Understanding Player's Environmental Perception in a Game Environment

Introduction

The prediction of human attitudes towards the environment has been a subject of exploration across diverse fields. In a similar context, this coursework aims to utilise response data from players of the Animal Crossing game, a simulation game that allows players to create their in-game environments and engage in activities similar to real life. To clarify the objectives, the initial step involves ensuring data quality for effective data preparation. Subsequently, exploratory data analysis is conducted to identify significant patterns among environmental perception, socio-demographic profiles, and in-game behaviour, with a particular focus on 'cutting down a tree'. Insights obtained from EDA lead to the training of a classification model to predict the level of environmental perception based on the socio-demographic profiles of the players, as well as identifying the most important variables. The dataset includes 640 game players from 29 countries, capturing the responses across six aspects: socio-demographic information, opinions of COVID-19, environmental perception, gaming habits, in-game behaviour, and game-playing feelings.

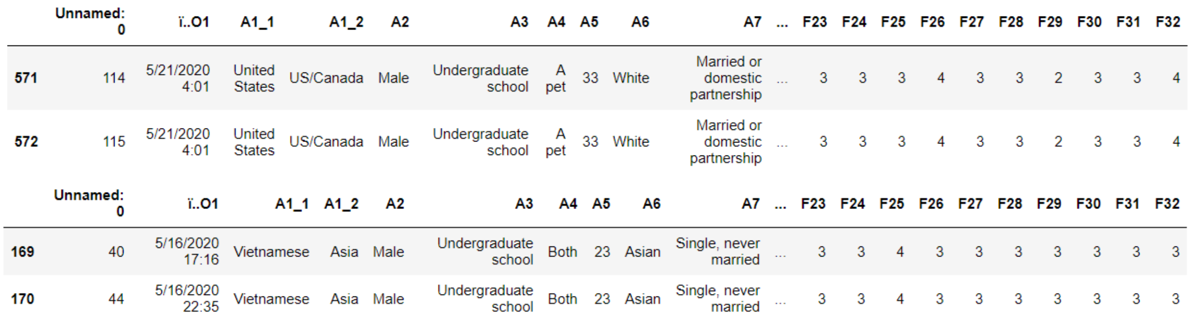
Data Quality

Data quality is a measure that determines the appropriateness of using data based on six aspects: completeness, accuracy, consistency, validity, uniqueness, and timeliness. It is crucial that the data meets these criteria to avoid poor decision-making. Therefore, it is important to address any current issues with the dataset before analysing it. The data quality of the given dataset has been reviewed based on those aspects mentioned above.

First, the completeness of data indicates whether it contains all the necessary information. It is important to check for any issues with missing values, duplicates, and coverage. In the given dataset, there are a total of 112 missing values. This is a relatively small number, considering the dataset consists of 640 rows and 96 columns. It is important to note that missing values in variables 'A4' and 'D4' are represented as zero values. Since these are categorical variables, they can be imputed with 'None', indicating that there is no pet or garden and 0 hours for 'A4' and 'D4', respectively, in this case. For other missing values, various methods can be used to impute them, depending on what the variables represent. These methods may include median, mean, and interpolation. Also, the paper titled 'A Multinational Data Set of Game Players' Behaviors in a Virtual World and Environmental Perceptions' aims to analyse the behaviours of multinational players. However, the variable

'A1_1', which denotes nationality, shows that more than half of the players who participated in the survey are American. This raises concerns about the representativeness of the data as it does not include a diverse range of nationalities. To improve the data, more data from players of different nationalities can be collected. Moreover, in the given dataset, there are two duplicate records. In this case, each row in the dataset represents a survey submission, and it appears that some respondents have submitted the survey twice, resulting in two identical entries with slightly different 'datetime'. Therefore, they can be removed to ensure the completeness of the data.

Secondly, accuracy represents whether data is correct and reliable. In this context, accuracy refers to issues such as misspellings, recording errors, incorrect data formats, and wrong data types. The variable 'A1_1' contains some misspelt values, such as 'Anerican' instead of 'American' and 'Vieynam' instead of 'Vietnam', which can be corrected. Additionally, 'A1_1' includes some values that are not nationalities, such as 'friendly' and '29'. Some values also have spaces at the beginning or end, which can cause the model to interpret them as different values from those without spaces. These errors are often caused by recording errors, and it is essential to find ways to reduce them, such as by using multiple-choice questions instead of allowing respondents to type their own answers. Moreover, the variable 'A5' contains values with the wrong data format, resulting in wrong data types. 'A5' represents the age of players and is supposed to be in integer format. However, due to the presence of values such as 'sub 28' and '30s', which are treated as strings, the data type of 'A5' is currently 'object'. This issue can be resolved by preprocessing the data and converting the data type of 'A5' into the correct data type.



	Unnamed: 0	ï..O1	A1_1	A1_2	A2	A3	A4	A5	A6	A7	...	F23	F24	F25	F26	F27	F28	F29	F30	F31	F32
571	114	5/21/2020 4:01	United States	US/Canada	Male	Undergraduate school	A pet	33	White	Married or domestic partnership	...	3	3	3	4	3	3	2	3	3	4
572	115	5/21/2020 4:01	United States	US/Canada	Male	Undergraduate school	A pet	33	White	Married or domestic partnership	...	3	3	3	4	3	3	2	3	3	4

	Unnamed: 0	ï..O1	A1_1	A1_2	A2	A3	A4	A5	A6	A7	...	F23	F24	F25	F26	F27	F28	F29	F30	F31	F32
169	40	5/16/2020 17:16	Vietnamese	Asia	Male	Undergraduate school	Both	23	Asian	Single, never married	...	3	3	4	3	3	3	3	3	3	3
170	44	5/16/2020 22:35	Vietnamese	Asia	Male	Undergraduate school	Both	23	Asian	Single, never married	...	3	3	4	3	3	3	3	3	3	3

Figure 1 Duplicates

Next, consistency refers to whether the data remains the same across the entire dataset. This includes inconsistent spelling. The variable 'A1_1' contains values that mean the same thing but are written differently. For example, 'american', 'American', 'america', and 'America' all refer to American nationality in this case. However, they are considered to be different values. This issue can be resolved by changing the method of collecting the data. Another issue with consistency is due to a violation of functional dependency. The variable ‘A1_2’ is based on ‘A1_1’, which indicates the regions of nationalities. However, as the nationality field includes mixed nationalities, it is improperly classified as one of the mentioned nationalities, ignoring the other. For instance, when the nationality is indicated as ‘Portuguese-Canadian’, the region is classified as ‘US/Canada’, ignoring the Portuguese information. This can be improved by finding a method to indicate both nationalities when they are mixed.

In addition, validity relates to whether data meets the structure and value requirements as defined in data rules. According to the data explanation, ‘D4’, ‘D5’, and ‘D6’ are in a multiple-choice format, but respondents were also given the option to type their own answers. Since manually typed answers can vary, they can be classified as ‘other’, which can improve the quality of the data.

Uniqueness implies whether data contains any duplicates. The dataset performs well in this regard, as there are no duplicates present. However, in cases where duplicates do exist, it is important to remove them, since they can make it difficult to understand the data when training a model.

Lastly, timeliness addresses whether data remains valid within its defined time frame. The survey was conducted between May 15th and June 9th, 2020, which is a relatively short period. To enhance the timeliness of the data, it would be better to collect it over a more extended period and update it regularly to reflect the current situation.

Detailed Analysis

Exploratory Data Analysis

Exploratory data analysis is an important step before training a model as it helps in understanding the data better. By analysing the age distributions of the players, we can gain some insights. Figure 2 shows that the distribution of ‘Age’ is slightly positively skewed, indicating that the age distribution is skewed towards younger players, with a significant number of players between the ages of 20 and 30. Further analysis in Figure 3 shows the age distributions of male and female players. It is evident that the age distribution of female players is highly positively skewed, whereas the age distribution of male players is negatively skewed.

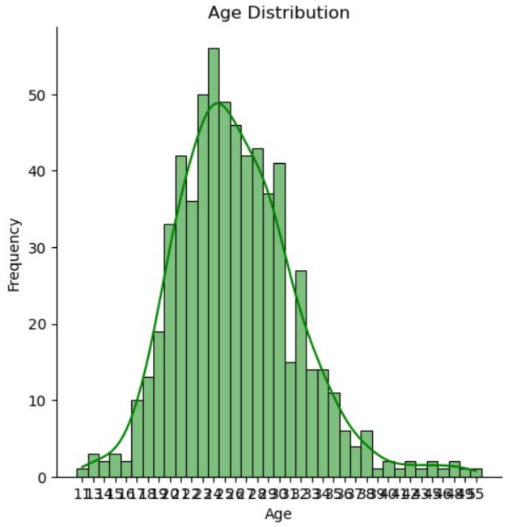


Figure 2 Age Distribution

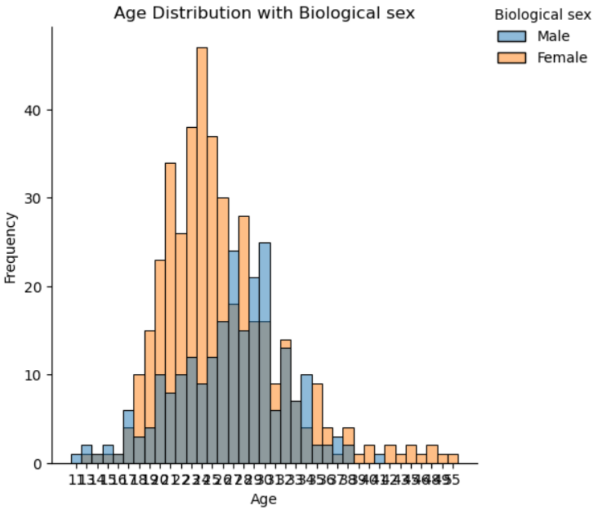


Figure 3 Age Distribution with Biological Sex

To analyse the relationship between biological sex and players' environmental perception, the mean values of variables from C1 to C15 are calculated. These variables consist of questions related to environmental perception, with even-numbered questions referring to human-centric questions and odd-numbered questions referring to environment-centric questions. By analysing the plots, it becomes evident that when answering human-centric questions, females tend to agree more with the environmental perspective compared to males.

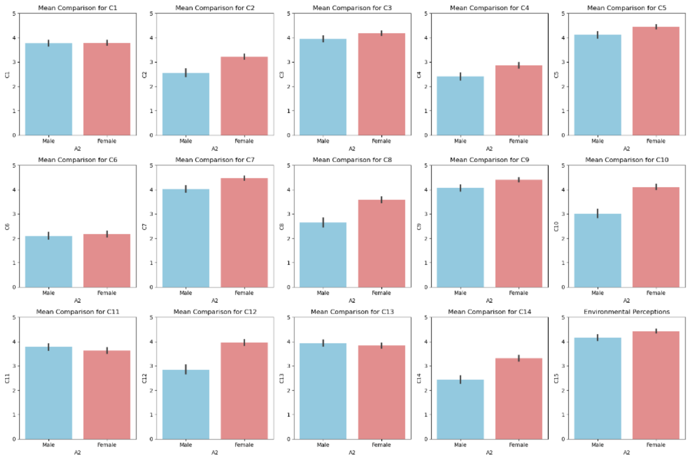


Figure 4 Environmental Perception

Finally, after comparing the in-game behaviour of male and female players in cutting down trees using bar plots, it becomes clear that the distributions for both males and females are similar. The distributions for both are negatively skewed, indicating that both males and females engage in cutting down trees in the game.

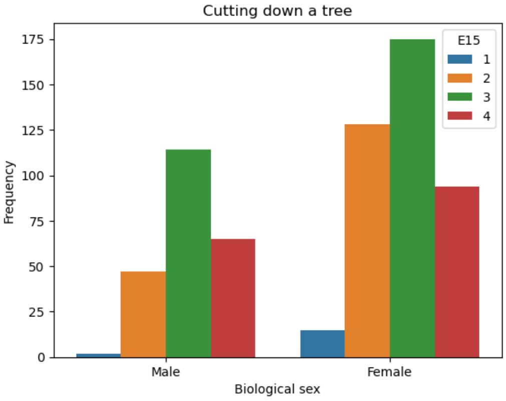


Figure 5 Cutting Down a Tree

Modelling

Using the previously conducted EDA, a classification model can be developed to predict players' environmental perceptions based on socio-demographic variables. In order to develop a classification model that predicts a player's environmental perception, the method of describing the environmental perception needs to be decided. The variables spanning from C1 to C15 can be used as they are related to environmental perception. These variables indicate how much a player agrees with environmental statements on a scale of 1 to 5, with 1 indicating strongly disagreeing and 5 indicating strongly agreeing.

Analysing distributions of each variable, a pattern represents that odd-numbered variables tend to have higher distributions compared to even-numbered variables. Upon fully understanding the variables, it is observed that players tend to answer even-numbered questions relatively low and answer odd-numbered questions relatively high. Therefore, a method is observed to represent players' environmental perception by taking an average for each of the odd-numbered questions and even-numbered questions and adding them together to indicate the environmental perception of each player. To classify players, the mean, median, and distribution of the values are observed. Considering a mean of 7.143359, a median of 7.107143, and a normally distributed distribution, as shown in Figure 7, it is evident that 7.1 can be used as a threshold to classify players. Consequently, players with a value less than or equal to 7.1 are classified as 0, indicating low environmental perception. Those with a value greater than 7.1 are classified as 1, indicating high environmental perception.

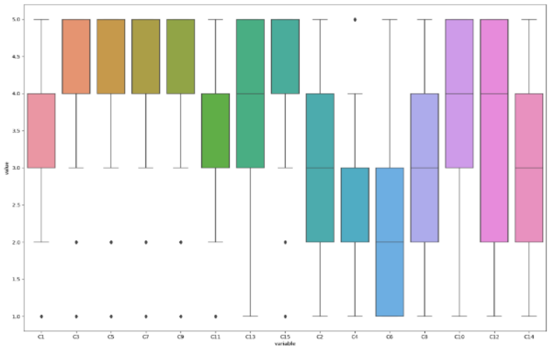


Figure 7 Odd- and Even-numbered Variables

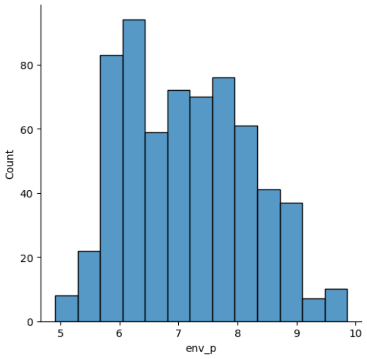


Figure 6 Distribution of Indicators

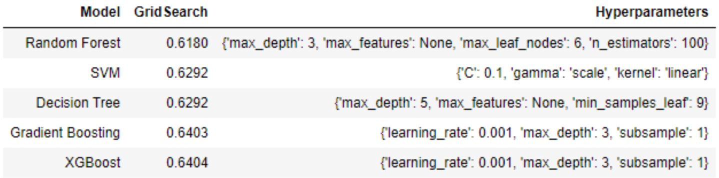
To train a model with the socio-demographic variables, from A1_1 to A8, preprocessing is required. As mentioned previously, there are some issues with these variables. To begin with, the nationality data, which is denoted by 'A1_1', needs to be preprocessed. This data is collected by manually typing the survey answers, which often leads to misspellings and inconsistent spellings. For instance, 'american', 'america' and 'anerica' all indicate 'american'. To address this issue, each value can be classified using a function which reduces the number of categories of 'A1_1' from 110 to 27. This improves the data, making it easier to train a model.

There is another issue with the data type of the 'A5'. This variable indicates the age of the players, but it includes string data. In specific, there are two values with '30s' and 'sub 28', which are considered strings. As a result, the data type of 'A5' is indicated as 'object'. To address this issue, '30s' is replaced with the value 30, which is the mode value of 30s, and 'sub 28' is replaced with the value 28. Then, the data type of 'A5' can be converted to 'int' instead of 'object', which improves the data.

As a next step, duplicates are removed. After preprocessing 'A1_1', exact duplicates within the socio-demographic variables are found. These 49 rows are removed as they can make it difficult to train the model.

Then, the categorical variables are encoded. Since 'A1_2', 'A4', 'A6', 'A7', and 'A8' have approximately five categories for each variable, they are encoded using one-hot encoding. 'A2', which indicates gender, is encoded using label encoding as there are only two categories. For 'A3', which indicates education level, it is encoded using label encoding in the order of Primary school, Secondary school, High school, Undergraduate school, and Graduate school and higher, as it is an ordinal variable. Lastly, 'A1_1' is encoded using label encoding as it contains over 20 categories. The only numerical variable, 'A5', is scaled using StandardScaler.

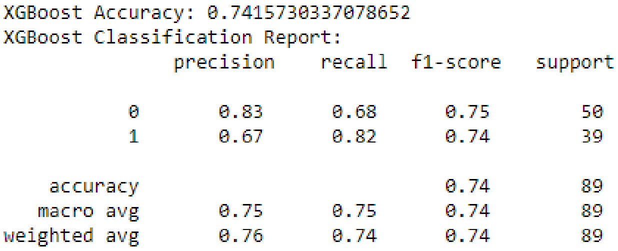
The dataset must first be divided into three separate datasets: training, test, and evaluation. The training dataset should make up 70% of the data, while the test and evaluation datasets should each make up 15%. Once the data is divided, several classification algorithms, including RandomForestClassifier, SVM, DecisionTree, GraidentBoosting, and XGBoost, can be applied to the training dataset to find the best model. To select the best model, GridSearchCV is used to find the optimal hyperparameters using the test dataset. This process helps to identify the best parameters for each algorithm, enabling the selection of the best model with the best parameters, as well as overcoming overfitting. Additionally, this process helps to overcome overfitting by cross-validating with five datasets.



Model	Grid Search	Hyperparameters
Random Forest	0.6180	{'max_depth': 3, 'max_features': None, 'max_leaf_nodes': 6, 'n_estimators': 100}
SVM	0.6292	{'C': 0.1, 'gamma': 'scale', 'kernel': 'linear'}
Decision Tree	0.6292	{'max_depth': 5, 'max_features': None, 'min_samples_leaf': 9}
Gradient Boosting	0.6403	{'learning_rate': 0.001, 'max_depth': 3, 'subsample': 1}
XGBoost	0.6404	{'learning_rate': 0.001, 'max_depth': 3, 'subsample': 1}

Figure 8 GridSearchCV Results

From the table above, one can see that XGBoost has achieved the best GridSearchCV result with the following hyperparameters: learning_rate of 0.001, max_depth of 3, and subsample of 1. To further evaluate the model, it is retrained using a dataset of both training and test datasets.



```
XGBoost Accuracy: 0.7415730337078652
XGBoost Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.68      0.75        50
           1       0.67      0.82      0.74        39

    accuracy                           0.74        89
   macro avg       0.75      0.75      0.74        89
weighted avg       0.76      0.74      0.74        89
```

Figure 9 The Final Result

Based on the evaluation dataset, the XGBoost model achieved an accuracy of 0.7416. In this case, accuracy is used as the metric for the model. Accuracy is a suitable metric when the data is balanced, and in this dataset, there are 318 instances of class 0 and 322 instances of class 1. Therefore, it is appropriate to rely on accuracy as the metric for this model.

The most important socio-demographic variable

While training the model, the most important socio-demographic variables are identified. The feature importance of variables is calculated for each model, and then visualised through bar plots to compare the feature importance scores. As shown in the plots, one has a significantly high score. All five previously trained models have assigned the highest feature importance to 'A2', which represents ' Biological sex'. This means that it is the most important socio-demographic variable indicating the environmental perception of the players, as the model is predicting environmental perception.

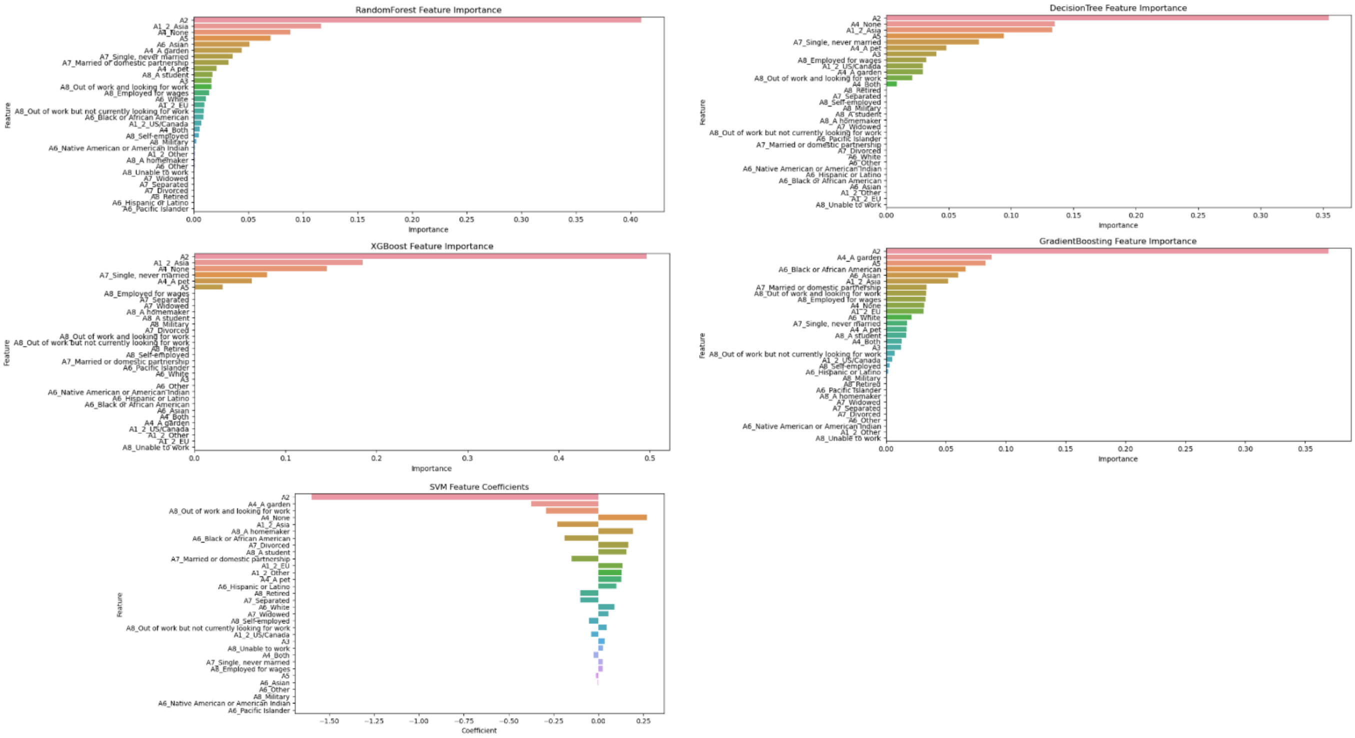


Figure 10 Feature Importance

Conclusions

In conclusion, the coursework aimed to explore and predict players' environmental perceptions. The data is evaluated in terms of data quality, based on six aspects: completeness, accuracy, consistency, validity, uniqueness, and timeliness. After gaining a thorough understanding of the data, some insights are obtained by exploratory data analysis, including the relationship between variables. Finally, a classification model is trained. By applying various preprocessing techniques and training the model with different algorithms and GridSearchCV, the best model is identified. The model demonstrated a 74.16% accuracy in predicting environmental perception based on socio-demographic variables, with the most important feature as 'A2', which represents biological sex.