Understanding Players' Environmental Perception in a Game Environment

I. Introduction

This study aims to analyse and draw insights from 640 players of *Animal Crossing: New Horizons (ACNH)* across 29 countries. The dataset covers socio-demographic profiles, COVID-19 concerns, environmental perceptions, gaming habits, in-game behaviours, and players' feelings while engaging with the game. This study utilises two key datasets: user survey responses and a complementary mapping file detailing the question description.

Prior to the analysis, it is important to ensure the data is consistent and accurate. Therefore, thoroughly investigating the variables of the dataset to identify missing data patterns, duplicate data, format, and type inconsistencies is important. The next step is to set clear objectives for the analysis, specifically targeting the exploration of the correlation between the in-game behaviour and socio-demographic features of ACNH players as well as the socio-demographic features and their environmental perspectives. In the detailed analysis, exploratory data analysis techniques are employed to depict the age distribution among players, investigate the relationship between biological sex and environmental perception, and compare the in-game behaviour (specifically, tree-cutting frequency) between male and female players, the analysis aims to identify key socio-demographic factors that influence players' environmental perceptions within the dataset. Finally, a classification model is developed using socio-demographic variables aimed at predicting a player's environmental perception and then the model is evaluated using appropriate metrics.

II. Data Description

Before data processing, the dataset consisted of 640 rows and 96 columns, this data having been stored in a CSV file with a file size of around 273kb. The dataset, as stated in the introduction, consists of a series of responses to a questionnaire regarding the video game ACNH. As seen in Table 1 below, just under half (47) of the columns are recorded in an ordinal format, the remainder being (mostly) nominal.

Data Classification	Type	Variable Name
Categorical	Nominal	A1.1 - A6 (Socio-Demographic Profiles)
		B1 - B3 (Covid-19 concern)
		D1 – D7 (Game playing habits)
	Ordinal	C1 – C15 (Environmental Perception)
		E1 – E28 (In-game behaviour)
		F1 – F32 (Game-playing feeling)
Numerical	Date/Time	Ô..O1

Table 1: Classification of the dataset's columns

There are 76 int-type columns, of which 75 contain answers to survey questions and are scaled using the Likert Scale from 1 (strongly disagree) to 5 (strongly agree). The only other int-type column is unnamed and consists of consecutive integers in the range 1 to 640, which can safely be presumed to be the index of the dataset. This column has been named **In**. A column that one would expect to be int-type is the Age column but instead is object type. The source of this issue and its rectification has been dealt with in the data quality section. Careful observation reveals that the dataset is **sorted by Age**. For the sake of order, the dataset was sorted by the **In** column (from 1 to 640).

III. Data Quality

Various checks and manipulations were performed to improve the dataset's quality. Firstly, the *isnull()* function was used to identify missing values across the dataset, which revealed that 16 rows from columns 'D1', 'D2', 'D3', and 'D7' had missing data.

To understand the missing data patterns, the ‘missingno’ library in Python was employed. The library creates a matrix and bar plots displaying missing data distributions across the missing columns which helps understand the distribution and prevalence of missing values within the dataset. As seen in Figure 1, it was interpreted that columns D1 and D2 exhibited possible related patterns in missing values, suggesting a Missing Not at Random (NMAR) scenario due to its blocked structure of missing values. Column D3 contained only one missing value, and hence, no specific pattern was identified. Column D7 appeared to have no pattern related to other variables, therefore Missing Completely at Random (MCAR). Since no rows were found with an entire row of missing data, no rows were removed from the dataset due to missing data. Additionally, the values in column D7 are not provided clearly or might contain unexpected or undefined entries like "??" which makes it challenging to interpret the data's quality or relevance. In terms of data quality, this column appears to have inconsistent or missing information.

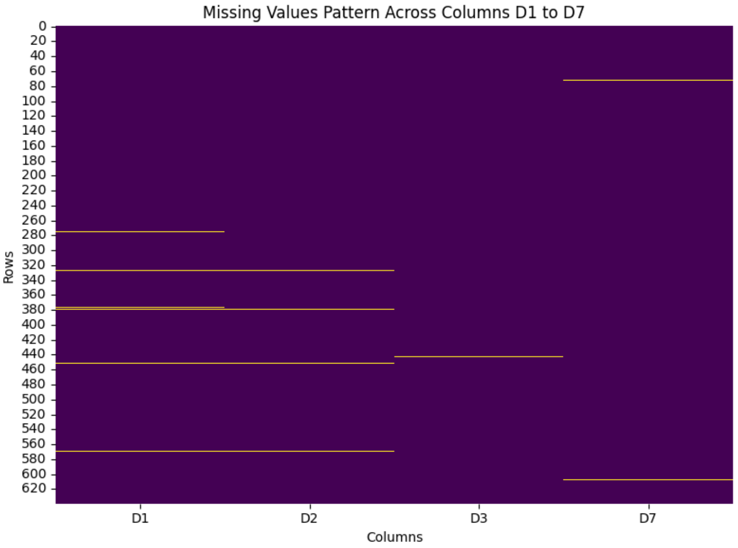


Figure 1: Missing values pattern across columns D1 and D7

Spelling errors, repetitions, and mis-categorizations were common in the dataset, mostly occurring in the D columns (Game playing habits) or the nationality columns as these had user-generated responses. For example, the nationality column has around 110 “unique” values before processing whereas below half of these refer to truly unique values, using basic data correction techniques it can be shown that at least half of these are redundant (being spelling errors, invalid response, or response that could be corrected). In D columns there were similar errors, though it was much harder to check how many could be stated as truly redundant before doing full pre-processing and the sheer number of values.

Similarly, in the dataset's column D6, there's a notable inconsistency in the way game-playing habits are listed. Some entries provide detailed, specific actions like "Decorating my island," while others offer vague or broad descriptions such as "Everything." This lack of uniformity makes it challenging to categorize or compare the activities effectively. Implementing fixed entries, such as a dropdown menu, could have resolved this issue by offering predefined choices, ensuring uniformity, and facilitating easier analysis and comparison of activities.

Special characters such as placeholder values, extreme values, and out-of-line values must be addressed to ensure uniform, high-quality data. Column A5 has two values where age is written as “30s" and “sub 28” whereas a numerical value is expected for the age column. As out of 600+ data entries only two of the rows have these values, it was appropriate to remove these values from the dataset.

The dataset is skewed towards female observations with 410 instances, while there are only 228 male instances. The dataset was found to have no outliers in any of the columns but there is a case of duplication where the entire tuple, even the time of form submission, is identical. However, this could also be a very rare coincidence. For the scope of this project, it was considered as a case of duplication and removed from the dataset.

In this project, the focus is towards A columns which are socio-demographic features including nationality, region, ethnicity, age, gender, highest education level, employment, marital status, and whether the players have a pet or a garden. Further, we focus on classifying the frequency of males and females in comparison to the E columns for in-game behaviours. Finally, we focus on the C columns which is the environmental perception of the players, the C columns are split into two factors: the even-numbered columns classifying anthropocentric behaviour and the odd-numbered columns classifying eco-centred behaviour.

IV. Detailed Analysis

Section A.1: Age distribution of players

The age distribution of players constitutes a vital aspect of understanding the data, as it is the only numeric-type feature other than the Likert scale answers. In Figure 2, it is observed that the age '24' is the modal class. The mean of the distribution is 26.09, while the median is 25.5. The distribution of age is skewed towards the ages 20 to 24. The youngest player on record is 11, while the oldest is 55.

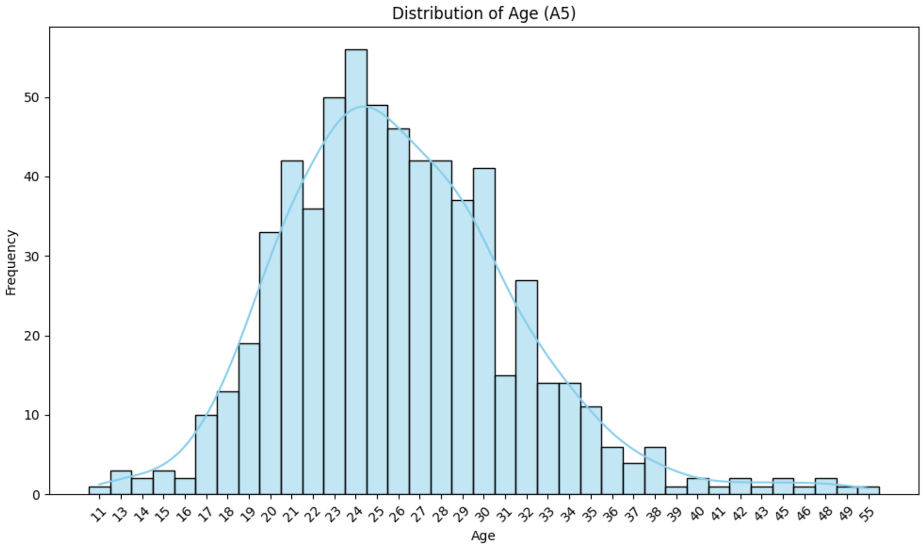


Figure 2: Age distribution

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

There are 228 males and 410 females in the dataset. The target variable here is the mean of the responses to group C of features, which concerns the environmental perception of the players. The higher the score, the more ecocentric the player. The histograms seen in Figure 3 display the distribution of the mean of environmental perception scores for each biological sex. It reveals that females tend to have a higher score than males. The mean for female environmental score is 3.76, while the mean for male environmental score is 3.32. The modal class for females is 3.75-4, while the same for males is 3-3.25. This clearly indicates that females generally tend to be more eco-centric compared to males.

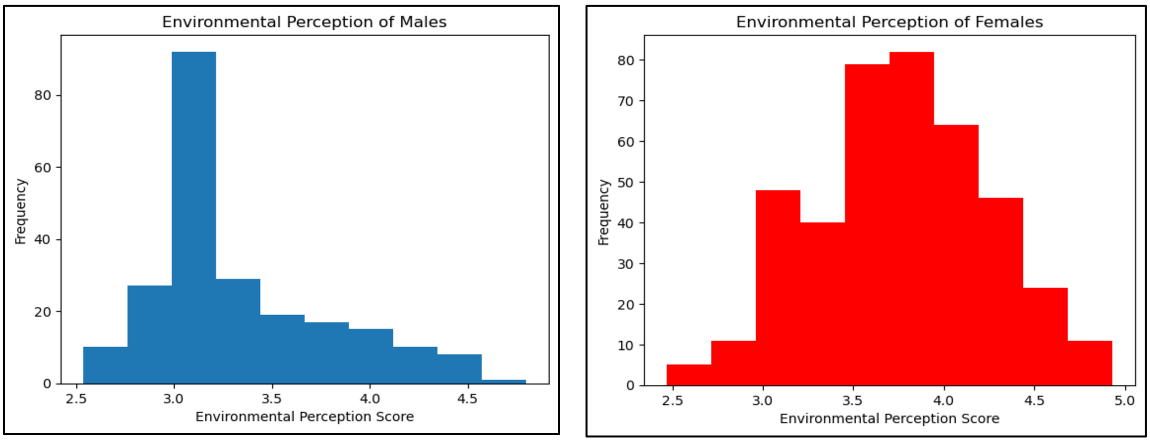


Figure 3: Environmental perception of Male and Female

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

When comparing the frequency of male and female players' likelihood of cutting down trees, from Figure 4 it is observed that neither males nor females strongly agreed (Likert scale 5) with cutting down trees often. The highest frequency of Likert score for males was 3 (94 individuals), indicating a moderate likelihood. For females, the highest frequency of Likert scores was 2 (159 individuals) suggesting a tendency to show less inclination towards cutting down trees compared to men.

Therefore, the overall conclusion based on this data is that a larger number of females showed a tendency to disagree or exhibit a lower inclination towards cutting down trees in comparison to males.

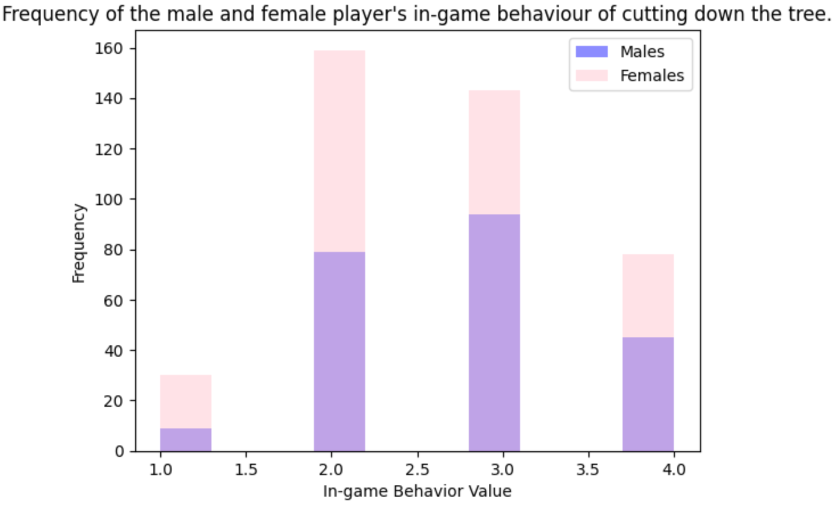


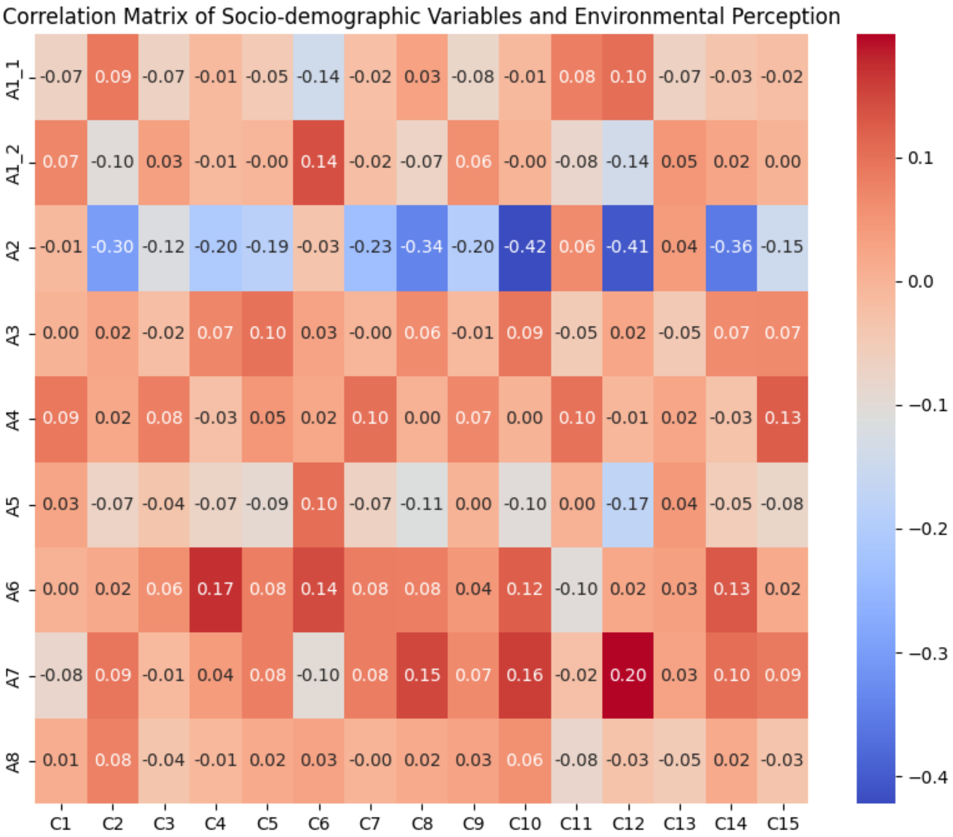
Figure 4: Frequency of Male and Female (Cutting down the trees)

Section B: Identify the most important socio-demographic variables to indicate the environmental perception of the players.

To distinguish the relationship between socio-demographic variables and the environmental perception of the players, the first step conducted is the calculation of the correlation matrix between these features.

Using the corr() function in Python a correlation matrix is created which is a statistical technique used to evaluate the linear relationship between these variables. Every cell consists of a correlation coefficient (strength of the relation) where 1 is considered a strong positive relationship, 0 is a neutral relationship and -1 indicates a strong negative relationship.

Examining this correlation matrix (Figure 5) gives insights into which socio-demographic variables might be more closely linked to aspects of environmental perception. For instance, the correlation between A2 and various environmental perceptions demonstrates notable negative associations across multiple dimensions: C2, C3, C4, C5, C7, C8, C10, C12, C14, and C15, with coefficients ranging from -0.185 to -0.422. This suggests an inverse relationship between variable A2 and how individuals perceive their environment concerning these aspects. Variables like A4 exhibit weak positive correlations with C1, C3, C7, C12, C14, and C15, ranging from 0.020 to 0.125, indicating a slight positive relationship. However, further investigation is required to understand which socio-demographic variables are **most** important to indicate the environmental perception of players.



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
A1_1	-0.07	0.09	-0.07	-0.01	-0.05	-0.14	-0.02	0.03	-0.08	-0.01	0.08	0.10	-0.07	-0.03	-0.02
A1_2	0.07	-0.10	0.03	-0.01	-0.00	0.14	-0.02	-0.07	0.06	-0.00	-0.08	-0.14	0.05	0.02	0.00
A2	-0.01	-0.30	-0.12	-0.20	-0.19	-0.03	-0.23	-0.34	-0.20	-0.42	0.06	-0.41	0.04	-0.36	-0.15
A3	0.00	0.02	-0.02	0.07	0.10	0.03	-0.00	0.06	-0.01	0.09	-0.05	0.02	-0.05	0.07	0.07
A4	0.09	0.02	0.08	-0.03	0.05	0.02	0.10	0.00	0.07	0.00	0.10	-0.01	0.02	-0.03	0.13
A5	0.03	-0.07	-0.04	-0.07	-0.09	0.10	-0.07	-0.11	0.00	-0.10	0.00	-0.17	0.04	-0.05	-0.08
A6	0.00	0.02	0.06	0.17	0.08	0.14	0.08	0.08	0.04	0.12	-0.10	0.02	0.03	0.13	0.02
A7	-0.08	0.09	-0.01	0.04	0.08	-0.10	0.08	0.15	0.07	0.16	-0.02	0.20	0.03	0.10	0.09
A8	0.01	0.08	-0.04	-0.01	0.02	0.03	-0.00	0.02	0.03	0.06	-0.08	-0.03	-0.05	0.02	-0.03

Figure 5: Correlation Matrix of Socio-demographic variables and Environmental Perception

For this, chi-squared analysis is performed. This method enables the selection of key socio-demographic attributes that might significantly influence how individuals perceive the environment using the best chi-squared scores. The reason behind choosing this method is because it is simple and can be easily applied for feature selection in machine learning, it is also computationally efficient and can be applied to large datasets with many features such as this dataset.

To extract the most relevant features, the A columns are compared to the C columns, and a list of each C column and their p-value is made after running. The test was deemed statistically relevant according to a 5% error level. The only A value dropped however was A3. Ultimately only column A3 (education level) was below this threshold, therefore all remaining A values were deemed important enough to keep.

Section C: A classification model that can be used to predict a player's environmental perception based on socio-demographic variables only.

The A group of columns includes socio-demographic profiles such as nationality, biological sex, education, employment, etc. Feature A1_1, which details nationality was dropped as the data was inconsistent, with multiple string values for the same nation. Additionally, the data in A1_1 was already in part encapsulated by feature A2_2, which contains the regions of the world to which the players belong. Feature A5, which contains the age of the players, was updated as mentioned in the Data Quality section.

The C group of columns contains information on the environmental perception of the players. To simplify the classification task, the mean of the 15 features was computed. Accordingly, the records were binarily classified into ecocentric (higher mean scores) and anthropocentric (lower mean scores).

The model selection phase involved understanding the data in terms of data types, the presence of multiclass variables, as well as the size of the dataset. A tree-based model, specifically a random forest, seemed to be our best bet based on the following factors:

1. Handling mixed data types: Tree-based models can effectively handle both numeric and categorical data, which is vital for our dataset.
2. Ensemble methods: As a step up from decision trees, a random forest involves building multiple decision trees with random subsets of the training data and voting on the best model.
3. Non-parametric: Considering the size of the dataset, and the mix of numeric and multi-class categorical variables, a model that doesn't assume a specific functional form can be very useful.

Nonetheless, while the factors above were very convincing, two other models were considered for testing, namely Logistic Regression and Naïve Bayes. For each model, relevant parameters were picked or calculated using training data. For testing, the only metric used for comparison was accuracy, as it is easy to calculate and can compensate for the high number of parameters. The accuracy scores came out 75% for random forest, 68% for logistic regression, and 65% for Naïve Bayes. Based on our initial qualitative understanding and the quantitative results from training and testing the models, the random forest was finally chosen as the closest to the ideal model for the data at hand.

The best parameters of the random forest (on the test set) are listed below:

1. Purity measure: Gini
2. Decision Trees: 100
3. Minimum Sample Size: 4

Some other parameter sets used that are worth mentioning include Set 1 (Entropy, 100, 6) and Set 2 (Gini, 300, 2). Set 1 worked second best on the testing set, while Set 2 fared the worst. The set chosen had the highest accuracy. It is worth noting that other parameter sets had the same accuracy as Sets 1 and 2. The performance of the model with the chosen set of parameters also had the best evaluation metrics, showing that overfitting was unlikely. Set 2 had higher evaluation accuracy than Set 1, showing that with lower testing accuracy, it did a better job at generalising the results.

A confusion matrix is used for the evaluation of the model's performance for a comprehensive performance assessment, True Positives, True Negatives, False Positives and False Negatives provide a detailed breakdown of the model's predictions. Calculating accuracy metrics from the confusion matrix such as precision and recall helps with making sure the model is not just giving out the most common output and ensures the model is improved by addressing class imbalances and refining feature selection.

To evaluate key performance metrics that provide insights into the behaviour and accuracy of the random forest classification model in the context where a positive outcome indicates that a player is ecocentric, Figure 6 reflects the performance of the model. Out of the instances evaluated, the model correctly identified 31 instances where players are ecocentric (True Positives). It also recognized 39 instances where players are more (True Negatives). However, there were 15 instances where the model incorrectly predicted whether the player was ecocentric or anthropocentric (False Positives). Additionally, the model failed to identify 11 instances where players were ecocentric (False Negatives).

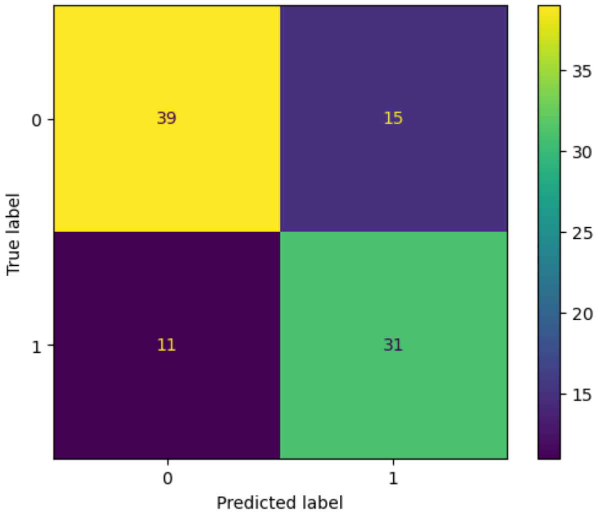


Figure 6: Confusion Matrix for Random Forest Model

Due to the inherent complexity of this task, creating a robust model which can confidently predict someone’s environmental perception requires consistent efforts and a lot of new research.

The accuracy of the model stands at 72.9% indicating overall correctness in classifying both positive and negative instances. The recall reflects the model’s capability to identify actual positive cases at around 72.9%. This suggests the model captures most instances representing ecocentric players. The precision of approximately 73.4% showcases the accuracy of the model’s positive predictions emphasising its reliability. The F1 score reinforces the model’s balanced performance at 72.9% showcasing a consistent model.

V. Conclusion

Using the multinational dataset of Player’s in-game behaviour in a game that involves virtual world and environmental perceptions, a classification model to predict players’ environmental perceptions based on their socio-demographic features was created. The data pre-processing was brief, involving removal of a few errors in A5 by dropping the rows with errors. Age and Gender were shown to be relevant metrics, with differences in distributions between genders and skew in the age suggesting these are relevant features. Chi-squared analysis was performed on the columns in A and C to find correlations, with a value of 5% used as the threshold, ultimately all columns in A had at least 3 strong correlations in C but column A3. It was decided Environmental perception was to be a binary class, which was determined by whether the sum of someone's response to C was greater than or lower than the mean. Our final model proved to be quite good at predicting these classes, with an accuracy of 73% and precision and recall scores of 73%. Through the use of test and evaluation sets during training, as well as checking a range of potential parameters, we are confident our results are not heavily overfitted and can be applied to similar situations in the future.