Essay Writing

Q3

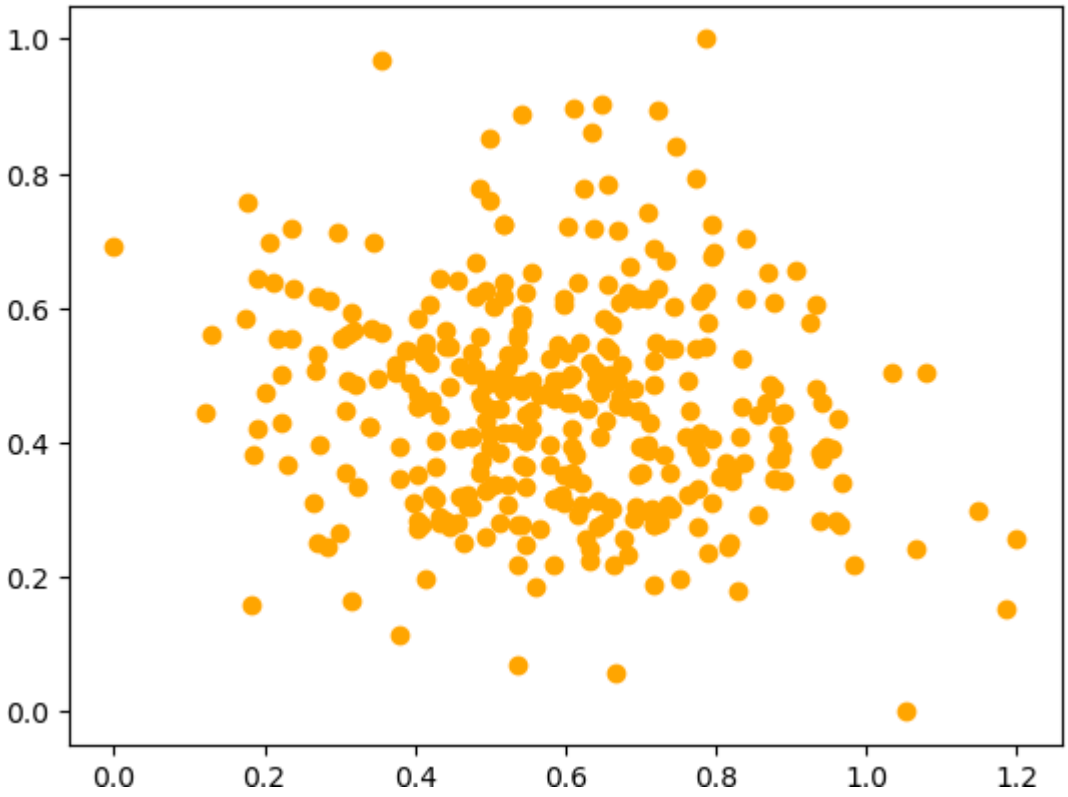
Develop and summarize a model to classify solution words by difficulty. Identify the attributes of a given word that are associated with each Using your model, how difficult is the word EERIE? Discuss the accuracy of your classification model.

A model is developed and summarized to classify solved words by difficulty. Identify the attributes of a given word associated with each classification. Using your model, how difficult is the word EERIE? Discuss the accuracy of your classification model.

Problem Analysis

- □□ Classification by difficulty (cluster analysis using percentages and included attributes)
- □□ Identify the attributes of a given word associated with each category (use word attributes to map the categories with a classification model)
- □□ How hard is the word EERIE? (put into classifier for classification)
- □□ Discuss the accuracy of your classification model (how well the model performed on the test set)

Scatterplot of key indicators



The third bit asks us to classify, but there is no pre-given difficulty level in the dataset.

So as a first step, we use cluster analysis to perform unsupervised Kmeans clustering of means and standard deviations by distance, and fit a histogram of probability distributions to visually classify the difficulty

In the second step, we use a random forest classification model for the five metrics of the word attributes (listed), and the clustering formed in the first step
target is mapped and a classifier is fitted to facilitate the subsequent prediction of words.

In the third step, we evaluate the word attributes of EERIE and then put them into the trained random forest model to make predictions and give a difficulty characterization of the word EERIE.

Model Principle

K-means clustering is a common unsupervised learning algorithm for partitioning a dataset into several different clusters (clusters), where data points within each cluster are more similar and those between different clusters are less similar. the optimization goal of the K-means algorithm is to minimize the sum of the distances between each data point and the cluster centers of the cluster to which it belongs (called SSE, Sum of Squared Errors). Because
This, the main steps of K-means clustering include choosing the appropriate k-value, initializing the cluster centers, assigning data points to the clusters, computing the cluster centroids, and iterative execution until convergence is reached.

The principles of random forest classification are the same as those of the random forest regression used in the second question, and will not be repeated here.

Model Assumptions

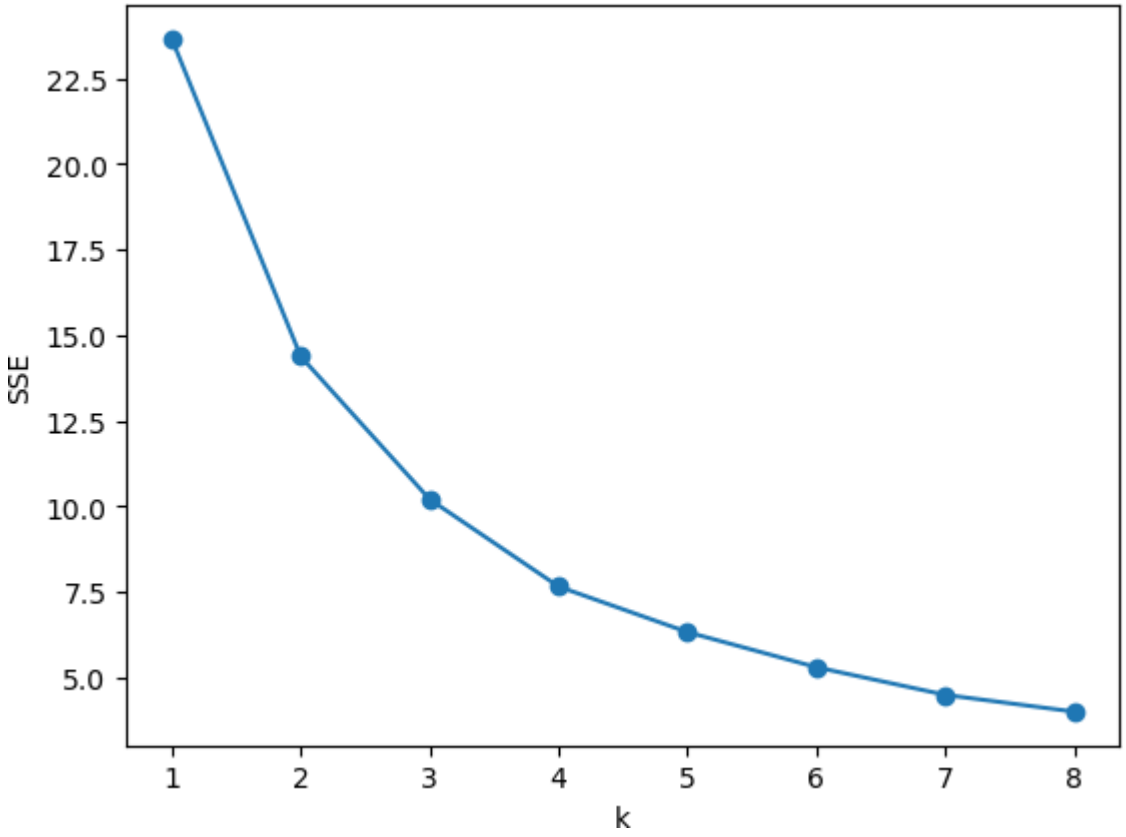
- □□ assumes that the difficulty of a word is described only by the percentage distribution of the number of reports, independent of other factors such as the percentage of difficulty patterns selected
- □□ assumes that the percentage distribution of the number of reports is close to a normal distribution and can be well described by the mean and standard deviation alone.
- □□ Assuming that the mean of the reported number distribution is a better indicator of word difficulty than the standard deviation, the weighting ratio of mean to standard deviation in Kmeans analysis is 6:5.
- □□ Assuming that noise due to chance does not mask the characteristics of the standard deviation and mean

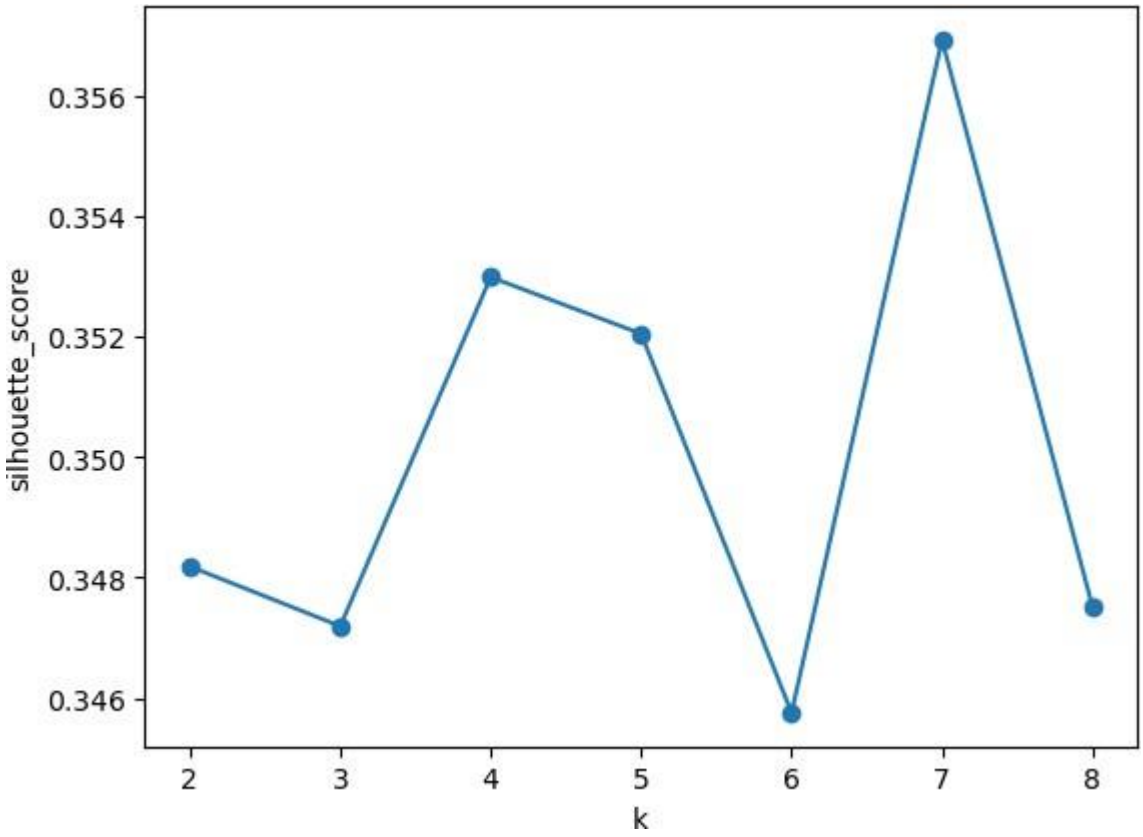
Main Models

Kmeans clustering analysis model is used first

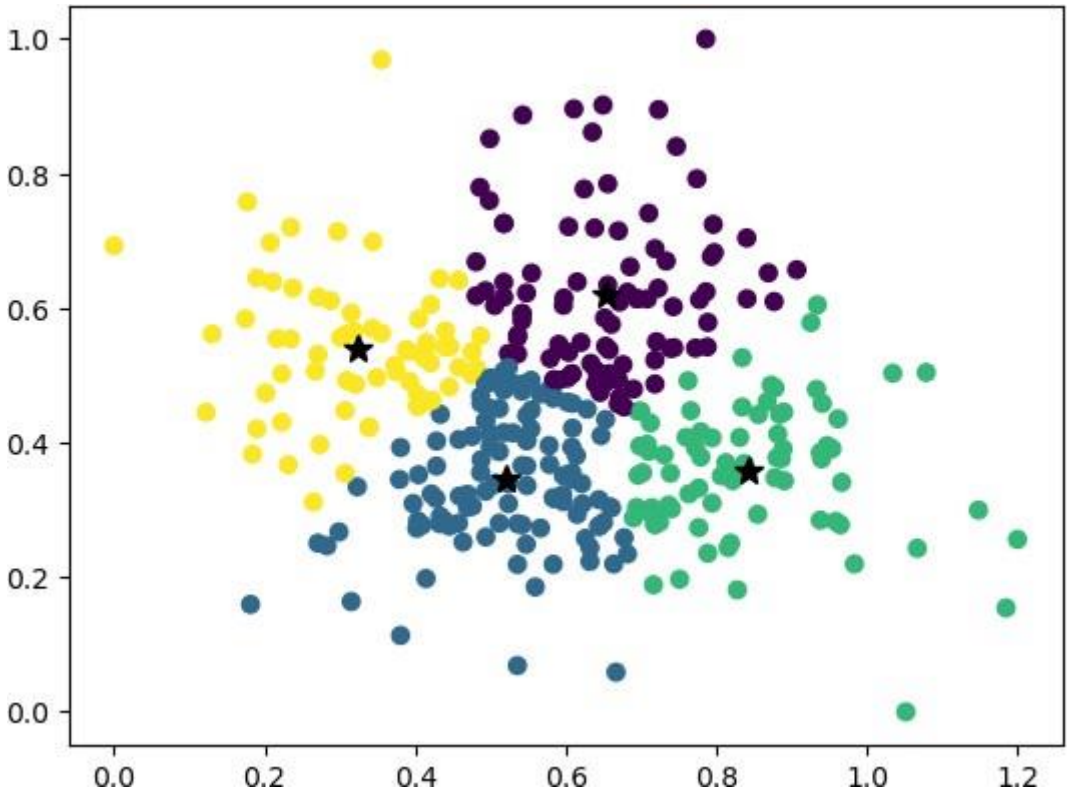
□□ data were normalized and the means were multiplied by 1.2, making the mean to standard deviation weighting ratio 6:5.

□□ Determine the optimal number of clusters as 4 using the elbow method with contour coefficients (one figure on each side)



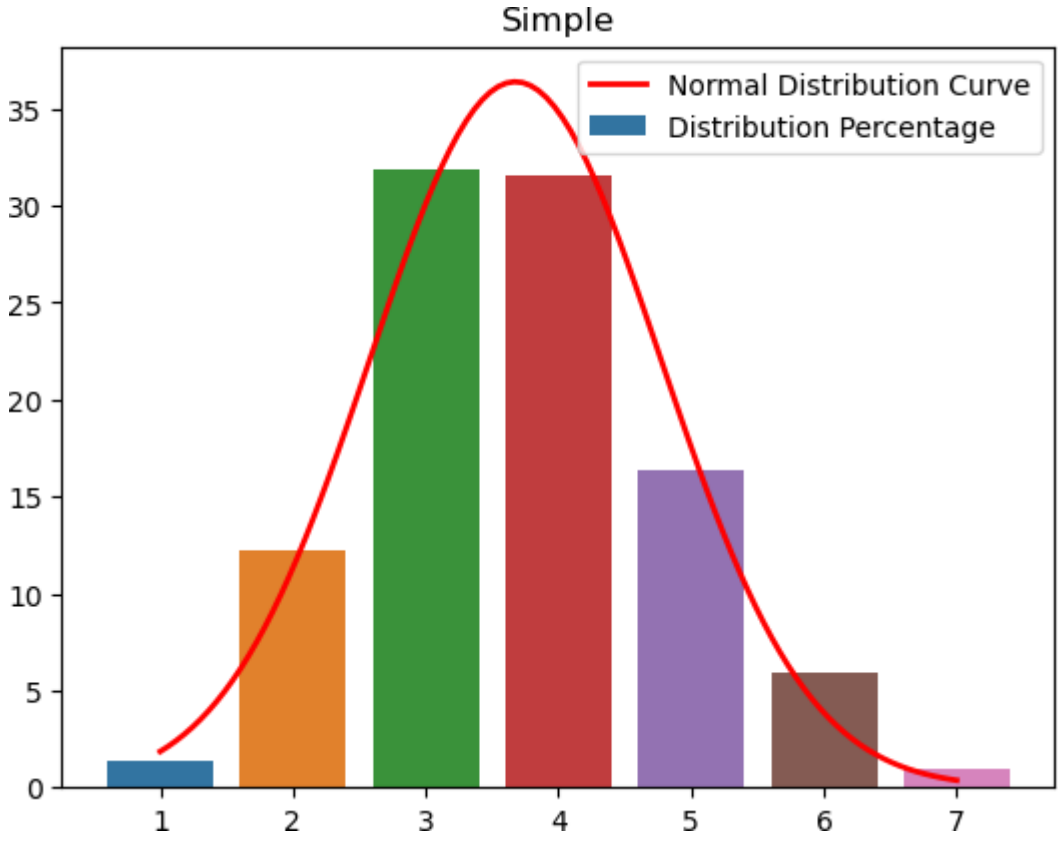


□□ Clustering results (points marked with an asterisk are clustering centers)

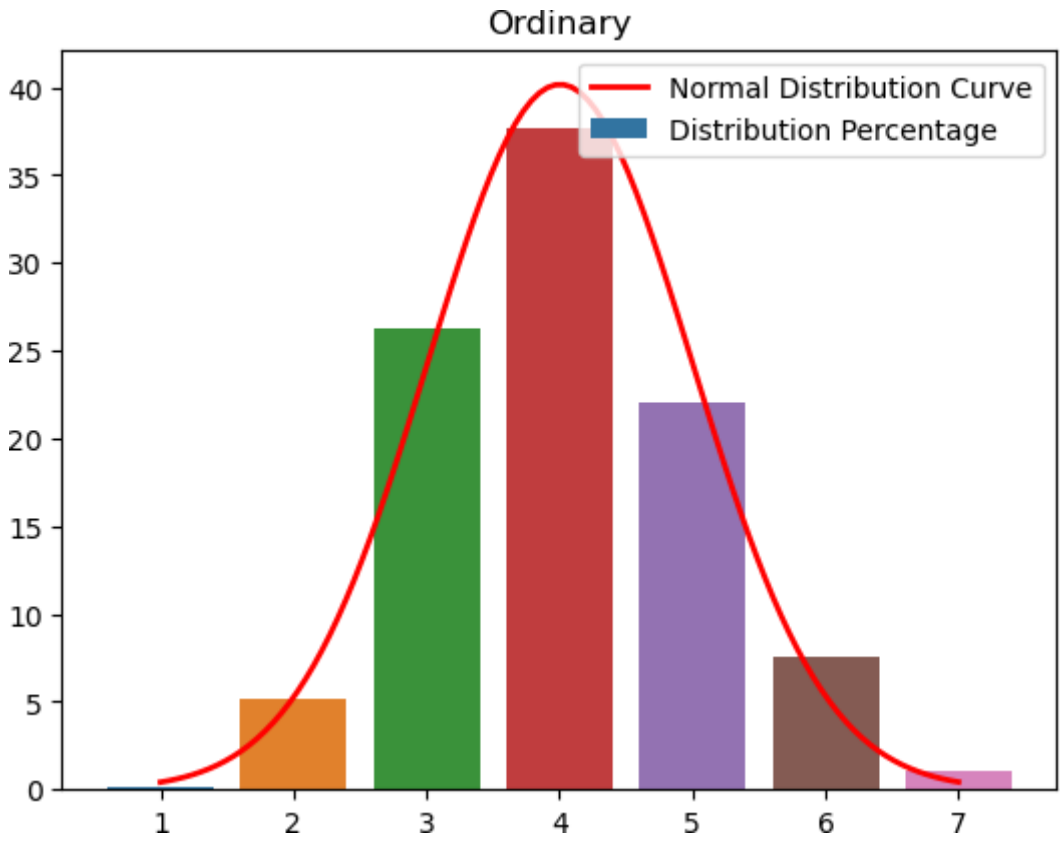


□□ summarize the different difficulty

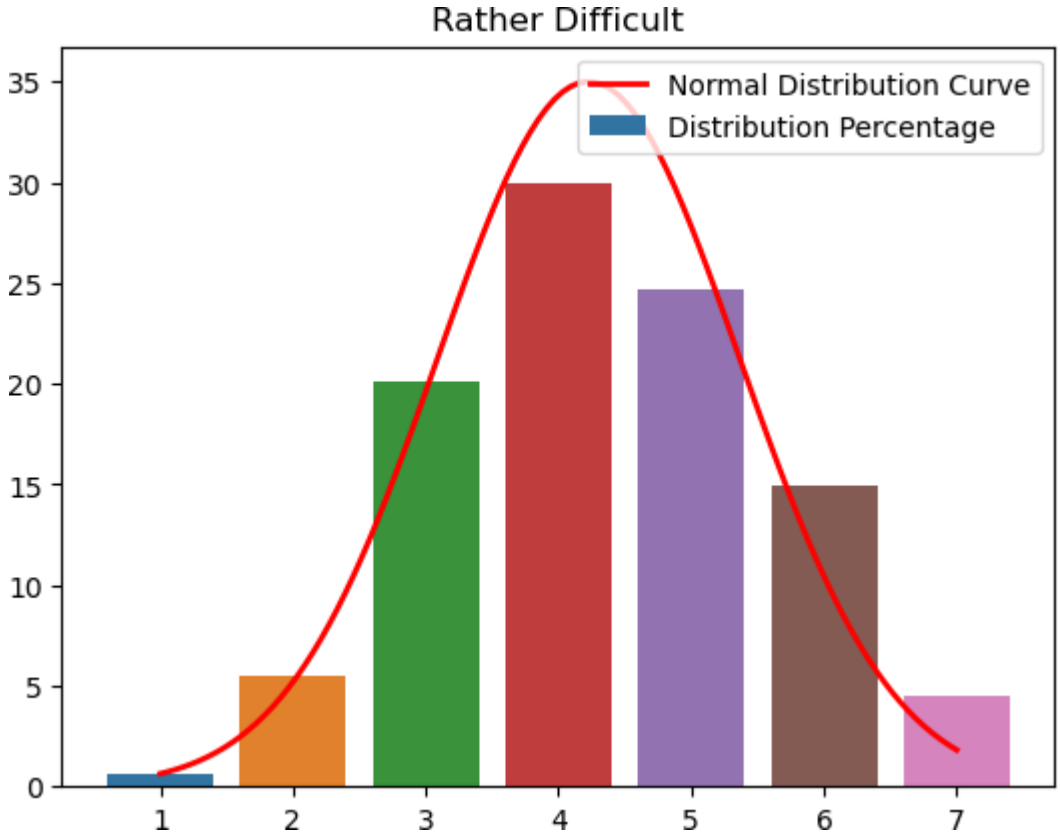
- distribution histograms with normal distribution simple



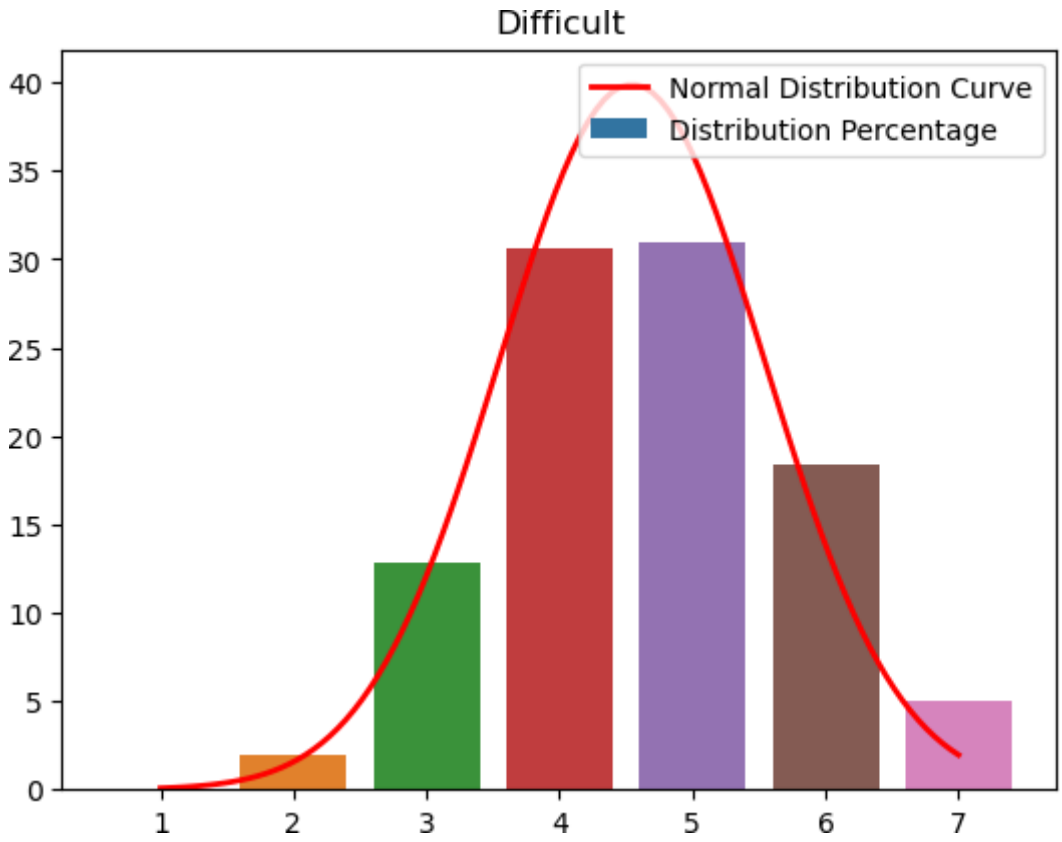
- General



- Harder



- Difficulties

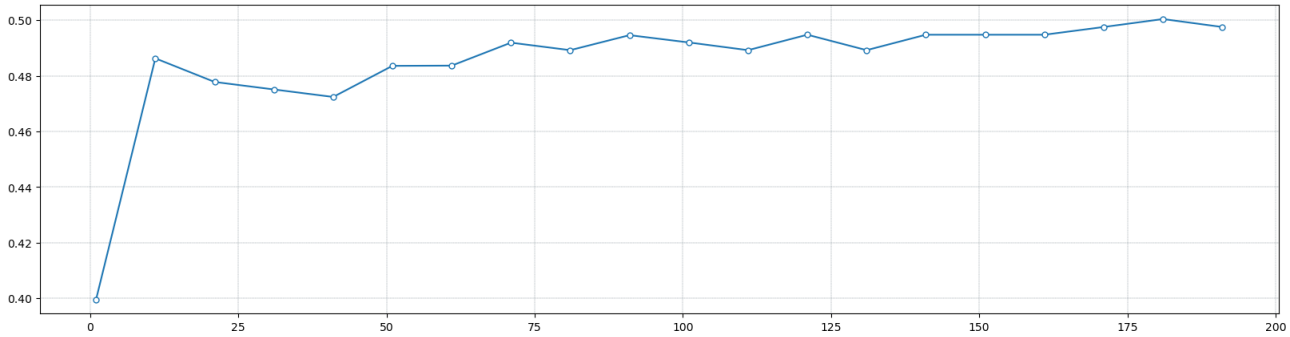


Again using the **random forest** model

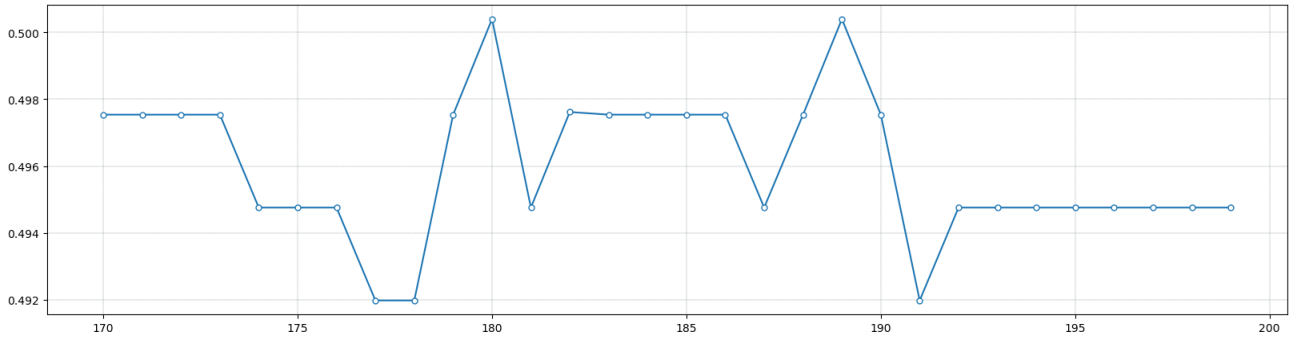
Call Parameters

Determine the optimal parameters of the random forest by grid search, learning curve, etc.

Large Depth Learning Curve

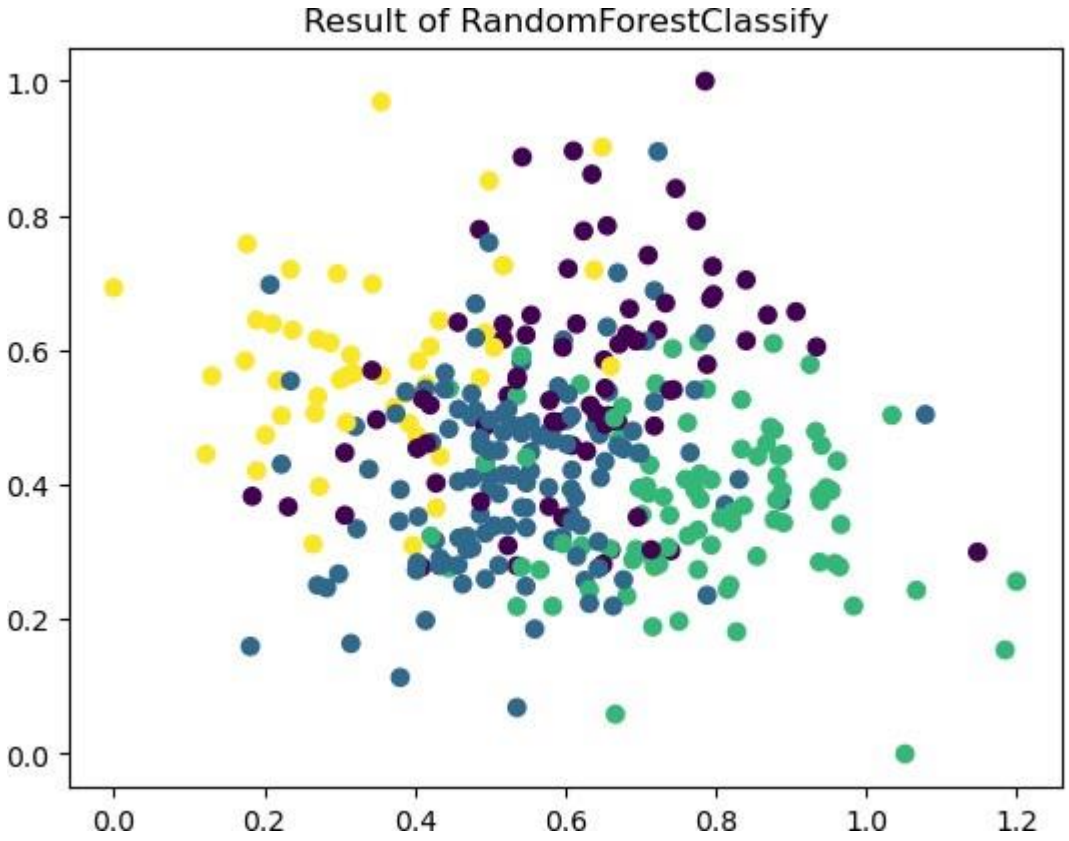


Small learning curve



Parameter	Parameter Value
n_estimators	189
max_depth	10
min_samples_leaf	6
min_samples_split	None

- random forest 를 classification image



Accuracy for the training set: 0.736 Accuracy for the test set: 0.523

- Importance of features

Feature Name	Importance %

entropy	37.54
text_rate	15.58

Feature Name	Importance %
green_score	21.12
yellow_score	14.30
conbine_score	11.44

Predicting Word Difficulty (EERIE)

Predicted classification:

Difficulty **probability of each classification**

Classification	Probability
Simple	9.54
General	14.09
Harder	13.63
Difficulties	62.73