Understanding The Impact of Social Isolation and Loneliness in a Game Environment

Abstract

This report explores the relationship between gaming behaviour in "Animal Crossing: New Horizons" (ACNH) and isolation duration. By analysing 640 data entities across six dimensions, potential correlations could be identified. Using Exploratory Data Analysis (EDA) and Pearson correlation coefficients, key gaming behaviour correlated with isolation duration were pinpointed. Data preprocessing involved label merging, encoding, normalization, and dimensionality reduction. The data were then modelled using Support Vector Machine (SVM) and K-Nearest Neighbours (KNN) algorithms, achieving accuracies above 60%. This study offers insights into how isolation duration affects gaming behaviour.

1. Introduction

The COVID-19 pandemic had led to prolonged isolation, restricting social interactions and increasing anxiety and depression. Against this backdrop, online games, like "Animal Crossing: New Horizons (ACNH)," became a popular way to relieve loneliness during that period. Previous studies have shown that online games can meet human social needs and relieve their feelings of loneliness (Ballard and Spencer, 2023). Especially during the pandemic, the number of gamers and game revenues increased significantly (Ellis et al., 2020). The purpose of this report is to analyse ACNH players' gaming behaviour during isolation and explore the relationships between isolation duration and in-game behaviour. We employed quantitative methods (including experiments and observations) to develop and optimize machine learning models to predict a player's length of being self-isolated/social distancing. The dataset used in this project includes information such as players' isolation duration, gaming activity levels, and personal feelings (Vuong et al., 2021), providing a basis for this analysis. This report has been divided into four parts: Exploratory Data Analysis, Analysis of Key Gaming Behaviors, Prediction Model Development, and Conclusions.

2. Exploratory Data Analysis

We first performed data preprocessing and checked for missing values in the dataset. The results showed that there were columns with missing values, but they irrelevant to the columns related to analysis objective. Therefore, no further processing of the data was needed.

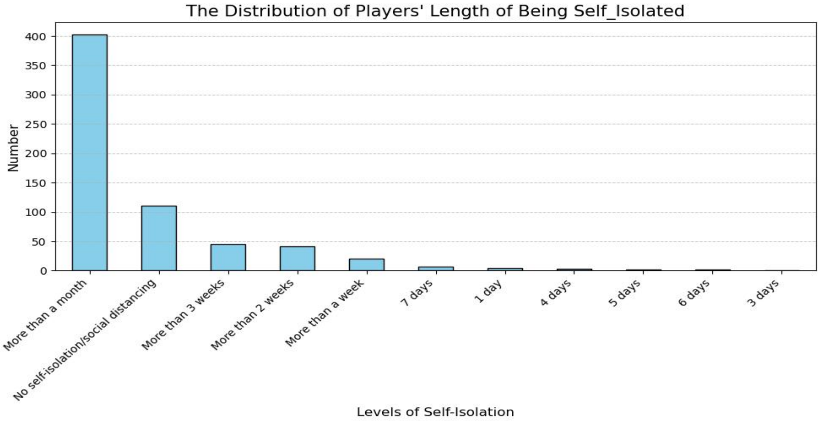


Figure 1. The distribution of players' length of being self-isolation/social distancing.

We used a bar chart to display the distribution of players' isolation durations. The horizontal axis represents isolation duration, and the vertical axis shows the number of individuals. As shown in Figure 1, the categories include over one month, non-isolated, over three weeks, over two weeks, over one week, seven days, six days, five days, four days, three days, and one day. The largest group, with nearly 400 individuals, has an isolation duration of over one month, significantly more than the second-largest group (non-isolated) with around 100 individuals. The other categories have fewer than 50 individuals each, indicating that those isolated for over one month constitute the majority.

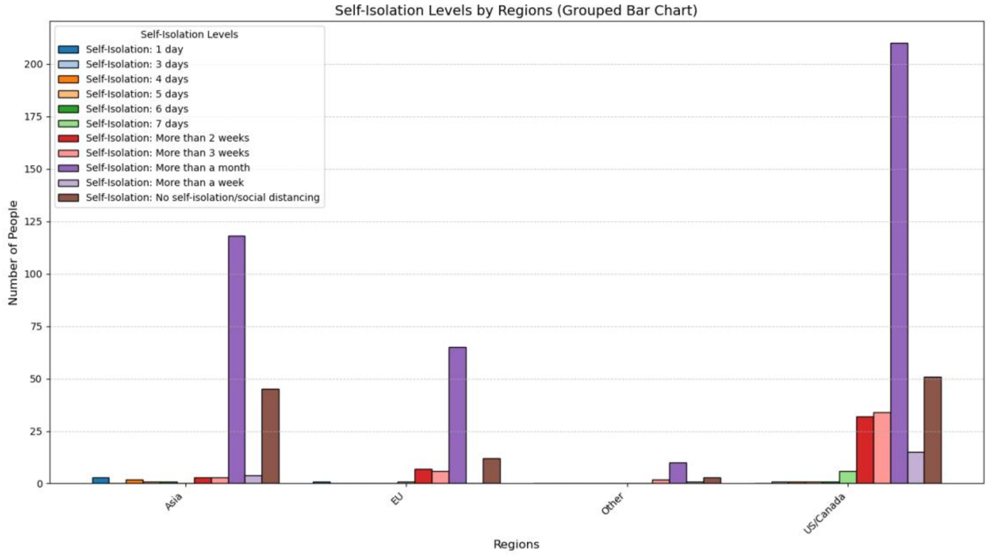


Figure 2. The regional distribution of players' length of being self-isolation/social distancing.

In Figure 2, the data is divided into four groups by region: Asia, Europe, US/Canada, and Others. The bar chart is constructed with regions on the horizontal axis and the number of people on the vertical axis, using different colors to represent different levels of isolation. According to Figure 2, in all regions, the number of individuals isolated for more than one month ranks first, far surpassing the second-highest category, which is the non-isolated group. This trend is particularly prominent in Europe and US/Canada, where the former has approximately four times the number of the latter.Therefore, it can be concluded that in different regions, individuals isolated for more than one month form the largest group and are more likely to experience loneliness and self-isolation.

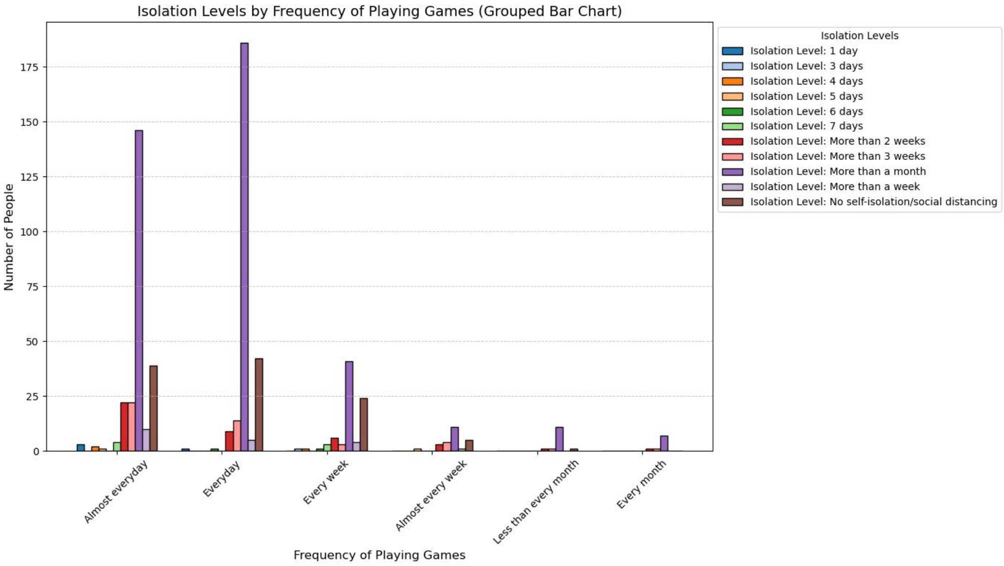


Figure 3. The isolation levels by frequency of playing games (Grouped Bar Chart).

From Figure 3, it is clear that in all game frequency groups, the number of players isolated for more than one month is the highest. Additionally, players with isolation periods of more than 2 weeks and more than 3 weeks account for relatively larger proportions in almost everyday and everyday groups compared with other groups. Thus, we can conclude that isolation duration is correlated with game frequency. Players in long-term isolation consistently dominate across different game frequency ranges, and their gaming frequency tends to be higher. This suggests that during extended isolation, players may increase their gaming frequency to alleviate feelings of loneliness.

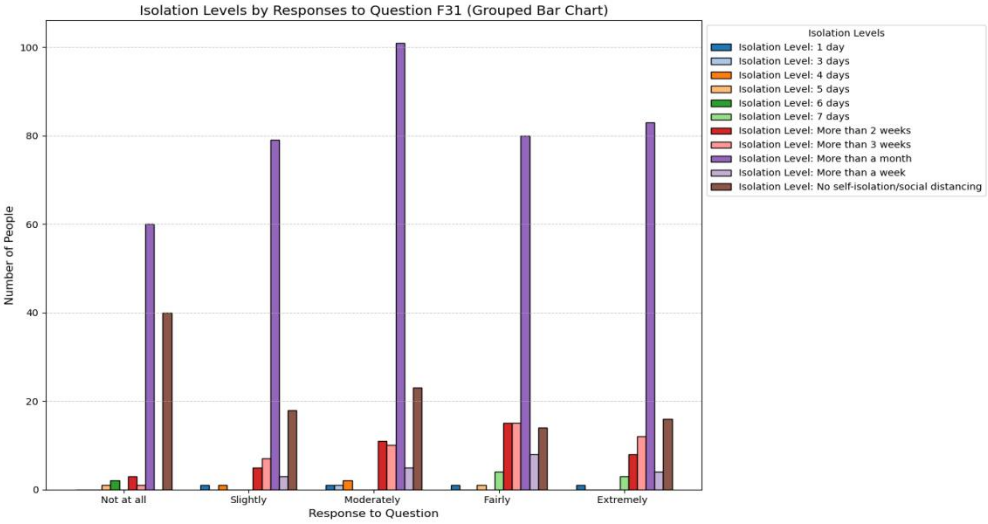


Figure 4. The comparison between different isolation levels and the response to question F31 (Game-playing feeling - I lost connection with the outside world).

From Figure 4, the distribution of responses to question F31 shows clear patterns. Most respondents across options have been in isolation for one month, with non-isolated individuals typically ranking second, except for the "Fairly" option, where those isolated for over two and three weeks exceed the non-isolated group. Additionally, among those choosing "Not at all," about 60% were in isolation, increasing to 80% for "Slightly," 85% for "Moderately," and 90% for "Fairly" and "Extremely." This highlights a significant correlation between longer isolation and a heightened perception of disconnection from the outside world.

3. Analysis of Key In-Game Behaviors Indicative of Players' Duration of Self-Isolation or Social Distancing

In the research on identifying the most important in-game behaviour that indicate the player's isolation duration, we have several important considerations. Firstly, it is of crucial importance to ensure the objectivity and credibility of the data, because only in this way can we guarantee that the research conclusions are authentic and reliable and avoid deviations caused by data issues. Secondly, we are well aware that the data contains abundant information. By comprehensively exploring the correlations between all in-game behaviour and players' isolation duration, we can avoid missing any key factors that might have a significant impact on the results. Finally, the encoding of non-numerical data and the results of correlation analysis will facilitate subsequent in-depth work such as the construction of prediction models. The detailed usage methods are as follows.

1. Data Extraction and Integration:

```
df_identify1 = df.iloc[:,36:64]
df_identify2 = df.iloc[:,12]
df_identify = pd.concat([df_identify2,df_identify1],axis=1)
```

The objective is to screen out all kinds of in-game behavior feature data that may be related to isolation duration, and integrate them into a new data frame named "df_identify" to prepare for the subsequent analysis.

2. Correlation Analysis:

```
correlation_matrix = df_identify.corr()
correlation_with_B2 = correlation_matrix['B2'].drop('B2')
top_correlated_columns = correlation_with_B2.abs().sort_values(ascending=False).head(5)
Top_correlated_columns
```

Calculating the correlation coefficient matrix of the integrated data frame "df_identify", and then extracting the correlation situations between other columns and the "'B2'" column. Sorting them to find out the top 5 columns with the strongest correlations, so as to determine which in-game behavior features are most closely associated with the length of players' self-isolation/social distancing.

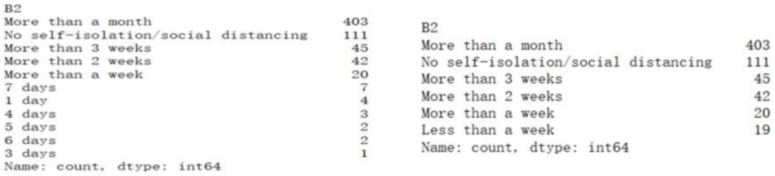
4. Developing a Machine Learning Model to Predict Players' Self-Isolation or Social Distancing Duration Based Solely on In-Game Behavior Variables

We aim to explore the relationship between "length of being self-isolated/social distancing" and "in-game behavior." Therefore, we use "length of being self-isolated/social distancing" as the label and "in-game behavior" as the features for subsequent research. The purpose of this project is to predict a player's "length of being self-isolated/social distancing" based on their "in-game behavior." Analyzing the label reveals that the dataset contains 11 different categories. Hence, this project involves a labelled classification problem, which falls under supervised learning. Based on this, we adopt KNN and SVM algorithms for modelling.

4.1 Data Processing

To facilitate subsequent modelling, we preprocessed the data in four steps: label merging, data encoding conversion, data normalization, and dimensionality reduction using LDA.

4.1.1 Merge Labels



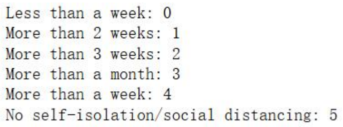
```
B2
More than a month                        403
No self-isolation/social distancing      111
More than 3 weeks                         45
More than 2 weeks                         42
More than a week                          20
7 days                                     7
1 day                                      4
4 days                                     3
5 days                                     2
6 days                                     2
3 days                                     1
Name: count, dtype: int64
```

```
B2
More than a month                        403
No self-isolation/social distancing      111
More than 3 weeks                         45
More than 2 weeks                         42
More than a week                          20
Less than a week                          19
Name: count, dtype: int64
```

Figure 5. The merge of some labels.

As observed in the label distribution (Figure 5), there is a severe class imbalance in the data across different isolation time categories, which could lead to poor model fitting. Therefore, we further processed the data by merging the categories with fewer samples into a single category while maintaining logical consistency. For example, we noticed that the number of samples with isolation times of 7 days or less is very small, so we decided to merge these different isolation times into a single category. As shown in Figure 5, we combined '1 day', '3 days', '4 days', '5 days', '6 days', and '7 days' into a new label—' Less than a week'.

4.1.2 Data Encoding



```
Less than a week: 0
More than 2 weeks: 1
More than 3 weeks: 2
More than a month: 3
More than a week: 4
No self-isolation/social distancing: 5
```

Figure 6. The encoding of labels.

```
label_encoder = LabelEncoder()
df_identify['B2'] = label_encoder.fit_transform(df_identify['B2'])
```

From Figure 6, we can see that the LabelEncoder was used to encode the 'B2' column, which represents the duration of players' self-isolation/social distancing or related classification results. This process converts non-numerical data into numerical form, facilitating subsequent numerical calculations, correlation analysis, and other operations, ensuring that the data can be processed on a unified numerical scale.

4.1.3 Data Normalization

Since both the KNN and SVM algorithms are highly sensitive to feature scale, we normalized the data to help the model better extract features and improve accuracy (Kumar et al., 2022).

4.1.4 Dimensionality Reduction（LDA)

The experimental dataset in this project contains 28 features, which results in high dimensionality that hinders the model's ability to extract features and also consumes significant computation time. Since the model aims to solve classification and supervised learning problems, we chose the LDA algorithm instead of PCA to reduce the dimensionality of the experimental data, making it easier for the model to converge in the subsequent steps (Benouareth, 2021).

4.2 Split Data Set

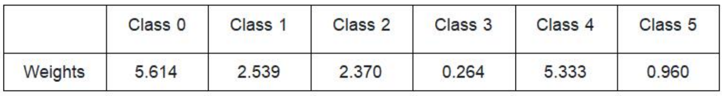
To mitigate the impact of sample imbalance, we use cross-validation during the training process. The entire dataset consists of 640 valid samples, divided into 512 for the train set and 128 for the test set. During training, we employ 5-fold cross-validation (Xiong et al., 2020), selecting 410 samples for training and 102 samples for validation in each fold, with the 128 test set samples used for testing. Afterwards, the model is evaluated using the data from the test set.

4.3 Model development

4.3.1 Parameter Settings

1) SVM Model: Based on the kernel='poly' and the analysis of the data features, it is evident that this model does not have linear characteristics, so 'linear' is excluded (Lu et al., 2023). After dimensionality reduction, the features become relatively simple, which is why 'rbf' is not used. Additionally, during model training, the class_weights parameter (Table 1) is set to assign different weights to the samples, which can further reduce the impact of data distribution imbalance.

Table 1. The weights are based on compute_class_weight function.



	Class 0	Class 1	Class 2	Class 3	Class 4	Class 5
Weights	5.614	2.539	2.370	0.264	5.333	0.960

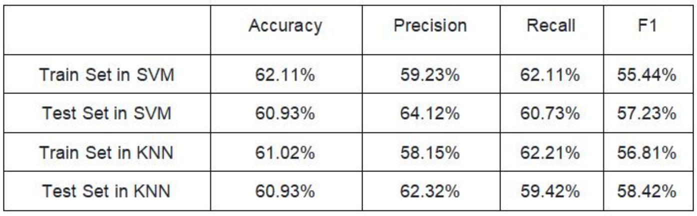
Finally, Grid Search was used during model training to explore and find the optimal parameters (Fuadah et al., 2022).

2) **KNN Model:** By setting n_splits = 6, the isolation time labels have been merged into 6 categories (Paramasivam et al., 2023).

4.3.2 Model Evaluation Metrics

Due to the lack of universality of the single Accuracy evaluation metric, we use four metrics, including Accuracy, Precision, Recall, and F1 score to comprehensively evaluate the model's performance.

Table 2. The evaluation metrics for two models.



	Accuracy	Precision	Recall	F1
Train Set in SVM	62.11%	59.23%	62.11%	55.44%
Test Set in SVM	60.93%	64.12%	60.73%	57.23%
Train Set in KNN	61.02%	58.15%	62.21%	56.81%
Test Set in KNN	60.93%	62.32%	59.42%	58.42%

From Table 2, it can be observed that in the evaluation of the SVM model, the training set demonstrates relatively high recall and accuracy, indicating that the model can effectively identify positive samples in the training set. However, the precision and F1 scores are relatively low, suggesting the presence of some false positives. On the other hand, the Test set shows higher precision compared to the training set, indicating that the model's predictions on the test set are more cautious, with fewer false positives.

For the KNN model, the training set has lower precision but higher recall than SVM, implying a higher false positive rate but a stronger ability to identify positive samples. On the test set, precision and recall are more balanced, and F1 score is slightly higher than the training set. Overall, the SVM model's strength lies in its better performance on the training set, with higher recall and accuracy, making it suitable for scenarios requiring stronger classification capabilities. Meanwhile, the KNN model's advantage is its slightly higher F1 score on the test set, reflecting more stable performance in identifying positive samples and better generalization ability on test data.

4.3.3 ROC Curve Analysis

Owing to the uneven distribution of sample categories, the Accuracy evaluation is no longer representative. Therefore, we introduce the Receiver Operating Characteristic (ROC) curve to observe the performance of models in various categories in more detail (Gneiting and Walz, 2022).

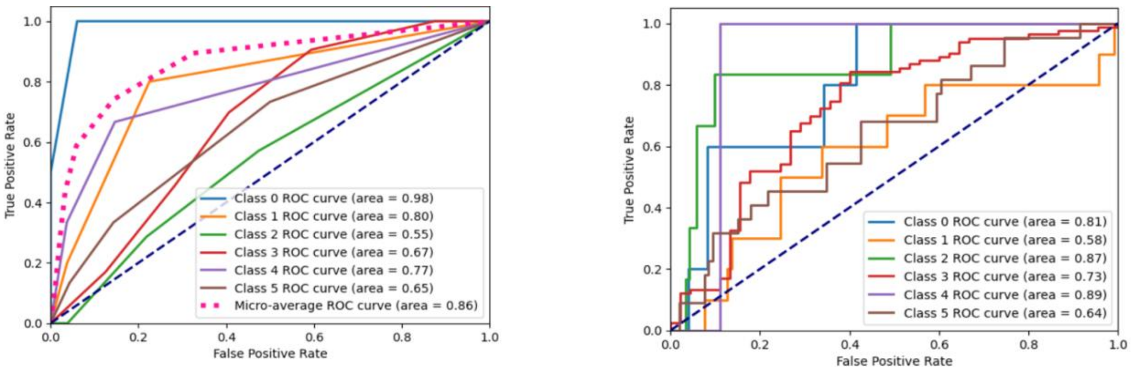


Figure 7. The analysis of the ROC curve on two models (KNN-left, SVM-right).

From Figure 7, it can be seen that the model performs exceptionally well in certain categories (e.g., Class 0) but delivers average performance in others (e.g., Class 2). Overall, the model's performance is acceptable but still has room for improvement, especially in categories with lower precision and recall.

5. Conclusions

This report focuses on the relationship between ACNH players' gaming behaviour and the duration of their isolation. The study primarily employs SVM and KNN algorithms to predict the possible isolation duration based on players' preferences for various gaming behaviour in ACNH. In general, the prediction results are not very satisfactory, mainly due to the following three challenges: Limited Data Volume with High Diversity and severely Imbalanced Samples, Low Feature Correlation with Labels, and based on the EDA. To improve prediction accuracy, we made several efforts, including merging categories to reduce the number of underrepresented samples, testing multiple machine learning algorithms for prediction, and evaluating the models comprehensively from various dimensions. These efforts aimed to mitigate the class imbalance bias caused by the uneven data distribution. As a result, while the accuracy of both algorithms remained around 60%, the F1 scores exceeded 55%. Moving forward, we plan to increase the total volume and balance of the data (if possible), further fine-tune the parameters, or explore other machine-learning methods. Additionally, despite using the same dataset, the results of the two algorithms differ, suggesting that each algorithm has different focal points. This discrepancy can significantly impact identifying specific data types, such as outliers. Therefore, for further research, it is essential to investigate the causes of these differences.

References

Benouareth, A. 2021. An efficient face recognition approach combining likelihood-based sufficient dimension reduction and LDA. *Multimedia tools and applications*. [Online]. **80**(1), pp.1457–1486. [Accessed 5 December 2024]. Available from: https://doi.org/10.1007/s11042-020-09527-9.

Ballard, M.E. and Spencer, M.T. 2023. Importance of Social Videogaming for Connection with Others During the COVID-19 Pandemic. *Games and culture*. [Online]. **18**(2), pp.251–264. [Accessed 5 December 2024]. Available from: https://doi.org/10.1177/15554120221090982.

Ellis, L.A., Lee, M.D., Ijaz, K., Smith, J., Braithwaite, J. and Yin, K. 2020. COVID-19 as 'game changer' for the physical activity and mental well-being of augmented reality game players during the pandemic: Mixed methods survey study. *Journal of medical Internet research*. [Online]. **22**(12), pp.e25117–e25117. [Accessed 5 December 2024]. Available from: https://doi.org/10.2196/25117.

Fuadah, Y.N., Pramudito, M.A. and Lim, K.M. 2022. An Optimal Approach for Heart Sound Classification Using Grid Search in Hyperparameter Optimization of Machine Learning. *Bioengineering*. [Online]. **10**(1), p.45. [Accessed 5 December 2024]. Available from: https://doi.org/10.3390/bioengineering10010045.

Gneiting, T. and Walz, E. 2022. Receiver operating characteristic (ROC) movies, universal ROC (UROC) curves, and coefficient of predictive ability (CPA). *Machine learning*. [Online]. **111**(8), pp.2769–2797. [Accessed 5 December 2024]. Available from: https://doi.org/10.1007/s10994-021-06114-3.

Kumar, S., Gupta, S., Arora, S. and Kumar, S. 2022. A comparative simulation of normalization methods for machine learning-based intrusion detection systems using KDD Cup'99 dataset. *Journal of intelligent & fuzzy systems*. [Online]. **42**(3), pp.1749–1766. [Accessed 5 December 2024]. Available from: https://doi.org/10.3233/JIFS-211191.

Lu, X., Xing, X., Hu, K. and Zhou, B. 2023. Classification and Evaluation of Tight Sandstone Reservoirs Based on MK-SVM. *Processes*. [Online]. **11**(9), p.2678. [Accessed 5 December 2024]. Available from: https://doi.org/10.3390/pr11092678.

Paramasivam, K., Sindha, M.M.R. and Balakrishnan, S.B. 2023. KNN-Based Machine Learning Classifier Used on Deep Learned Spatial Motion Features for Human Action Recognition. *Entropy*. [Online]. **25**(6), p.844. [Accessed 5 December 2024]. Available from: https://doi.org/10.3390/e25060844.

Vuong QH., Ho MT., La VP., Le TT., Nguyen, T.H.T. and Nguyen MH. 2021. A multinational dataset of game players' behaviors in a virtual world and environmental perceptions. *Science Data Bank*. [Online]. [Accessed 1 December 2024]. Available from: https://doi.org/10.11922/sciencedb.j00104.00098.

Xiong, Z., Cui, Y., Liu, Z., Zhao, Y., Hu, M. and Hu, J. 2020. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Computational materials science*. [Online]. **171**, p.109203. [Accessed 5 December 2024]. Available from: https://doi.org/10.1016/j.commatsci.2019.109203.