| Problem Chosen | 2023 MCM/ICM Summary Sheet | Team Control Number |
|---|---|---|
| C | | 2318039 |

# Based on Machine Learning Model to Predict the Wordle Results

Wordle is a popular puzzle currently offered by the New York Times. We conducted data feature mining on users' game reports on Twitter, used KNN, random forest and other models to predict the number of reports, the percentage of attempts and the difficulty of specific words. The research on this issue can provide reference suggestions for game developers to improve their games and help them to become more popular with Wordle.

For question 1, the analysis shows that the Number of reported results is significantly correlated with contest number and word difficulty. Then, KNN, random forest and Bagging algorithms are used to establish regression prediction models for the Number of reported results. Considering the effects of time fluctuations and word difficulty, the combination of the three models results in several reports between 20,100 and 21,000 on March 1, 2023. Finally, the influence of time factor on the percentage of difficult mode report was excluded, and the processed data was quantitatively expressed into different levels of difficulty level of the word in Hard Mode, and the attributes of words were subdivided into four indicators. Through Pearson correlation analysis of attributes and the difficulty level of the word in Hard Mode, the specific impact of four indicators on the reported percentage of difficult mode was obtained.

For question two, by using a control variable for the average number of attempts, we know that the percentage correlation of attempts is closely related to the average number of attempts. Then, we take the attribute of a word as the independent variable and predict the average number of attempts through machine learning [1] models such as random forest, KNN and decision tree, so as to indirectly get the relevant percentage of a specific word. For the word "EERIE", the predicted correlation percentage is 0%, 2%, 13%, 30%, 32%, 19%, 4%.

For problem 3, this paper divides the difficulty of words into four categories according to the average number of attempts, and uses the four indexes of words to make classification prediction based on the above machine learning algorithms. The result was that all three models rated the word "EERIE" at a difficulty level of 4 (hard).

For question 4, we found that the popularity of the game increased and then decreased in 2022, but the loyal players of the game continued to increase. In addition, the game's average attempt data fluctuates within a small range, suggesting that word difficulty does not vary much from day to day. Finally, we found that players were more likely to post better results on Twitter.

**Key Words:** Wordle; KNN; Correlation; Random Forest; Regression

# Contents

# I. Introduction

## 1.1 Background

In some Western countries, crossword have a very long history. In essence, crossword is not just a means of entertainment, but plays a role in cultural transmission and knowledge accumulation, because players need not only a rich English vocabulary, but also the ability to think quickly, simply by chance is not likely to become a high score player.

Wordle was originally designed by a software engineer as a time-wasting game for his partner. After being posted on family chats and the Internet, it attracted more and more attention, until New York Time bought the game, which further promoted Wordle's wider spread. With simple rules, no barriers to entry and suitable for all demographics, including students and office workers, Wordle has attracted more than 2 million users in just a few months from its launch with less than 100 users, frequently appearing on various social media tweets. With only one opportunity per day, it also increases the curiosity and challenge of players, and one round of games does not require a lot of time. This combination of advantages has led to an increase in the number of hits on Wordle and a wave of charades. At the same time, due to the rapid increase in popularity of Wordle, more and more followers of this new genre appeared, and even other countries launched similar versions of Wordle, with a trend of global popularity.

## 1.2 Restatement of the Problem

**Question 1**: According to the results of the report provided in the attachment, build a mathematical model to analyze the daily change trend of the number of Wordle related indicators. The model developed is used to predict what results the report will present on March 1, 2023, and to indicate whether word attributes in the hardship model will have an impact on the reported metrics of interest, with specific explanations.

**Question 2**: Building a mathematical model to classify the difficulty of words using different criteria to identify the properties of different categories of words, and the model was used to analyze the accuracy of the model based on the criteria for classifying the difficulty of the word EERIE.

**Question 3**: Building a mathematical model to predicted number of percentages occupied by (1, 2, 3, 4, 5, 6, X) at a future point in time, and using a specific example of the percentage of the

word EERIE on March 1, 2023 to analyze the uncertainty of the constructed mathematical model and explain the rationality of the prediction.

**Question 4**: According to the data provided, analyze, and summarize other possible features. And write a letter to Puzzle Editor of New York Times based on all the above results.

## 1.3 Our Work

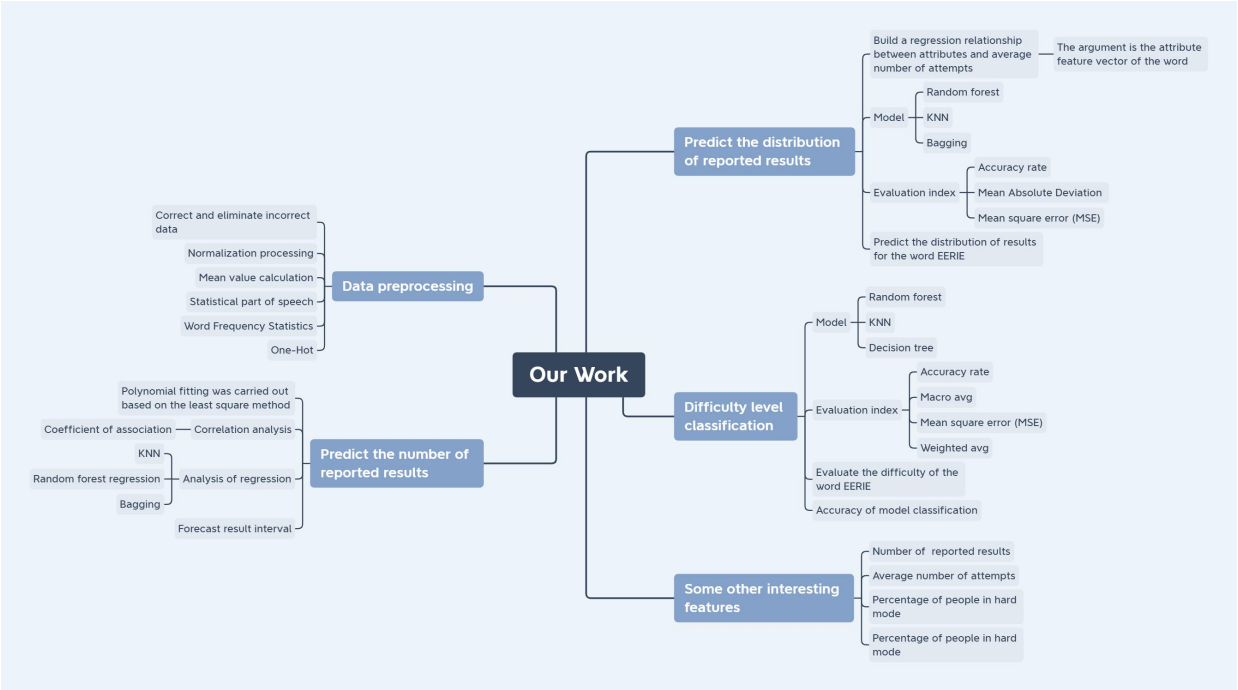



Fig1. Our work

Firstly, the data set is preprocessed, the data that does not meet the requirements is corrected and eliminated, and the percentage of each attempt times is normalized. Calculate the required average value; Statistics of word parts of speech and word frequency; One-Hot coding and so on.

Secondly, for problem 1, the data set was analyzed and the time series diagram of the number of reports was drawn. By observing the time series diagram, it was found that the change trend was first rising and then declining. From the highest point, the data is divided into two sections for discussion, and the first part of the data is taken for polynomial fitting and correlation analysis based on the least square method to verify the hypothesis that the Number of reported results is closely related to the number of contest number and the average number of players' attempts to solve the riddle. Based on this assumption, the data set is divided into the training set and the test set in the ratio of 8:2, and KNN, random forest and Bagging machine

learning are used for regression analysis and prediction. By combining the above three algorithms, the prediction range of the Number of reported results on March 1, 2023 is obtained.

The polynomial regression analysis of contest number and the percentage of the number of reports in Hard Mode shows that there are fluctuations. Therefore, it is reasonable to believe that the attributes of words have an impact on the percentage of the number of reports in Hard Mode. Compare The properties of The Word (Word Frequency, Number of parts of word, The maximum number of occurrences of a word) with the difficulty level of the word in Hard Mode to do correlation analysis. You can tell if certain attributes of a word affect the percentage of Hard Mode reports.

Then, for the second problem, ignoring the differences among game players, the correlation percentage of the number of attempts by players to solve the riddle is only related to the properties of the words themselves. The frequency of a word, the number of parts of speech a word has and the maximum number of times a letter appears in a word are selected as the attributes of a word. Combined with random forest, KNN and decision tree algorithms, the feature vector of word attributes and the average number of attempts were regression analyzed, and the percentage distribution was finally predicted by combining the data set.

Finally, for question 3, the difficulty of words is closely related to the attributes of words, so we establish a model to classify the difficulty of words by using the attributes mentioned in Question 2. The difficulty level is measured by the average number of attempts λ. KNN, decision tree and random forest algorithms are selected to analyze the attributes and difficulty levels. Accuracy, macro avg and weighted avg are selected as evaluation indicators to obtain the difficulty level of specific words.

# II. Assumptions and Justifications

1. **Assume that the maximum number of letters in a word reflects the difficulty factor of the word.** The rules of the game set, as long as the correct letter appears then its tile will be shown as green or yellow, then if a word has two or more of the same letter, after correctly guessing this letter, if its position is correct, then just guess the remaining three or fewer letters can be; If the location is not correct, you can change its position and continue the above steps to get the correct answer.
2. **Assume that the average problem-solving ability of the game population remains the**

**same in 2022.** The factors of individual players are difficult to control and there is no supporting data. In addition, due to the large population base of game players, the differences of game players groups can be ignored.

3. **Assumed that the frequency of words used in everyday life is directly related to the difficulty of solving the problem.** The more frequently a word is used in everyday life, the easier it is for the player to try it in the first place, and the less difficult it is to solve the problem. On the other hand, some unusual words in professional fields are less likely to be tried by players.

# III. Notations

**Table 1 Notations**

| *Symbol* | *Description* |
|---|---|
| $X_n$ | The percentage of crossword that the player guessed for the Nth time before the data was processed |
| $r_i$ | Correlation coefficient |
| $r^2$ | Decision coefficient |
| $x_i$ | Fitted value |
| $i$ | Average number of attempts in ascending order |
| $j$ | The correlation percentage of solving the crossword correctly for the Nth time |

# IV. Part I: Predict the Number of Reported Result

## 4.1 Data Preprocessing

The attached file contains some of the Wordle's game data that players reported on Twitter from January 7, 2022 to December 31, 2022, including the date, contest number, word of the day, the number of people reporting scores that day, the number of players on Hard Mode, and so on.

Considering errors that may occur during data collection, first, we preprocess the data given, and modify or eliminate some data that do not meet the requirements. For example, words with contest number 525 and 314 are only composed of four letters. So we change "tash" to "trash" and "clen" to "clean"; The number of reported results with a contest number of 529 was 2569, which was obviously lower than normal, so it was changed to 25690; The percentage with contest number of 281 is 126% after the sum. It is not in line with reality, so it will be removed.

Considering the rounding of the data, the percentage sum is not 100%, so we need to adjust the percentage sum of the data to 100%. Adjust the percentage $X_n$ of the player's Nth guess of the crossword according to the following formula.

$$X_n = \frac{X_i}{\sum_{i=1}^{7} X_i} * 100\% \quad (1)$$

where for $i \leq 6$, $X_i$ represents the percentage of players who guessed the puzzle for the ith time before the data was processed, $X_7$ represents the percentage of players that could not solve the puzzle in six or fewer tries.

## 4.2 Analyze the Influencing Factors of Report Quantity Variation

By analyzing this data set and plotting a time series of columns of the number of reported results, we can see that from January 7, 2022 to December 31, 2022, the number of people reporting their scores on that day showed an upward and then downward trend and finally stabilized over time. The figure takes mid-February 2022 as the time node, which can be divided into two curves: the rising period and the declining period which eventually become stable. The number of reported results did not decline smoothly over time, but there was a large fluctuation, which we guessed may have a greater relationship with the difficulty of the words on that day.

Take February 9 as the node to divide the data into two paragraphs for discussion. Our guess is that the actual number of reported results is made up of two parts, one is the number of people playing the game that changes smoothly over time, and the other is the fluctuation caused by the difficulty of the words. Next, we will test the hypothesis by taking data from January 7 to February 9.

Firstly, the contest number and number of reported results were polynomial fitted based on the principle of least square method [2], and the number of prediction reports after fitting was denoted as $x_1$. After fitting, the correlation coefficient $r_1$ of the two is 0.91, but it can be seen from Fig2 that there are still many points outside the fitting curve, and this difference is the fluctuation caused by the difficulty of words. In this case, we measured the difficulty of a word by the average number of attempts the player made to guess the word. The formula for calculating the average number of attempts is as follows:

$$N_i = i * \sum_{i=1}^{7} X_i \quad (2)$$

Then, the difference is obtained by subtracting the real value from the fitted value, and the correlation analysis is made between the average number of attempts of players to guess the crossword and the difference, and the correlation coefficient $r_2$ is 0.752. Finally, the polynomial fitting between the average number of attempts and the difference is continued, and the fitted value is denoted as $x_2$. $x_3 = x_1 + x_2$ is the final predicted value of the reported number, as shown in Fig 2. It can be seen that the predicted curve has a high coincidence degree with the real value. By testing the degree of fit between $x_3$ and the real value, the coefficient of determination $r^2$ is 97.8%.

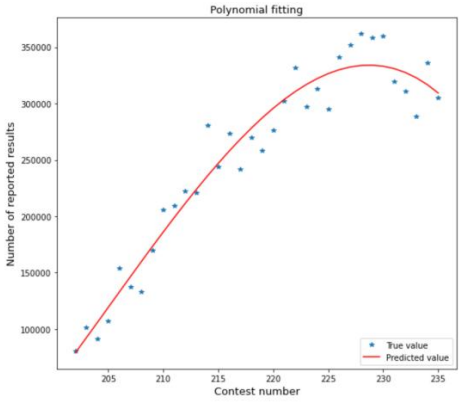



Fig2. Polynomial fitting

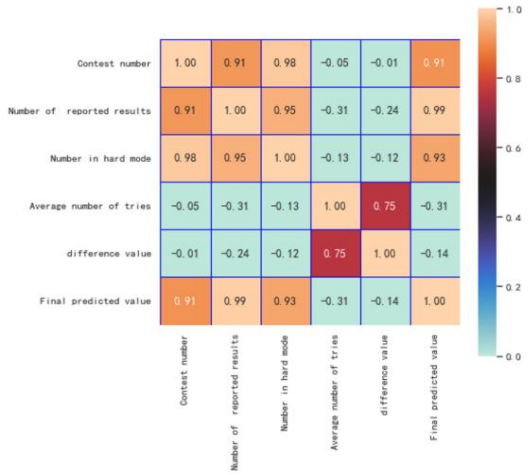



Fig3. Correlation Analysis

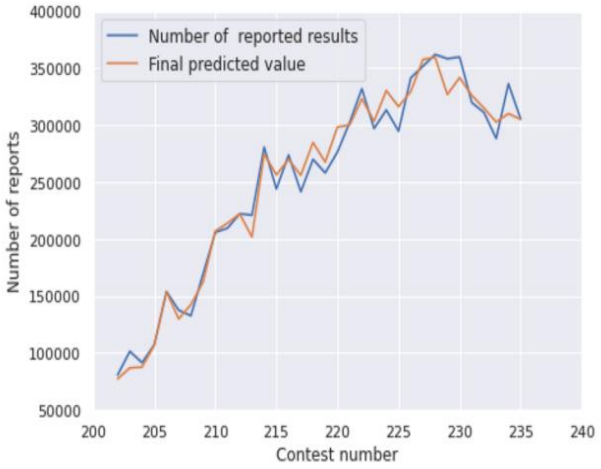



Fig4. Fitting visualization

## 4.3 Regression Prediction of the Total Number of Reports

Based on the attempts in the first part, we can conclude that the above guesses (Number of reported results is closely related to the number of contest and the average number of attempts of players to solve the riddle) are correct. Based on this, contest number and the average number of

players' attempts to solve the riddle in the data from February 10 to December 31 were taken as independent variables, and Number of reported results were taken as dependent variable for regression.

This part of data is divided into training set and test set with a ratio of 8:2, and three machine learning models of KNN, Random forest [3] and Bagging [4] are used for regression analysis respectively. Since contest number on March 1, 2023 was 620, we took x from 618 to 623 in consideration of fluctuations, we take contest number to be 618 to 623, and the average number of attempts on that day is 3.8, 4.2, and 4.6 respectively for regression prediction.

Finally, the verification accuracy of the three models is above 97.5%, so it can be considered that the prediction of the three models is relatively accurate. The Bagging model predicts that the minimum value of number of reported results is 20500 and the maximum value is 21000. The Random Forest model predicts that the minimum value of number of reported results is 20100 and the maximum value is 20750. The KNN model predicted a minimum of 20,700 and a maximum of 20,800 for the number of reported results that day.

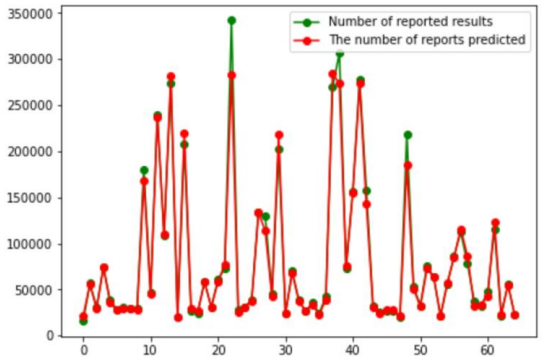



Fig5. Bagging Regressor' s score：0.983032

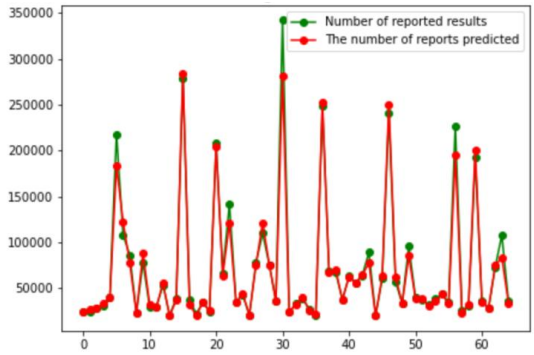



Fig6. Random Forest Regressor's score：0.975511

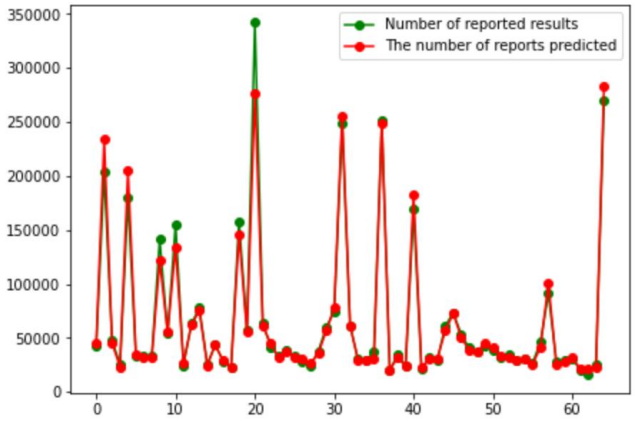



Fig7. KNeighbors Regressor's score：0.975511

Combining these three models, we believe that the forecast range for the number of reported results on March 1, 2023, is 20100 to 21000.

## 4.4 The Effect of Word Attributes in Hard Mode

According to the data set, we can observe that the percentage of Hard Mode reports increases and fluctuates slightly over time. We took contest number as the independent variable and the percentage of Hard Mode reports as the dependent variable, and conducted polynomial regression analysis on them. The results are shown in the figure below:

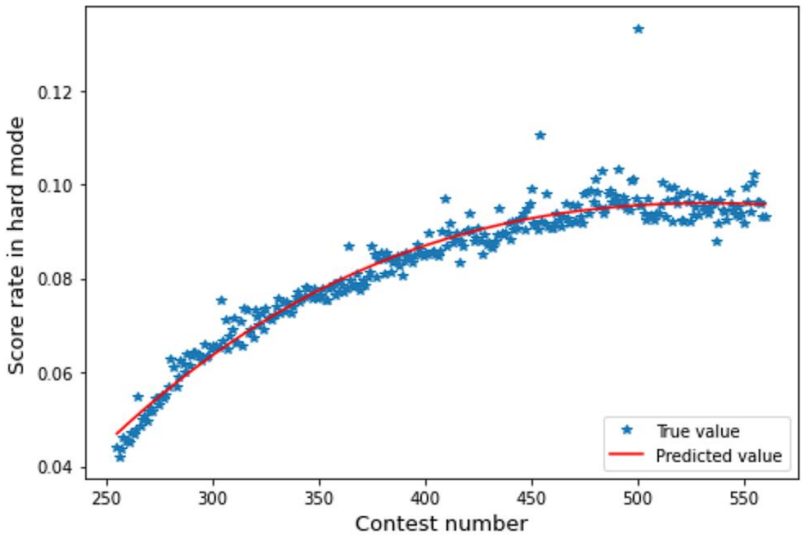



Fig8. Polynomial fitting

It can be seen that most of the data do not fall on the curve, but slightly fluctuate up and down the fitting curve. Therefore, combined with Section 4.1, it is reasonable to believe that the attribute of the word causes the slight fluctuation.

We measure the difference between score rate in Hard Mode and the fitting curve to see how far the real data point deviates. The point where score rate in Hard Mode is above 2% on the fitted curve defines the difficulty level of the word in Hard Mode as 4; 0% to 2% is 3; -2% to 0% is 2, and less than -2% is 1.

Here, we subdivide the attributes of words into the following four categories: frequency of word occurrence [5], simplicity of word, maximum number of occurrences of the same letter in a word, and number of parts of speech of a word.

The frequency of words refers to the specific division of frequency in the authoritative English Dictionary Collins English Dictionary, which divides the frequency of words into five levels from low to high, with the lowest being 1 and the highest being 5. The degree of simplicity

of a word is determined according to the requirements of Chinese curriculum standards for words at different stages. If the word is a primary school curriculum standard word, the degree of simplicity is 5; Similarly, the simplicity levels of junior high school curriculum standard words, senior high school curriculum standard words, CET-4 and CET-6 curriculum standard words and words that do not appear above are respectively 4, 3, 2 and 1. The data of word parts of speech still refer to the Collins English Dictionary, which divides the data given into the most common four categories: verbs, adjectives, nouns and adverbs.

Analyze the correlation between the word attributes and the difficulty level of the word in Hard Mode, as shown in the figure, the vertical and horizontal represent: Contest number, The difficulty level of the word in Hard Mode, Word frequency, The maximum number of occurrences of a word, Number of parts of word.

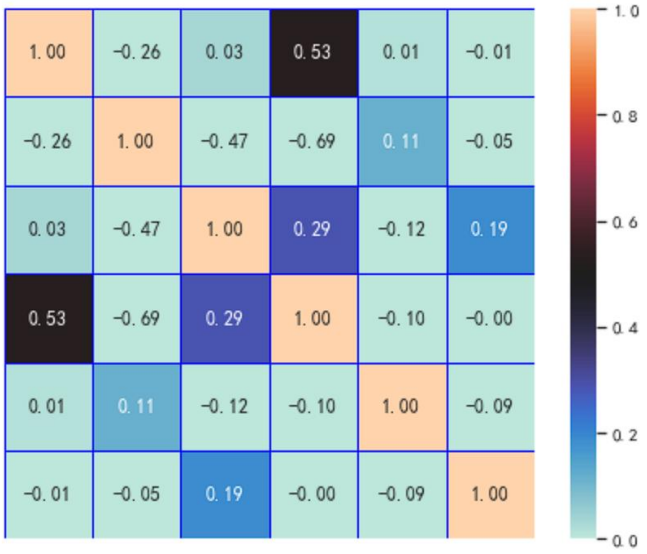



Fig9. Correlation Analysis

It can be seen that the correlation between the frequency of words and the difficulty level of the word in Hard Mode is -0.47, according to Pearson correlation coefficient [6], the two are moderately correlated. The simplicity of the words and the difficulty level of the word in Hard Mode showed a strong correlation of -0.69. The correlation between the number of parts of word and the difficulty level of the word in Hard Mode is -0.05. The maximum number of occurrences of a word and the difficulty level of the word in Hard Mode are 0.11, which can be considered as very weak correlation between them.

Therefore, it can be concluded that some attributes of words will affect the reported number percentage of Hard mode. Among the four attributes of words, the simplicity of the words and

the frequency of words have a great influence on it, and the other two attributes have little influence on it. And the larger the simplicity of the words and the smaller the frequency of words, the lower the percentage of Hard Mode reports.

# V. Part Ⅱ: Predict the Distribution of Reported Results

## 5.1 Word Attribute Division and Basis

The number of Wordle players has changed over time, with several players quitting and new players joining. Based on hypothesis 2, it can be assumed that the average problem-solving ability of players will remain the same in 2022 since population factors are difficult to control. At the same time, the correlation analysis between contest number and the average number of attempts of players shows that the correlation coefficient is -0.05. The two are very weakly correlated, so hypothesis 2 can be considered correct.

Therefore, it is reasonable to assume that the correlation between the percentage of attempts a player makes to solve the riddle is only related to the properties of the word itself. According to 4.3, the attributes of words are divided into the following four types: frequency of words, simplicity of words, maximum number of the same letter in a word, and number of parts of speech of words.

Also, we noticed that when the average number of attempts was similar, the percentage of attempts used to guess the word was similar. For example, the words with contest numbers 247, 345, 407, 459, 492, and 544 all have an average number of attempts of 4.00, moreover, the corresponding percentage of players' answers to the riddle at the Nth $(n = 1, 2, 3, 4, 5, 6, 7)$ times is tightly clustered within a cell. However, several groups of data with a relatively large gap in the average number of attempts, such as words with contest number of 322,363,411,425, have an average number of attempts of 4.14. They correspond to the percentage of players who got the answer to the riddle the Nth time, also in one cell, but the data distribution of the percentages differs from that of the example above.

Comparing multiple sets of data and calculating the standard deviation [6] shows that groups with similar average number of attempts have similar percentages of the number of attempts the player used to guess the word. And because the relevant percentages add up to a constant value of 100%, when the difficulty of the word increases, the proportion of attempts less than or equal to 3 will decrease, and the proportion of attempts greater than or equal to 4 will increase.

The explanation for this is as follows: Due to the large number of Wordle players, the average number of attempts per word is a good indicator of the average difficulty of the word for all players. However, according to hypothesis 2, we ignore the difference of game player groups, that is, we believe that the distribution of players with different problem-solving abilities is unchanged, so when the difficulty of the word is the same or very similar, the percentage of players who get the answer right the Nth time should be similar.

Therefore, we map the predicted average number of attempts to the percentage of the corresponding player's Nth correct answer as follows:

1. If the predicted average number of attempts occurs in a given data set, the corresponding 7 correlation percentages in the data set are averaged respectively as the predicted correlation percentage under the average number of attempts.
2. If it is predicted that the average number of attempts does not appear in the given data set, we average the average number of attempts in the data set and the seven correlation percentages corresponding to the two groups of data closest to this group respectively as the predicted correlation percentage under this average number of attempts, and then obtain the percentage under this number of attempts according to formula 1. The formula is:

$$x_{ij} = \frac{x_{i-1,j} + x_{i+1,\ j}}{2} \qquad (3)$$

Where i is the serial number in ascending order of average attempts, and j is the relevant percentage of players' guessing the answer to the riddle at the Nth ($n = 1, 2, 3, 4, 5, 6, 7$) times.

## 5.2 Using the Random Forest, KNN, Decision Tree for Regression Analysis

Based on the above analysis, the problem is simplified to a regression relationship between the attributes of words and the average number of attempts of words. The four indexes of the word were used to construct the attribute feature vector of the word, which was used as the independent variable of regression analysis. The average number of attempts made by the players was used as the dependent variable in the regression analysis.

Firstly, the data set was divided into the training set and the test set in the ratio of 8:2. Choose the decision tree, KNN and random forest three models training in the training set, and then test the test set, accuracy, the mean square error (MSE) and mean absolute error (MAE) as evaluation indexes. The experimental results are shown in the figure:

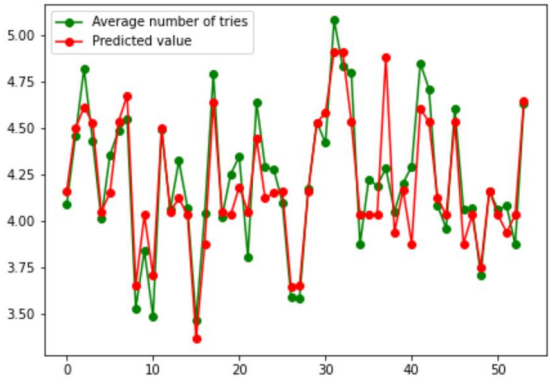



Fig10. Decision Tree Regressor's score:0.798037

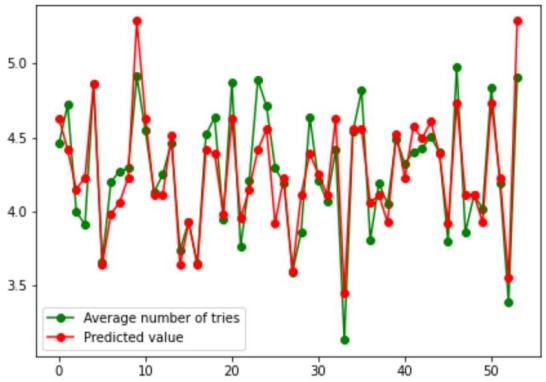
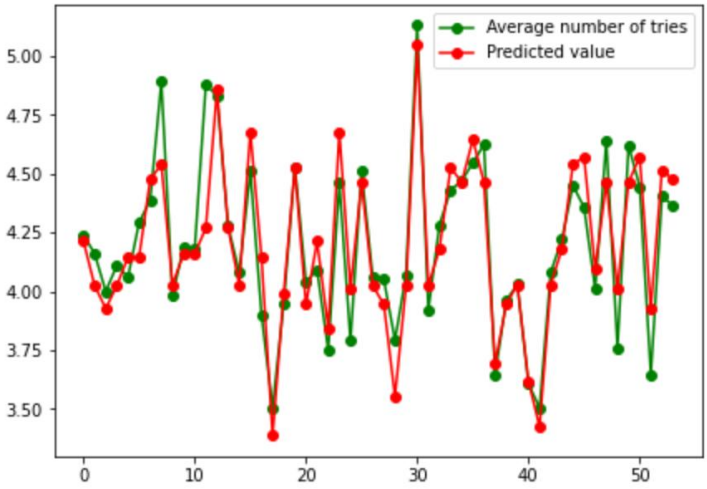



Fig11. KNeighbors Regressor's score:0.799474




A

Fig12. Random Forest Regressor's score:0.816819

The MSE and MAE of decision tree, KNN and random forest are shown in the following table:

**Table 2 The relevant data of the model**

| | Accuracy | MSE | MAE |
|---|---|---|---|
| **Decision Tree** | 0.798037 | 0.0278 | 0.129 |
| **KNN** | 0.799474 | 0.0282 | 0.133 |
| **Random Forest** | 0.816819 | 0.0234 | 0.113 |

## 5.3 Predict the Relative Percentage of EERIE

For the word EERIE query, the Collins English Dictionary has the frequency level of two stars, the difficulty level of 1 (CET-8 vocabulary), the number of repeated letters of 3, and the number of parts of speech of 1. The average number of attempts to predict using the three models to do regression is 5.21, 4.59, 4.67 respectively. Here, we select the KNN model with the highest

accuracy and predict that the correlation percentage of player attempts corresponding to the word EERIE is 0%, 2%, 13%, 30%, 32%, 19%, 4% through the mapping rules specified in 5.1.

According to the accuracy rate of KNN on verification set and the standard deviation value calculated for the relevant percentage data near it, the average value standard deviation is 1.33. Therefore, we have reason to believe that the error between the relevant percentage and the true value of this prediction should be within 1.5.

# VI. Part Ⅲ: Difficulty Level Classification

## 6.1 Word Classification and Regression Analysis

For question three, we need to build a model to classify the difficulty of words. The difficulty of a word is closely related to its attributes, so we use the attributes mentioned in 4.3 to classify the difficulty of words.

The average number of attempts used by the players to solve the riddle is denoted as λ. Based on the 358 sets of data, the average number of times players used to guess the word in this period of time is 4.20 times. Considering the accuracy of classification, the difficulty of words can be divided into four categories according to the size of λ : when λ⩽3.7, the difficulty level is 1 (simple); When $3.7 < \lambda \leq 4.1$, the difficulty level is 2 (relatively easy); When$4.1 < \lambda \leq 4.4$, the difficulty level is 3 (relatively difficult); When $4.4 < \lambda$, the difficulty level is 4 (difficult).

Next, using the four indicators of word attributes, construct word attribute characteristic vector, as the regression analysis of the independent variable; The feature vector that measures how difficult the word is for the player to solve the problem is the difficulty level of the processed word, and serves as the dependent variable for regression analysis.

The data set is divided into training set and test set at a ratio of 9:1. We still choose decision tree, KNN [8] and random forest models for training on the training set, and make classification prediction on the test set. Accuracy, macro avg and weighted avg [9] are selected as indicators of the evaluation model. Specific experimental results are shown in the figure below:

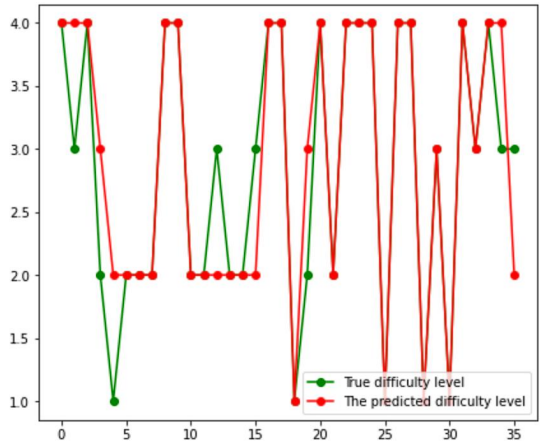



Fig13. KNeighbors Classifier's score:0.777778

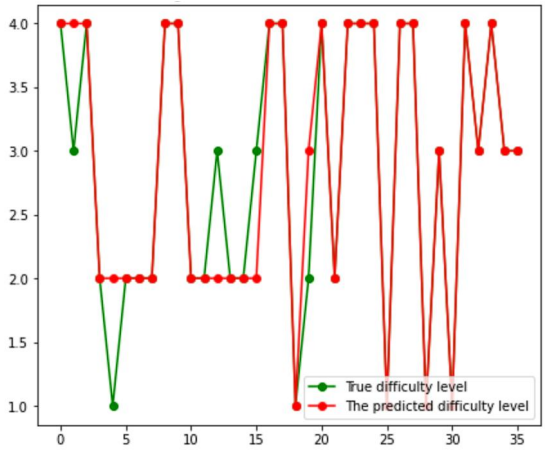



Fig14. Decision Tree Classifier's score:0.861111

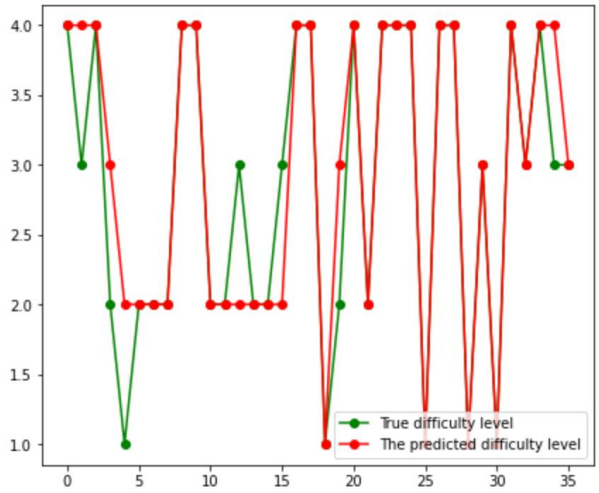



Fig15. Random Forest Classifier's score:0.805556

**Table 3 The relevant data of the model**

| | Accuracy | macro avg | weighted avg |
|---|---|---|---|
| **Decision Tree** | 0.78 | 0.73 | 0.76 |
| **KNN** | 0.86 | 0.83 | 0.86 |
| **Random Forest** | 0.81 | 0.77 | 0.80 |

The attribute of the word EERIE has been explained in Part Ⅱ. Using the above three models to do classification prediction, it turns out that the difficulty of the word is 4 (hard).

# VII. Part Ⅳ: Other Salient Features of the Data Set

## 7.1 Number of Reported Results

By analyzing the data set and drawing the time series graph of the Number of reported results column, as shown in Figure 16, we can see that from January 7, 2022 to December 31, 2022, the number of people reporting scores on that day showed a trend of first rising and then declining over time. Starting in January 2022, the popularity of the game increased and the number of people reporting scores continued to rise, peaking on February 4 at 359,679, before gradually declining and eventually leveling off.

As can be seen from Figure 16, the curve is not smooth in the process of rise and fall. Therefore, we guess that many factors, such as whether the day is a working day or not and the difficulty of words on that day, have an impact on the number of people reporting scores.

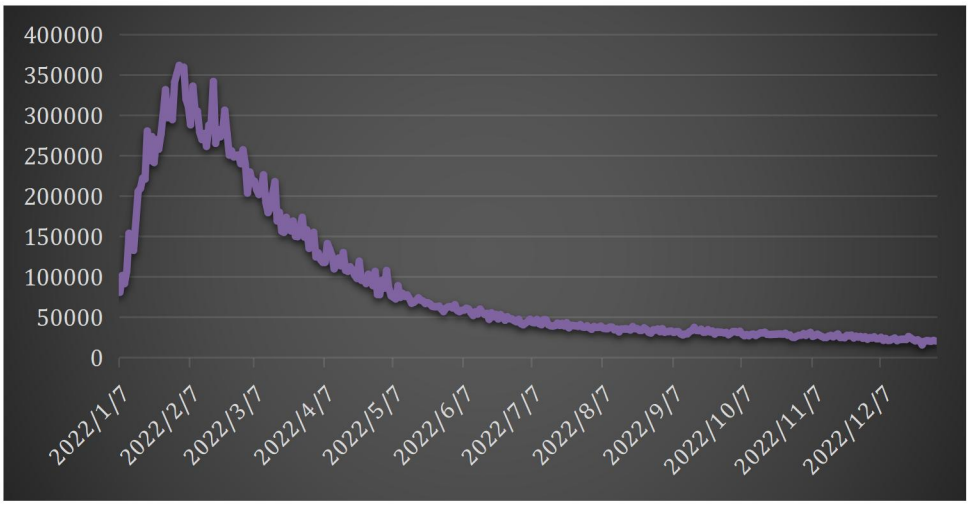



Fig16. Number of reported results

## 7.2 Average Number of Attempts

Plotting the average number of attempts per day， as shown in Figure 17, we found that the average number of attempts for the game fluctuated within a small range, with an average of around 4. This shows that the difficulty of the word does not vary much from day to day, which makes it easier to attract players than games with varying difficulty, which may be one of the reasons for the game's rapid popularity.

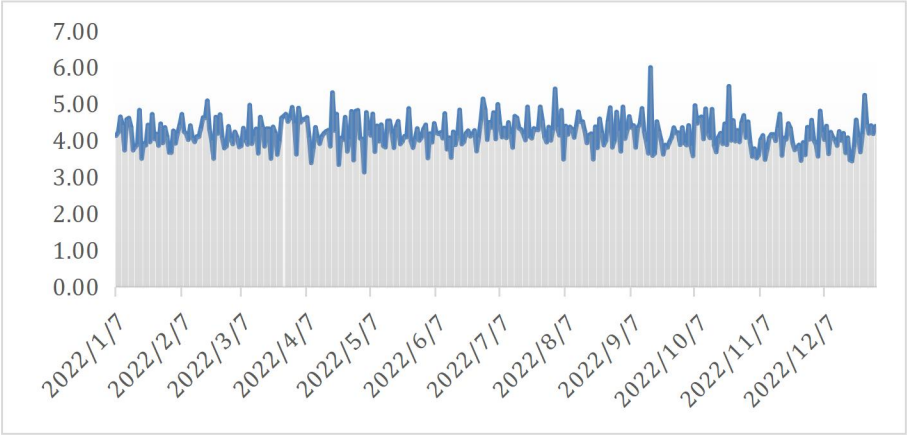



Fig17. Average number of attempts

## 7.3 Percentage of People in Hard Mode

The percentage of players in Hard Mode is shown in Figure 18. Over time, the percentage of players entering Hard Mode generally increased in 2022, initially increasing gradually from 2%, then dropping slightly towards the end of the year, but eventually settling at around 10%. This shows that the player has some experience and has mastered some techniques. And those who end up staying in Hard Mode are loyal players of the game.

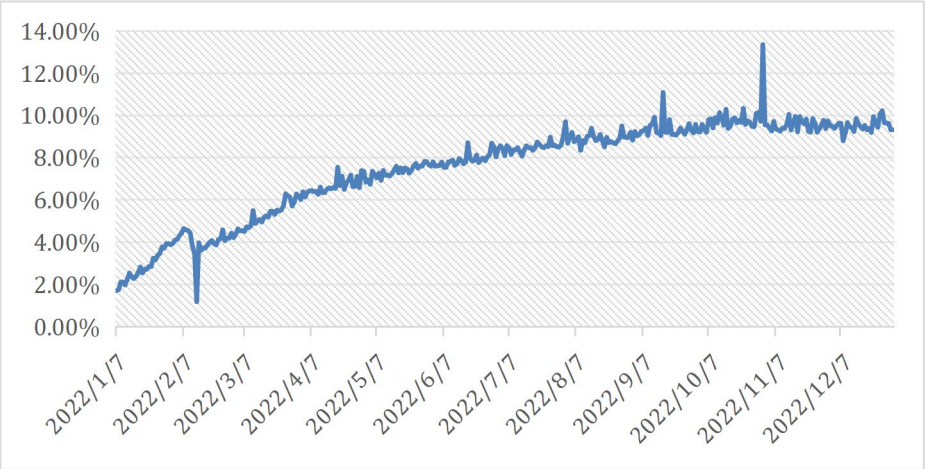



Fig18. Percentage of people in Hard Mode

## 7.4 The Proportion of Attempts in Shared Data

The average number of attempts within the statistical period is used as the representative to draw a pie chart of the proportion of various attempts,as shown in Figure 19. We found that players are more likely to post good results on Twitter. The percentage of people who tried more than seven times in the published data was small, with most of the data concentrated on four or five attempts.

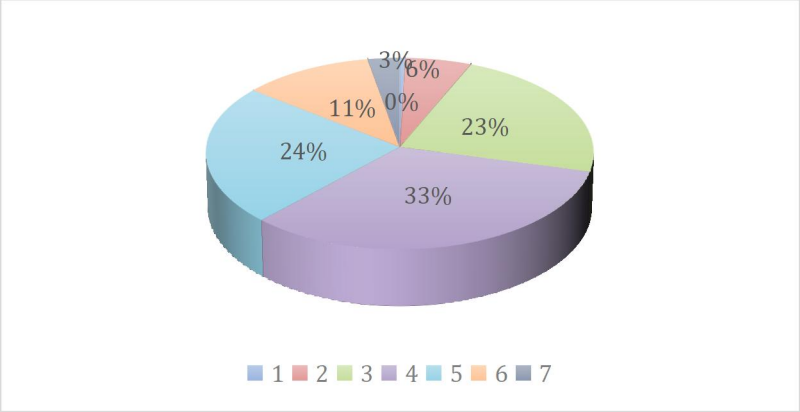



Fig19. The proportion of attempts in shared data

# Ⅷ. Stability Analysis

In question 2, considering that the indexes of words may have different values under different references, we randomly selected 18 (5%) groups of data from 357 groups to raise or lower the Frequency of use or Degree of difficulty by one level. The regression analysis between the attributes of a word and the average number of attempts is shown in the following figure:

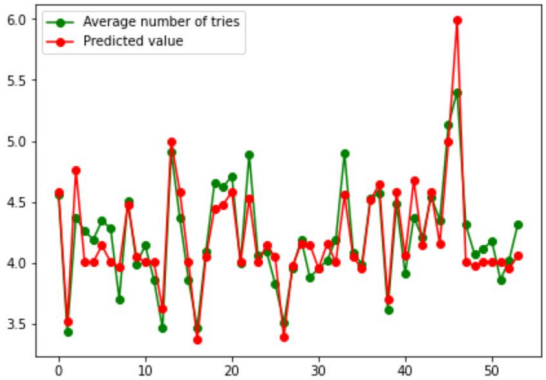



Fig20. Decision Tree Regressor's score:0.789071

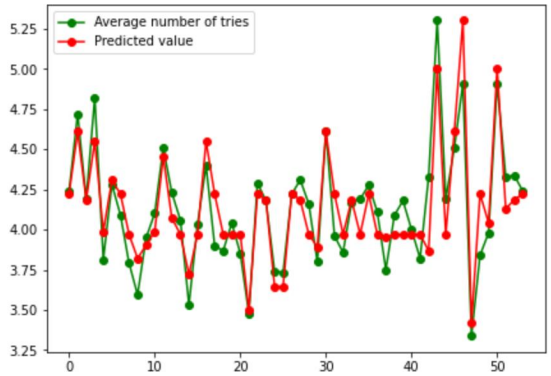



Fig21. KNeighbors Regressor's score:0.787504

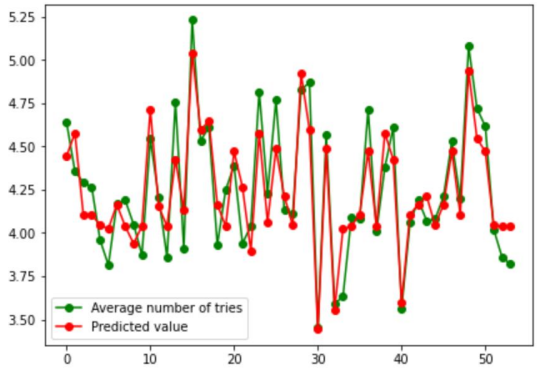



Fig22. Random Forest Regressor's score:0.814852

The MSE and MAE of decision tree, KNN and random forest after adding disturbance are 0.0355 and 0.149; 0.0290, 0.135; 0.0280, 0.140 respectively.

**Table 4 The relevant data of the model**

| | Accuracy | MSE | MAE |
|---|---|---|---|
| **Decision Tree** | 0.789071 | 0.0355 | 0.149 |
| **KNN** | 0.787504 | 0.0290 | 0.135 |
| **Random Forest** | 0.814852 | 0.0280 | 0.140 |

# Ⅸ. Model Evaluation

## 9.1 Strengths

1. The least square method can be realized by a simple computer program, which reduces the difficulty in the fitting process and makes it easier to calculate unknown data. Moreover, it can obtain the best functional relationship by minimizing the sum of squared errors.
2. The KNN model has a relatively low complexity and simple structure, which can be used for classification or regression and is not easily affected by outliers in the data. Due to the small data set, the proximal point has the highest accuracy in the above experiments.
3. Random forest model has fast training speed, can mention the importance of variables, and has good anti-noise ability, which is not easy to overfit.
4. The Bagging model can sample from the dataset in a put-back manner, improving generalization error and reducing overfitting by reducing the variance of the base classifier.

## 9.2 Weaknesses

1. When using the above model to make regression or classification prediction, the accuracy of the model is not very high due to the low correlation between the feature vectors of the independent variable and the dependent variable. In the future, the feature vectors of word attributes can be further mined.
2. When the data set is randomly divided into training set and verification set, the accuracy of the model will fluctuate slightly due to different data.

# Ⅹ. A Letter to the Editor of New York Times

Dear Puzzle Editor,

We are writing to share the results of our team's analysis of the New York Times Wordle Game user statistics file of daily results. In view of the problem you raised, we used relevant knowledge of data mining to preprocess the data of the file, and then combined KNN, random forest, decision tree and other machine learning algorithms to build regression and classification models for detailed processing and analysis, and finally got relatively reliable results. The specific analysis process and conclusions are as follows:

For problem 1, the data set is analyzed and the time series diagram of the number of people reporting is drawn. From the highest point, the data was divided into two sections for discussion. The first part of the data was selected for polynomial fitting based on the least square method, and the correlation analysis showed that the Number of reported results was closely related to the number of contest and the average number of players' attempts to solve the riddle. Based on the above relationship, the data set is divided into the training set and the test set in the ratio of 8:2, and KNN, random forest and Bagging machine learning are used for regression analysis and prediction. By combining the above three algorithms, the prediction range of the Number of reported results on March 1, 2023 is between 20100 and 21000.

Excluding The influence of time factor on the reported percentage of difficult mode, Pearson correlation analysis was made on the four indicators of attributes and The difficulty level of the word in Hard mode, and the specific influence of the four indicators on the reported percentage of difficult mode was obtained. The conclusion is as follows: Some attributes of a word will affect the percentage of reported hard mode quantity. Among The four attributes of a word, The simplicity of the words and The frequency of words have great influence on it, while the other two attributes have little influence on it. And The simplicity of the words and The frequency of words are smaller, the percentage of the number of hard mode reports is lower.

For the second problem, the differences among game players are ignored, and the correlation percentage of players' attempts to solve the riddle is only related to the properties of the words themselves. The frequency of a word, the number of parts of speech a word has and the maximum number of times a letter appears in a word are selected as the attributes of a word. Combined with random forest, KNN and decision tree algorithms, the feature vector of word attributes and the average number of attempts were regression analyzed, and the percentage distribution was finally predicted by combining the data set. We selected the KNN model with

the highest accuracy and predicted that the correlation percentage of player attempts corresponding to the word EERIE was 0%, 2%, 13%, 30%, 32%, 19% and 4% through the mapping rules specified in 5.1.

For question 3, the difficulty of words is closely related to the attributes of words, so we establish a model to classify the difficulty of words by using the attributes mentioned in Question 2. The difficulty level is measured by the average number of attempts λ. KNN, decision tree and random forest algorithms are selected to analyze the attributes and difficulty levels. Accuracy, macro avg and weighted avg are selected as evaluation indicators to obtain the difficulty level of specific words. The attribute of the word EERIE has been explained in problem 2. Three algorithms are used to classify it, and all three algorithms consider it to be 4 (difficult).

For problem four, we drew many graphs to analyze other interesting features. We found that the popularity of the game increased and then decreased in 2022, but the number of loyal players continued to increase. In addition, the game's average attempt data fluctuates within a small range, suggesting that word difficulty does not vary much from day to day. Finally, we found that players were more likely to post better results on Twitter.

The above is the result of our mathematical modeling and solution for the question you raised. We believe that the research on these issues can provide some useful suggestions for game developers to improve their games and help them become more popular with Wordle.

Thank you very much for your consideration!

Sincerely Yours,

Team #2318039

# References

[1] Zhihua Zhou. Machine Learning [M]. Beijing: Tsinghua University Press, 2016

[2] Xinyu Qian. Research on Phase Operation of Synchronous Generator Based on least Square Method [D]. Shenyang university of technology, 2021. DOI: 10.27322 /, dc nki. Gsgyu. 2021.000058.

[3] Xinhai Li. Application of stochastic forest model in classification and regression analysis [J]. Chinese Journal of Applied Entomology,2013,50(04):1190-1197. (in Chinese)

[4] Jiajia Zhang, Fuliang Yi, Hui Yang et al. Multi-classification prediction of Alzheimer's disease progression based on Bagging [J]. Chin J Health Statistics,202,39(05):675-679+684.

[5] Hang Zhao, Ji Ma, Fuchun Zhang et al. Application of Variance Analysis to study the use frequency of different initial letter words in CET [J]. English square (academic), 2012, No. 024 (12) : 87-88. The DOI: 10.16723 / j.carol carroll nki yygc. 2012.12.091.

[6] Xiaoyong Yang. Analysis of Variance Analysis -- Single Factor Analysis of Variance [J]. Experimental Science and Technology,2013,11(01):41-43.

[7] Ying Chang. Study on laser ultrasonic characteristics of solids with different structures and Defect Detection based on Pearson Correlation Coefficient [D]. Northwestern polytechnical university, 2019. DOI: 10.27406 /, dc nki. Gxbgu. 2019.000456.

[8] Ying Bin Sang. Research on Classification Algorithm Based on K-Nearest Neighbor [D]. Chongqing University,2009.

[9] micro avg, macro avg, and weighted avg, and their differences -- blog weighted avg and CSDN blog weighted avg