Coursework 1: Understanding Players' Environmental Perception in a game Environment

Group Members

1. Shubham Dubey (Student ID #201705949, Email: sc23sd@leeds.ac.uk)
2. Rishabh Bezbaruah (Student ID #201777433, Email: mm23rb@leeds.ac.uk)
3. Reshma Ruby Nediyakalayil (Student ID #201782385, Email: sc23rrn@leeds.ac.uk)
4. Aishwarya Patil (Student ID #201750917, Email: mm23abp@leeds.ac.uk)

Section 1: Introduction

The report examined a dataset comprising of responses collated through an online survey conducted among players of Nintendo's Animal Crossing: New Horizons (ACNH) game and aimed to derive significant insights into the in-game attitudes of players with their environmental perspectives. The survey, which took place from the 15th to the 30th of May 2020, leveraged the online communities comprising of ACNH players from several social media platforms. The dataset under consideration has six major categories of the data, from its 640 surveyed respondents from 29 countries, namely, Socio-demographic Profile, COVID-19 Concern, Environmental Perception, Game-Playing Habit, In-Game Behaviour, and Game-Playing Feeling. In this report, an in-depth data quality assessment was conducted initially, to ensure data accuracy, completeness, and consistency. A detailed exploratory data analysis was undertaken which explored the age distribution of the surveyed players, while investigation was conducted to see the relationship between the biological sex of the players and their environmental perception. Additionally, in the same segment, a gendered-analysis was conducted with the in-game tree cutting behaviour between male and female players. Thirdly, the report identified crucial socio-demographic variables which determined the environmental perception of the players, and then a classification model was developed to forecast a player's environmental perception, based on their socio-demographic variables.

Section 2: Data Quality

2.1 Data Source

The dataset under consideration was taken from a public dataset titled 'A multinational dataset of game players' behaviours in a virtual world and environmental perceptions' from the Science Data Bank (ScienceDB) data repository platform. The dataset, titled as 'data_640_validated.csv' was published on 9th October 2021, and distributed under the Creative Commons 4.0 License. The dataset is in the Comma Separated Value (CSV) format, consisting of responses from the surveyed ACNH players. It also consists of a codebook, titled 'Data description_validated.xlsx', which is a Microsoft Excel Spreadsheet format. The codebook is a detailed guide for the survey questionnaire and has been extensively referred to while analysing the dataset. The dataset can be accessed here: https://www.scidb.cn/en/detail?dataSetId=cb5d36cce29f4e5695a586c9b85d04b6

2.2 Methods of Investigation

The following methods were employed to assess the quality of the data under consideration:

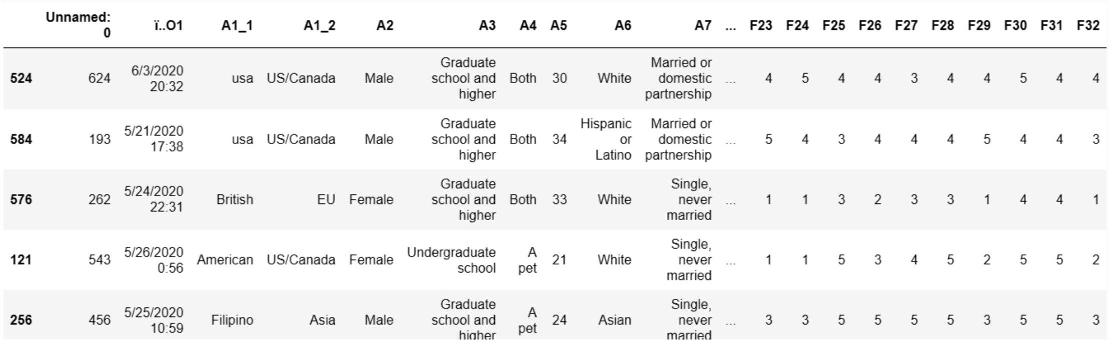
i. Data Exploration: Tools such as Tableau and Microsoft Excel were employed to explore the data visually and check for any inconsistencies.
ii. Data Profiling: Python was utilised to profile the data, such as generation of statistical summaries, identification of duplicate and missing values, conversion of data types, standardisation of inconsistent values, among other data quality checks.

2.3 Data Overview

i. The dataset, initially detected with 'Windows-1254' encoding and a confidence level of ~0.55 suggesting Turkish language in the dataset, faced import issues. 'Latin-1' (ISO-8859-1) encoding led to successful import of the dataset for the data quality checks.
ii. The dataset, formatted as a CSV file, has a file size of 279,051 bytes (~272 kB).

iii. The dataset has a total of 640 rows and 96 columns. An overview of the dataset can be seen in Figure 1 below:

Figure 1 Overview of dataset under consideration



	Unnamed: 0	ï..O1	A1_1	A1_2	A2	A3	A4	A5	A6	A7	...	F23	F24	F25	F26	F27	F28	F29	F30	F31	F32
524	624	6/3/2020 20:32	usa	US/Canada	Male	Graduate school and higher	Both	30	White	Married or domestic partnership	...	4	5	4	4	3	4	4	5	4	4
584	193	5/21/2020 17:38	usa	US/Canada	Male	Graduate school and higher	Both	34	Hispanic or Latino	Married or domestic partnership	...	5	4	3	4	4	4	5	4	4	3
576	262	5/24/2020 22:31	British	EU	Female	Graduate school and higher	Both	33	White	Single, never married	...	1	1	3	2	3	3	1	4	4	1
121	543	5/26/2020 0:56	American	US/Canada	Female	Undergraduate school	A pet	21	White	Single, never married	...	1	1	5	3	4	5	2	5	5	2
256	456	5/25/2020 10:59	Filipino	Asia	Male	Graduate school and higher	A pet	24	Asian	Single, never married	...	3	3	5	5	5	5	3	5	5	3

iv. The codebook of the dataset can be referred to understand the meaning of each column.

2.4 Data Quality Checks

2.4.1 Data Completeness

i. The dataset does not contain any duplicate records, ensuring a high level of uniqueness.

ii. Several columns of the dataset had missing values. These columns are listed in Table 1. The missing values in column A4 was detected, as one of the responses to the corresponding question is 'None', which was identified as zero value in Python. This was resolved by substituting 'None' responses with 'Neither' string value. Since columns 'D1', 'D2', 'D3' and 'D4' were utilised in the subsequent analysis, it is recommended that the missing values of these columns can be left as it is.

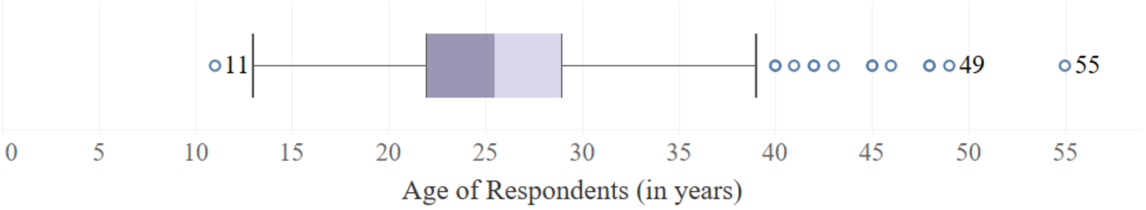
Table 1 Number of missing values in columns

Variable	Question	Number of missing values
A4	Do you have a pet or a garden at home?	86
D1	How long have you been playing video games?	6
D2	Which genre of video games do you often play?	5
D3	How often do you play computer/video games?	1
D4	How many hours on average did you spend playing game a day in the last two weeks?	1
D7	Which style/theme do you prefer when you can terraforming?	13

2.4.2 Data Accuracy

i. The datatype of data corresponding to column 'A5' which denotes the age of the players, was 'object' datatype, and had to resolved. Additionally, there were two entries present as '30s' and 'sub 28' under A5, which, if left unresolved, can cause problems in the analysis.

Figure 2: Boxplot concerning Age of Respondents



To resolve these anomalies in the recorded values, a boxplot was plotted with the age of the respondents, as shown in Figure 2. As there are outliers present in the age of the respondents, mean imputation was not preferred. Therefore, median imputation of the ages was applied to impute the invalid values with the median value (25.5 years). Concurrently, the datatype was transformed into a ‘float’ datatype (numeric) for better representation and ease of analysis of the ages of the respondents.

ii. Additionally, Figure 2 indicated the presence of outliers, where the lower age outlier corresponds to 11 years (the youngest age), while the upper age outlier corresponds to ages above 38 years, with the oldest age being 55 years. It must be noted that ACNH has a rating of PEGI-3 (suitable for players above 3 years), hence it was entirely possible for the ages of the respondents to be as low as 11 years to as high as 55 years, i.e., under the valid spectrum. Therefore, these outliers were removed from the dataset.

iii. A column named as ‘Unnamed:0’ contained discrete integral values from 1 to 640, as per Figure 1. This corresponded to the number of respondents of the survey. Hence it is recommended to treat the column as an ‘Identification Number’ to ensure the cleanliness of the dataset.

iv. The 'ï..O1' column, upon transformation to date-time format, represented values ranging from 15 May 2020, 14:20 to 15 June 2020, 18:51 in DD-MM-YYYY HH:MM format. As there were no sources available to cross-reference, it is advisable to remove this column from the dataset, as it did not affect the subsequent project analysis.

2.4.3 Data Consistency

i. For the column ‘A1_1’ representing the respondents’ nationalities, there were several inconsistencies, which were addressed. This involved consolidating various spellings and terms for the same nationality, such as 'America,' 'American(United States),' 'United States,' 'US,' into the standardized form 'American.' Additionally, misspellings were corrected (such as 'Autsralian' to 'Australian'), dual nationalities were formatted with a dash (such as 'Korean/American' to 'Korean-American'), and inconsistent entries (such as ‘Mixed nationalities’, ‘Mixed’) were grouped into a new category labelled 'Mixed.' Furthermore, leading and trailing spaces in entries were removed to ensure uniformity during analysis.

ii. Column 'D6,' representing respondents' favourite activities in the game, exhibited inconsistent data with special characters like 'Online social interaction, visiting other peopleâ€™s Island' and 'Collecting furniture and D.I.Yâ€™s.' Similarly, column 'D7,' indicating respondents' preferred style/theme for in-game terraforming, had inconsistencies due to open-ended responses. Since these columns were not utilized in subsequent analyses, data cleaning processes were not undertaken, as they fell outside the project's scope. After following the processes, the dataset was used for subsequent analysis.

Section 3: Detailed Analysis

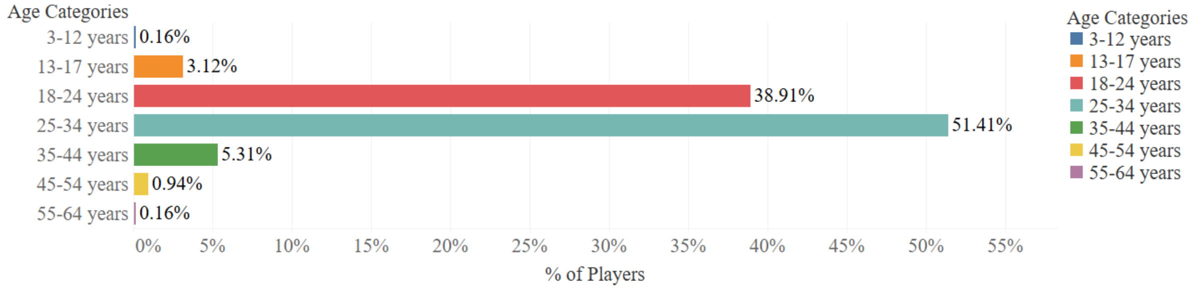
3.1 Exploratory Data Analysis

3.1.1 Age Distribution of Players

The age distribution of players in the dataset has been shown in Figure 3, where young adults in the 18-34 years are the core audience of the game. For a more thorough analysis, the ages have been categorised as ‘3-12 years’, ‘13-17 years’, ‘18-24 years’, ‘25-34 years’, ‘35-44 years’, ‘45-54 years’ and ‘55+ years’. This categorisation has been undertaken given the PEGI-3 rating of the game, where it is reasonable to have young children in the ‘3-12 years’ range. The ‘13-17 years’ accounts for the teenage players, while the ‘18-24 years’ group accounts for the players who are young adults. For older adults, the age intervals have been divided in the range of 10 years.

According to Figure 3, the game was significantly popular among adults aged early twenties to early thirties (51.41% or 361 players), possibly due to disposable income and nostalgia associated with the production company of the game, Nintendo. The young adult group (18-24 years) also constituted a substantial player base (38.91% or 202 players) as they had started gaining autonomy, following their teens. Beyond 34 years, the player proportion declined, yet the game still appealed to older audiences, with the oldest player at 55 years. Only a small fraction (0.16% or 1 player) accounted for the children

Figure 3: Frequency analysis of of respondents by age categories

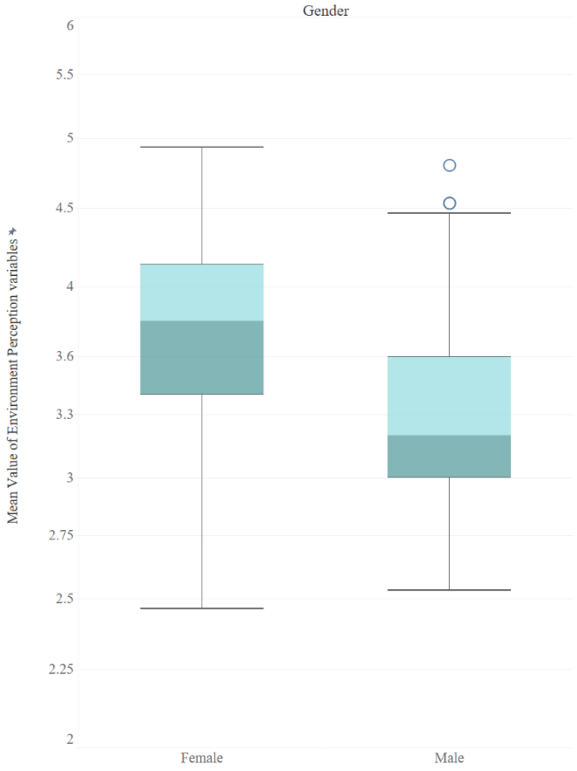


category (3-12 years), which can be attributed to the survey's complexity, and inability of children to comprehend the same. Teenagers represented 3.12% (10 players) within the 13-17 years category.

3.1.2 Relationship between biological sex and players' environmental perception

According to Figure 4, it was observed that female players are more ecocentric, compared to their male counterparts. The boxplot, provided in Figure 4, was calculated taking the mean of the environment perception scores from columns 'C1' to 'C15' at the y-axis, and the gender at x-axis. A few outliers were present in male players with a higher environment perception score; however, these outliers can be considered as individual experiences differ from the experiences of 'male' group.

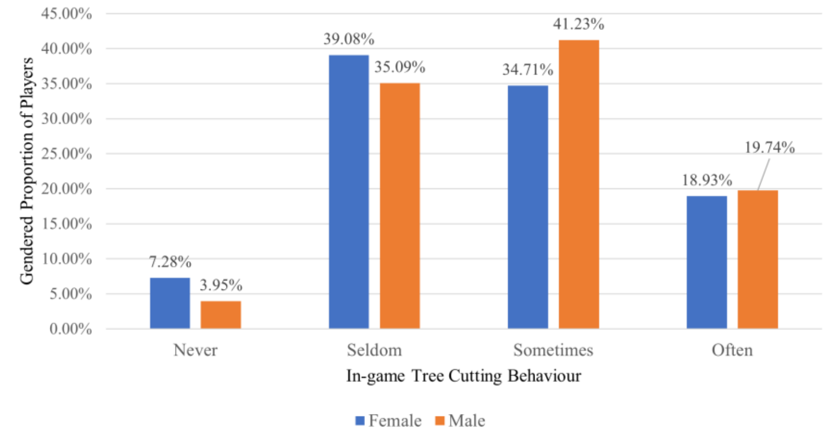
Figure 4: Box-Plot of Environment Perception Variables by Gender



3.1.3 In-Game Behaviour of cutting down a tree: Frequency analysis between males & females

Before undertaking this analysis, it was observed that the number of male and female players in the dataset were not represented proportionally, where there were more females (412 players) compared to males (228 players). A simple frequency analysis between male and female players cannot be a fair representation of their in-game behaviour of cutting down a tree in ACNH, hence a gendered-proportional analysis was undertaken.

Figure 5: Gendered Analysis of In-game Tree-Cutting Behaviour



From Figure 5, it can be inferred that male players were more likely to engage in tree-cutting activity in-game ('Sometimes' & 'Often') than their female counterparts. At the other end of the spectrum, more female players preferred not to cut down trees, compared to male players. Additionally, more female players reported cutting down trees sporadically than male players, which may imply their ecocentric perspective.

3.2 Identification of most important socio-demographic variables to indicate environmental perception of players

A Decision Tree Model was employed to identify the most significant socio-demographic variables that can indicate a player’s environmental perception. A hyperparameter tuning was conducted through GridSearchCV. The primary motivation behind choosing this approach was to recognise the optimal configuration of hyperparameters for the decision tree, which was used for feature selection later.

The analysis proceeded with employing the grid search technique, which involved specifying a grid of hyperparameter values to explore, and it systematically tested each combination to determine which set of hyperparameters produced the best model performance. The hyperparameter values that were considered, in this case, were criterion, maximum depth and splitter. While training a decision tree, it is possible to compute how much each feature's impurities decrease; the greater the decrease in impurities, the more important the feature is. After training a decision tree with the hyperparameters mentioned above, it was found that the optimal set of hyperparameters that provided the best performance of 72.09% were the Gini impurity criterion, a maximum depth of 5, and employing a splitter at random.

Table 2: Corresponding Gini Gain Values for features ‘A1_1’ to ‘A8’

Sl. No.	Variables	Gini Gain Value
1	A1_1	0.1234036289653159
2	A1_2	0.049269045061216994
3	A2	0.27170199292675135
4	A3	0.05316100991253001
5	A4	0.13572046458448067
6	A5	0.24763033281206717
7	A6	0.040557530302294253
8	A7	0.019745823842456107
9	A8	0.05881017159288737

Variables ‘A1_1’ (Nationality), ‘A2’ (Sex), ‘A4’ (presence of pet or garden at home), and ‘A5’ (Age of respondents) were identified as the most important features. This determination was made by considering responses rounded to two decimal places, and the variables exhibiting the highest values can be referred from the Table 2. Therefore, it can be inferred that the signification socio-demographic variables influencing the environmental perception of players encompass ‘A1_1’, representing nationality; ‘A2’, denoting biological sex; ‘A4’, indicating the presence of a pet or garden at home; and ‘A5’, reflecting the respondent's age. These variables, identified through their numerical importance, play a crucial role in shaping the model's outcomes and would offer valuable insights into the players' environmental perceptions.

3.3 Development of Classification Model

To develop a classification model, the initial step involved the selection of an optimal classifier. The Random Forest classifier was chosen for its ability to handle categorical variables, manage high-dimensional features, deliver robust predictions through ensemble learning, assess feature importance, and exhibit versatility across different types of datasets and tasks. The newly created Random Forest classifier underwent evaluation with the use of various metrics.

The initial step was to randomise the data, which is a fundamental practice in machine learning. This randomization helps to promote fairness, improve model generalisation, mitigate overfitting, reduce sensitivity to data order, and ensure statistical validity in evaluating model performance. This randomised model was trained with using the most important socio-demographic variables obtained in Section 3.2. The randomised data was partitioned into two different sections accordingly, where the initial section, consisting of 70% of the dataset, served as the training dataset for the model, while the remaining 30% was allocated for testing purposes. The latter was further split into two equal sections - the first section (15%) was used for testing, and the other section (15%) was used for evaluation.

The entire data was treated as nominal and categorical data types that signified categories with no inherent order or ranking. Hence, during preprocessing, label encoding was used to convert nominal data into numerical inputs since the machine learning algorithm ‘Random Forest’, used here for model creation, operates effectively with numerical inputs. The Random Forest classifier was initialized and evaluated using various metrics, including precision, recall, F1-score, and accuracy. The performance of the model was assessed in two stages - evaluation and test.

3.3.1 Evaluation Classification Report

i. ‘Ecocentric’ vs ‘Anthropocentric’: As per the ecocentric perspective, importance is placed on the ecosystem as a whole; whereas as per the anthropocentric perspective, it is considered that humans are superior to all other organisms. These terms were used as the basis for classification.

ii. For the class ‘Ecocentric’: The precision of 0.64 indicated that 64% were correctly classified. The recall of 0.78 indicated that the model captured 78% of the actual instances of ‘Ecocentric’. The F1-Score of 0.70 provided a balanced measure of precision and recall. Considering the support of 50 instances, these metrics collectively suggested that the model had performed reasonably well in identifying ‘Ecocentric’ instances (with reference to Table 3).

iii. For the class ‘Anthropocentric’: The precision of 0.69 indicated that 69% were correctly classified. The recall of 0.52 indicated that the model captured 52% of the actual instances of ‘Anthropocentric’. The F1-Score of 0.59 provided a balanced measure of precision and recall. With a support of 46 instances, these metrics indicated a reasonable performance, but it is important to note that the recall was comparatively lower, suggesting that the model might have missed some instances of ‘Anthropocentric’ (with reference to Table 3).

With respect to the support values, a high precision might be less impressive if the support is low. For this project, while the precision values were decent for both classes, it is essential to consider the recall and overall F1-Score to understand the model's performance, especially when dealing with imbalanced datasets or classes with varying support.

Table 3: Evaluation Classification Report

	Precision	Recall	F1-score	Support
Ecocentric	0.64	0.78	0.70	50
Anthropocentric	0.69	0.52	0.59	46
Accuracy			0.66	96
Macro Average	0.66	0.65	0.65	96
Weighted Average	0.66	0.66	0.65	96

Table 4: Test Classification Report

	Precision	Recall	F1-score	Support
Ecocentric	0.75	0.89	0.82	47
Anthropocentric	0.88	0.71	0.79	49
Accuracy			0.80	96
Macro Average	0.81	0.80	0.80	96
Weighted Average	0.81	0.80	0.80	96

3.3.2 Test Classification Report

i. For the class ‘Ecocentric’: The precision of 0.75 indicated that 75% were correctly classified, indicating a relatively high precision. The recall of 0.89 indicated that the model captured 89% of the actual instances of ‘Ecocentric’, which suggested an effective identification of these instances. The F1-Score of 0.82 provided a balanced measure of precision and recall, indicating a strong performance in identifying ‘Ecocentric’ instances. Considering the support of 47 instances, these metrics collectively suggested that the model performed reasonably well in identifying ‘Ecocentric’ instances (with reference to Table 4).

ii. For the class ‘Anthropocentric’: The precision of 0.88 indicated that 88% were correctly classified, demonstrating high precision. The recall of 0.71 indicated that the model captured 71% of the actual instances of ‘Anthropocentric’, indicating a relatively good identification of these instances. The F1-Score of 0.79 provided a balanced measure of precision and recall, indicating a strong performance in identifying ‘Anthropocentric’ instances. With a support of 49 instances, these metrics indicated a commendable performance, with high precision and a balanced trade-off between precision and recall. The model performed well in identifying instances for both classes, contributing to an overall accuracy of 80%. The macro and weighted averages further affirmed the model's effectiveness across classes (with reference to Table 4).

When comparing the evaluation and test classification reports, several key observations highlighted the improved performance of the classification model in the latter. Across both ‘Ecocentric’ and ‘Anthropocentric’ classes, there was a notable enhancement in precision, recall, and F1-Score in the test set compared to the evaluation set. The improved recall for the ‘Ecocentric’ class, rising from 0.78 in the evaluation set to 0.89 in the test set, signified more accurate identification of ‘Ecocentric’ instances in the latter data. Concurrently, the ‘Anthropocentric’ class maintained consistently high precision in both sets, with an improvement in recall across the test set. Overall, the model's performance appeared more robust in the test set, suggesting strong generalisation capabilities beyond the training data.

In conclusion, the classification model consistently demonstrated enhanced and reliable performance in predicting a player's environmental perception based on socio-demographic variables. This is evident through the balanced metrics across classes and the improved recall observed in the test set. This indicated the model's potential effectiveness in real-world applications.

Section 4: Conclusion

The overall aim of the project was to develop a classification model that could be used to predict a player’s environmental perception, based on their socio-demographic characteristics. Every step of the project focused on achieving the said goal, which started with the process of data quality check. The accuracy, completeness and consistency of the data was ensured during the process before the project delved into the data exploration stage. Data exploration investigated the age distribution of players in the dataset, along with how the sex of the players influenced their respective environmental perceptions and with the tree-cutting behaviour in game. It was found that young adults between the age group of 25-34 years constituted a major portion of the players, while female players leaned more towards ecocentric views. Building on the processes and findings of the data exploration stage, key socio-demographic variables, including nationality, biological sex, the presence of a pet or a garden, and the age of the respondents, were found to be influential in determining one’s perspective towards the environment. This finding was then utilised with the Random Forest Classification Model, which used the same. The built model demonstrated robust performance, particularly in improved metrics, such as recall in the test set. This emphasised the effectiveness of the model in predicting the players’ environmental perception.