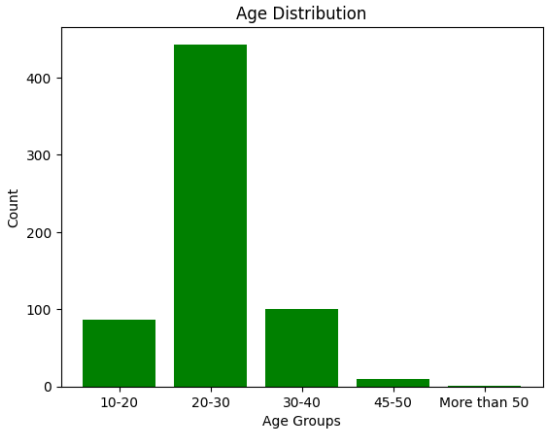


Figure 2 : Histogram for Age groups

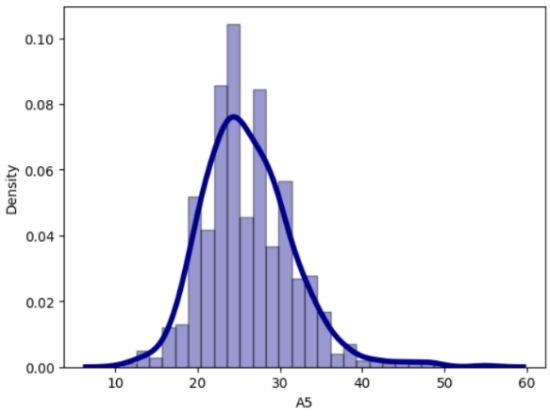
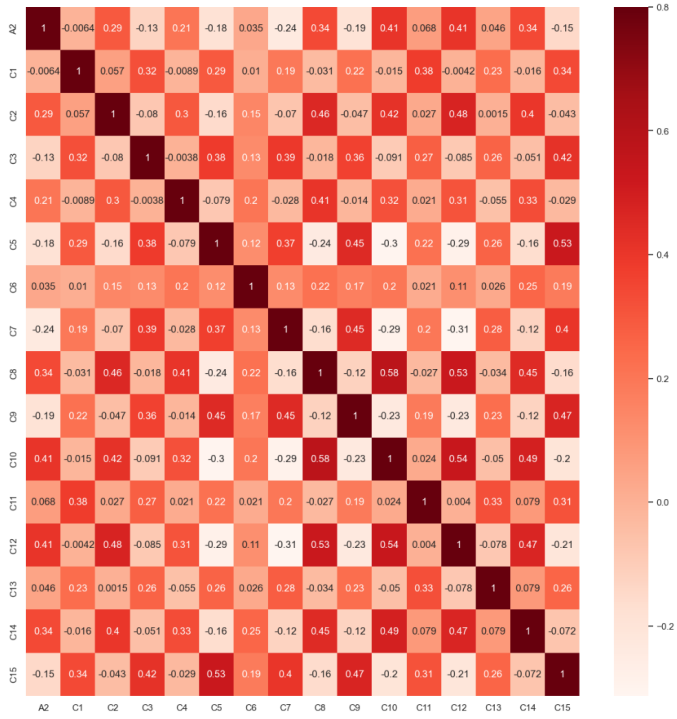


Figure 3: Age Distribution

The graphs suggest that the data is approximately normally distributed as indicated by the bell-shaped curve. It shows that the majority of the players are between the ages of 20 and 30, with the highest density at around 25 years old. The density decreases as the age increases, with very few players above the age of 50. This suggests that the video game is more popular among younger players.

2. **The relationship between the biological sex as defined by the dataset and the players' environmental perception:**

We conducted a correlation analysis between the gender feature and all environmental perception variables to determine to what extent the gender is linearly related to environmental perception. The following heatmap visualises the correlation of gender with each of the environmental perception variable:



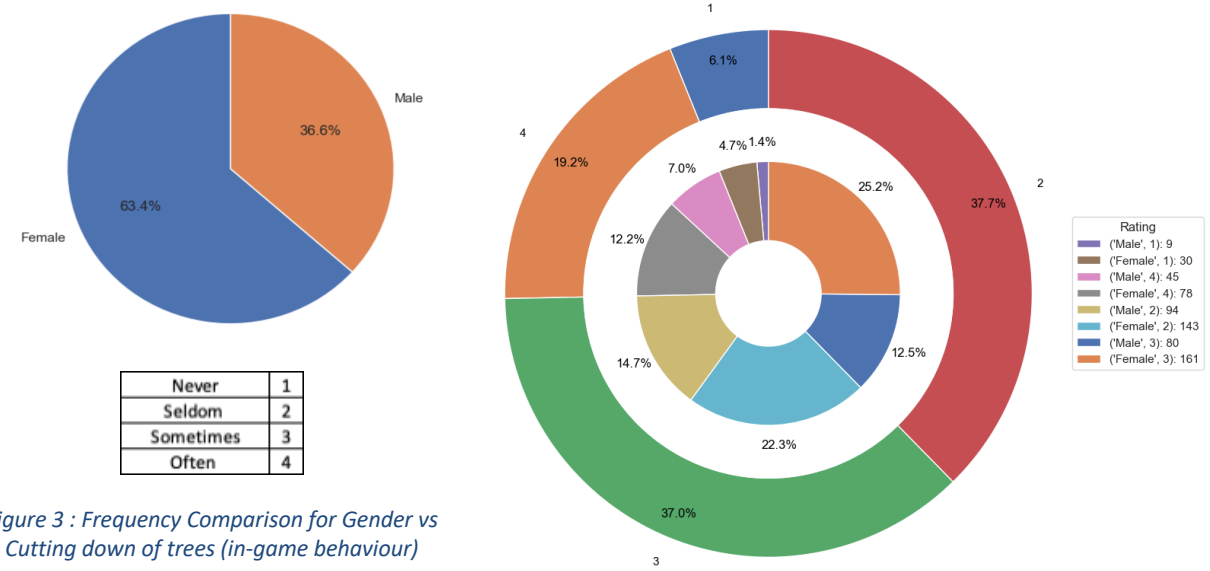
Environmental Perception Variable	Correlation Coefficient
C1	-0.0063
C2	0.2948
C3	-0.1268
C4	0.2071
C5	-0.1836
C6	0.0347
C7	-0.2353
C8	0.3425
C9	-0.1904
C10	0.4105
C11	0.0681
C12	0.4099
C13	-0.0464
C14	0.3422
C15	-0.1455

Figure 2: Heatmap for Age vs Environmental Perception Variables

Analyzing the correlation coefficients reveals that, overall, there is not a notably strong relationship between a player's biological sex and their environmental perception.

3. **A comparison of the frequency of the male and female players' in-game behaviour (cutting down the tree) :**

The dataset does not have equal samples for male and female players. The distribution of players based on their biological sex is as follows:



Never	1
Seldom	2
Sometimes	3
Often	4

Figure 3 : Frequency Comparison for Gender vs Cutting down of trees (in-game behaviour)

We utilized a nested pie chart to illustrate the distribution of responses from both men and women regarding in-game behavior. According to the chart, it seems that women are more inclined to cut down a tree compared to men. The majority of women indicated they would either seldom or sometimes cut a tree, with the highest percentage choosing 'sometimes.' In contrast, the highest percentage of men are likely to ‘seldom’ cut down a tree. Those who answered 'Never' comprised 1.4% male and 4.7% female, while those who answered 'Often' included 7% male and 12.2% female.

B. Identifying the most important socio-demographic variables to indicate the environmental perception of the players

We employed four distinct methodologies to discern the socio-demographic variables indicative of players' environmental perception:

- Correlation Matrix:

Utilizing a heatmap, we visually examined the correlation between socio-demographic and environmental perception variables. Looking at the matrix, we cannot clearly say which is the best feature.

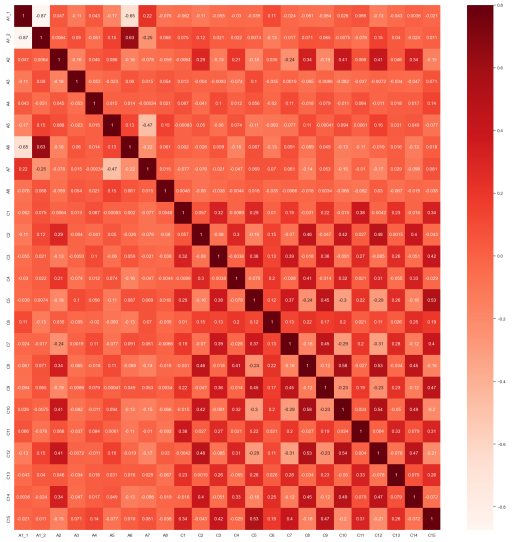


Figure 4 : Correlation Matrix plot

- Information Gain:

Information gain guided the prioritization of socio-demographic attributes based on environmental perception. Information Gain measures how well a feature separates the data into different classes. The higher the Information Gain, the more relevant the feature is in predicting the target variable. The table displays the weighted proportional entropy for each feature when splitting on all values, revealing Age and Nationality as possessing the highest weighted proportional entropy.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
Nationality	0.09	0.1	0.07	0.13	0.08	0.15	0.11	0.13	0.09	0.11	0.12	0.13	0.09	0.11	0.07
Region	0.02	0.02	0.02	0.03	0.04	0.04	0.02	0.04	0.02	0.03	0.03	0.06	0.02	0.02	0.02
Gender	0.01	0.07	0.02	0.03	0.03	0.01	0.05	0.09	0.03	0.13	0.01	0.13	0.01	0.09	0.02
Education Level	0.02	0.02	0.03	0.02	0.02	0.02	0.03	0.03	0.01	0.02	0.03	0.01	0.01	0.01	0.02
Pet or Garden	0.02	0.04	0.03	0.03	0.02	0.01	0.03	0.04	0.02	0.03	0.01	0.04	0.01	0.02	0.02
Age	0.14	0.18	0.15	0.18	0.14	0.17	0.13	0.23	0.13	0.18	0.14	0.21	0.12	0.19	0.12
Ethnicity	0.03	0.04	0.02	0.07	0.04	0.05	0.03	0.05	0.02	0.06	0.05	0.08	0.02	0.07	0.04
Marital Status	0.02	0.05	0.01	0.01	0.03	0.02	0.02	0.05	0.01	0.04	0.01	0.05	0.02	0.04	0.03
Employment Status	0.03	0.06	0.03	0.04	0.05	0.03	0.03	0.04	0.02	0.07	0.06	0.05	0.05	0.04	0.03

Figure 5 : Different values of Information gain based on Socio-demographic variables.

- Feature Selection using Random Forest:

Random Forest's embedded method for feature selection involved constructing 4 to 12 hundred decision trees, each built on a random subset of observations and features. This approach ensured that trees remained decorrelated, mitigating the risk of overfitting. The identified influential features, in order, are:

Age,
Do you have a pet or garden?,
Employment status,
Nationality,
Highest education level.

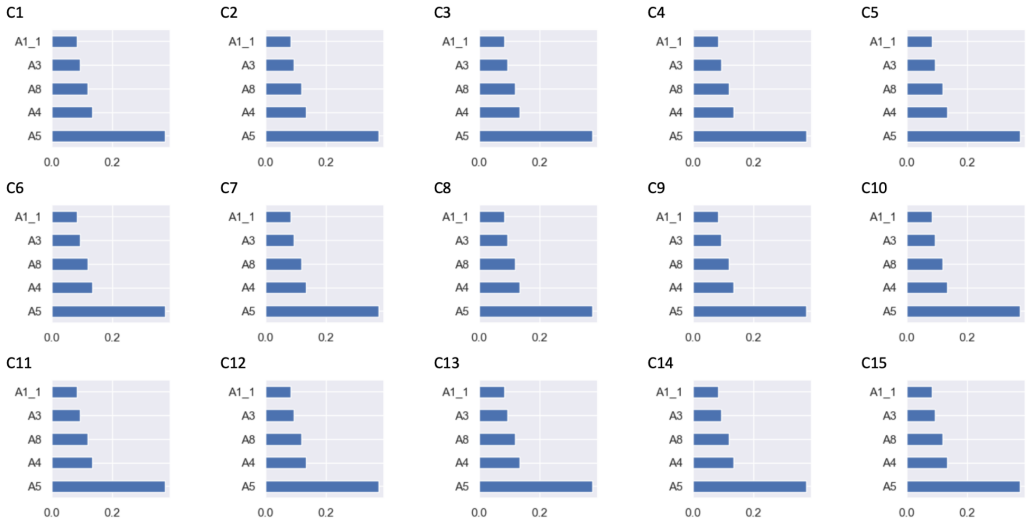


Figure 6 : Plots for different C columns and the best features selection for them.

C. Classification model to predict a player's environmental perception based on socio-demographic variables

In the exploration of our dataset, various issues were identified, as detailed in the Data Quality section. Prior to proceeding with splitting the data and classification model development, our initial task involved data transformation to rectify these discrepancies.

Our first focus is on the Nationality column, which is A1_1. Where multiple inconsistencies in the spelling, representation, and irrelevant entries not pertaining to nationality were encountered.

To address this, we systematically addressed each region, creating lists for various representations of the same country and standardizing them. For example:

```
#US/Canada Region
list1=['Canadian',     'Canadian     ',     'Candadian','Portuguese-
Canadian','canadian']
for i in range(0,len(df_new['A1_1'])):
    if df_new["A1_2"][i]=="US/Canada" and ( df_new["A1_1"][i] not in
list1):
        df_new["A1_1"][i]="American"
    elif df_new["A1_2"][i]=="US/Canada" and ( df_new["A1_1"][i] in
list1):
        df_new["A1_1"][i]="Canadian"
```

Similar procedures were applied to other regions.

Regarding irrelevant entries such as "29" or "Friendly" in the Nationality column, we grouped the A1_1 and A6 columns to determine Nationality based on their Ethnicity. Upon examining the grouped data, it was observed that the individual who had inputted "29" had identified their ethnicity as white. Given that the majority of individuals with the same ethnicity were American, we subsequently updated the nationality from "29" to "American." Similarly, a similar adjustment was made for the entry "Friendly," as the person's ethnicity was identified as Asian, and the majority of Asians in the grouped data were found to be from the Philippines.

In the case of an individual mentioning their nationality as "Asian," a review of the grouped dataset revealed that the predominant nationality for Asians was Filipino. Consequently, we opted to modify the nationality entry to "Filipino."

We also noticed inconsistencies in the age representation, and that the datatype was object instead of numeric. We corrected the inconsistencies by:
1. Replacing “sub 28” with 28
2. Replacing “30s” with a mid value of 35

Post corrections, the Age column – A5 was converted to numeric form.

Unnecessary columns were dropped and the resultant dataset is as follows:

	A1_1	A1_2	A2	A3	A4	A5	A6	A7	A8	C1	...	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
0	Filipino	Asia	Male	Secondary school	A pet	11	Asian	Single, never married	A student	4	...	2	4	3	3	4	5	5	3	4	4
1	American	US/Canada	Male	Secondary school	Both	13	White	Single, never married	A student	4	...	3	5	5	5	5	5	5	5	3	5
2	American	US/Canada	Male	Secondary school	Both	13	White	Single, never married	A student	5	...	3	5	5	5	5	5	5	5	3	5
3	Filipino	Asia	Female	High school	A pet	13	Asian	Single, never married	A student	4	...	1	5	4	4	3	5	5	5	2	5
4	American	US/Canada	Male	Secondary school	A pet	14	Hispanic or Latino	Single, never married	A student	3	...	1	3	4	5	3	3	4	4	2	4

The dataset has 22 unique values for Nationality:

```
In [26]: df_new['A1_1'].unique()

Out[26]: array(['Filipino', 'American', 'British', 'Vietnamese', 'Canadian',
                'European', 'New Zealander', 'Singaporean', 'Australian',
                'Mixed Nationalities', 'South Korean', 'Hispanic ', 'Burmese',
                'Chinese', 'Indonesian', 'Caucasian ', 'English ', 'Japanese',
                'Ashenkazi Jewish', 'Latina', 'Colombian', 'Argentina'],
               dtype=object)
```

Hence, nationalities were categorized to facilitate model training and testing using the following code:
`df_new['A1_1']=df_new['A1_1'].astype('category').cat.codes`

Similar procedures were performed for all columns to make training of the classifier model easier.

We group the environmental perception questions (C1 - C15) based on their positive and negative outlook to understand a person's environmental perception:
Positive questions – C1, C3, C5, C7, C9, C11, C13, C15
Negative questions – C2, C4, C6, C8, C10, C12, C14

Examining C2, which asks, "How much do you agree with the following statements? [Humans have the right to modify the natural environment to suit their needs]," reveals that a majority of respondents provided responses exceeding 3. This suggests a disagreement with the statement, positioning them on the positive side as individuals who care about the environment.

In summary, our dataset has been categorized into Socio-demographic variables (A1_1 to A8), Positive questions (C1 to C15, with values ranging from 1 to 5 where 5 corresponds to Strongly Agree and 1 corresponds to Strongly Disagree), and Negative Questions (C2 to C14, valued from 1 to 5 where 5

represents Strongly Disagree and 1 represents Strongly Agree). Additionally, a new column was introduced by calculating the rounded mean of responses to Positive and Negative questions. The individual's environmental perception was then determined by the mid-value between the rounded mean of Positive questions and the mean of Negative questions.

Categories were established for environmental concern to incorporate into the Categorical Perspective column:

- Category A – Cares about the Environment: Mid value falls between 4 and 5.
- Category B – Neutral: Mid value falls between 3 and 4.
- Category C – Does not care about the Environment: Mid value falls below 3.

To establish this categorization, we have referenced the following table:

Positive_mean	5	5	5	5	5	4	4	4	4	4	3	3	3	3	3	2	2	2	2	2	1	1	1	1	1
Negative_mean	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Mid_value	3	4	4	5	5	3	3	4	4	5	2	3	3	4	4	2	2	3	3	4	1	2	2	3	3
Categorical Perspective	B	B	A	A	A	C	B	B	A	A	C	C	B	B	A	C	C	C	B	B	C	C	C	C	B

Once more, we categorised the Categorical Perspective column within our dataset for analysis. Where A was categorised as 0, B as 1 and C as 2.

The transformed dataset was then split into training (70%), testing(15%), and evaluation(15%) sets. This was accomplished by using the train_test_split method provided by the sklearn library. The X variable was defined as the dataframe containing socio-demographic variables from columns A1 to A8, while the y variable encompassed two columns— one containing the mean of all columns C1 to C15 and the Categorical Perspective column.

Initially, the dataset was divided into 70% and 30%. The 70% portion was allocated to the training dataset (X_train, y_train), while the remaining 30% was assigned to temporary variables (X_temp and y_temp). Subsequently, these temporary variables were once more split into equal halves of 50%, resulting in X_test, y_test (representing 15% of the entire dataset), and X_eval, y_eval (constituting another 15% of the complete dataset).

Next, the training dataset was applied to a MultiOutputClassifier that consisted of various other classifiers such as RandomForestClassifier, DecisionTreeClassifier, Support Vector Machines. This step was done to evaluate which model most accurately fits the data.

Given our dataset's multiple inputs and outputs, the initial model selection leaned towards decision trees. However, ensemble classifiers seemed like a more favourable choice due to the number of labels/classes in the y dataset. The classifier we decided on was the Random Forest Classifier. This ensemble method combines the predictions of many different individual models (Decision Tree Classifiers) to improve the generalization and the overall performance. Random forest tends to see more unseen data, and it combines various trees reducing the chances of overfitting the model.

```
MultiOutputClassifier(RandomForestClassifier(n_estimators=10,
random_state=42))
```

After applying the data to the aforementioned model, it was determined that the Random Forest achieved the highest accuracy, reaching a value of 0.63. The accuracy was computed individually for each column and subsequently averaged.

We visually represent the performance of your prediction model by using a confusion matrix:

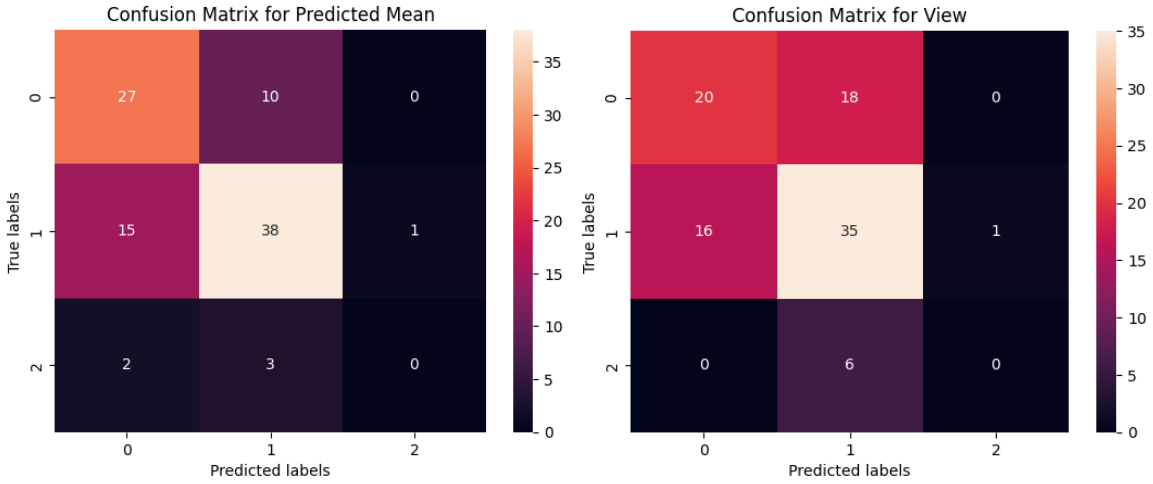


Figure 7 : Confusion Matrix for Multi-Output Classifier

There were virtually no instances of false negatives for both columns, suggesting that the model provides accurate answers approximately 63% of the time.

We evaluated the model's ability to capture all relevant positive instances by examining the recall score, which was computed as 0.61. Subsequently, we found the accuracy of the positive predictions through the precision score, yielding a value of 0.59. The F1_score for our model was determined to be 0.612, suggesting that the model strikes a reasonable balance between precision and recall but does not excel in either aspect.

CONCLUSIONS:

In conclusion, our report meticulously addressed data quality concerns and employed various analytical techniques to derive meaningful insights. The exploratory data analysis illuminated a predominant age group of 20-30 among the players. The correlation analysis indicated that gender, overall, does not significantly influence the environmental perception of players. Nevertheless, when assessing the likelihood of cutting down a tree in the game, women exhibited a slightly higher preference. Utilizing correlation analysis, information gain, and random forest methods, we identified key socio-demographic variables that strongly correlate with players' environmental perception. These variables include Age, Pet or Garden Ownership, Employment Status, Nationality, and Highest Education Level.

These critical factors were integrated into a multi-output classification model, which successfully predicted environmental perception of a player with an accuracy rate of 63%. This comprehensive analysis provides valuable insights into the nuanced relationship between socio-demographic variables and players' environmental perceptions in the context of the game.