Report for Assignment 1

by XINGLONG LI – 21115109

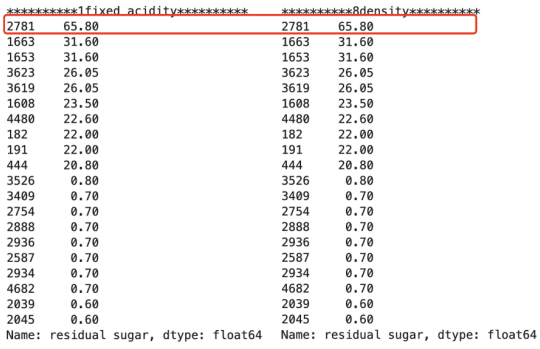
The wine quality dataset has thirteen columns for the features and one (‘color’) for the target and the abalone dataset has eight columns for features and one (‘Rings’) for the target. According to the official description of the datasets, there are no missing value in each column (but we still need to guarantee the point by exploring the real dataset). Our aim is to compare the classification performance (accuracy) of different models, including the KNN Classifier, Decision Trees Classifier, and Random Forest Classifier.

1. Address Abnormal and Null Data

1.1. Wine dataset

As run and shown in jupyter lab, there are no missing values in the dataset, proving its official description. We sort each feature and select the top 10 largest and smallest values for each feature, 20 values in total, and find that the row of data with index 2781 is 'abnormally' larger than others, so we try removing it from the dataset. It is shown in Figure 1 (the row shows abnormal data in each feature, but we only show two features to save space in the report, others are shown in jupyter lab). In addition to the row, we found no other abnormal first 10 maxima and minima.

Interestingly, we find that if we remove the row data that looks 'abnormal', the accuracy in kNN model drops from 0.9508 to 0.9485! Thus, we eventually decide to reserve it.



**********1fixed acidity**********

2781	65.80
1663	31.60
1653	31.60
3623	26.05
3619	26.05
1608	23.50
4480	22.60
182	22.00
191	22.00
444	20.80
3526	0.80
3409	0.70
2754	0.70
2888	0.70
2936	0.70
2587	0.70
2934	0.70
4682	0.70
2039	0.60
2045	0.60

Name: residual sugar, dtype: float64

**********8density**********

2781	65.80
1663	31.60
1653	31.60
3623	26.05
3619	26.05
1608	23.50
4480	22.60
182	22.00
191	22.00
444	20.80
3526	0.80
3409	0.70
2754	0.70
2888	0.70
2936	0.70
2587	0.70
2934	0.70
4682	0.70
2039	0.60
2045	0.60

Name: residual sugar, dtype: float64

Figure 1: Abnormal row data.

1.2. Abalone dataset

Based on the discussion in 1.1, we do the same operation. But we need to use one-hot encoding to convert the categorical feature ‘Sex’ into a numerical format, allowing models to effectively utilize this feature. We also compare accuracy with another way of encoding, which is categorical encoding. Detailed comparisons are shown in part 7.

2. Select Features to Use

2.1. Wine data set

If a feature is false information, for example if we add a column filled with random numbers as a feature column, the performance of our model fitted or trained on data with that feature is likely to be worse than without it, because false leads tend to make the situation worse. Moreover, if a feature is useless information, it may not make the model worse but removing it may make the fitting or training process faster, which is also quite important for models, especially deep learning models based on big data. However, there is often no quantifiable and specific value or criterion for determining which features are harmful or useless, so we must try to determine them using some qualitative data.

We compute the moments or summary statistics on the data (mean, median, variance, skewness, kurtosis) in Jupyter Lab. They are shown in Table 1. The distribution with kurtosis > 3.0 appears as a curve with long tails (outliers) [1]. We can see that the kurtosis of 'chlorides' is 50, well above 3.0, which may adversely affect the performance of the models. In addition, the variance of 'density' is 0.000009, we know that if a column of data has a variance of 0, it means that all the data in that column have the same value and it will not be able to help the classification. Thus, we remove 'Chlorides' and 'Density' and experiment with the default KNN parameters. However, we must be aware that these empirical data processing methods and qualitative analyses of the data do not necessarily bring positive benefits to the model, as mentioned above. (In fact, the result in Table 3 shows that removing 'chlorides' or 'density' reduces the score of the models on the test dataset).

Table 1: Summary statistics of wine dataset.

	Max	Min	Mean	Median	Variance	Skewness	Kurtosis
fixed acidity	15.9000	3.80000	7.215217	7.00000	1.680947	1.723432	5.060898
volatile acidity	1.5800	0.08000	0.339570	0.29000	0.027049	1.493267	2.821732
citric acid	1.6600	0.00000	0.318590	0.31000	0.021108	0.471879	2.401040
residual sugar	31.6000	0.60000	5.433944	3.00000	22.079212	1.169869	0.550836
chlorides	0.6110	0.00900	0.056031	0.04700	0.001227	5.399956	50.895822
free sulfur dioxide	289.0000	1.00000	30.528787	29.00000	315.011565	1.220155	7.908617
total sulfur dioxide	440.0000	6.00000	115.737762	118.00000	3194.910320	-0.000890	-0.371630
density	1.0103	0.98711	0.994690	0.99489	0.000009	0.015011	-0.561341
pH	4.0100	2.72000	3.218474	3.21000	0.025852	0.387217	0.368374
sulphates	2.0000	0.22000	0.531244	0.51000	0.022143	1.797910	8.656972
alcohol	14.9000	8.00000	10.491615	10.30000	1.422556	0.566116	-0.531096
quality	9.0000	3.00000	5.818350	6.00000	0.762687	0.189705	0.231892
color	1.0000	0.00000	0.246151	0.00000	0.185589	1.178858	-0.610481

We also use a pairs plot to look at the whole of the dataset. For convenience, we here only choose a subset, just some features that seem important including 'fixed acidity', 'volatile acidity', 'total sulfur dioxide' and 'pH', and show it in Figure 2. As we can see from the following pairs plot, there are two labels: the blue and orange labels represent '0' and '1' respectively, it is obvious that one label is well distinguished from another in each plot, and

these features could thus be a key distinguishing factor between the two ‘color’ labels. The distributions (KDE plots) highlight noticeable differences in these variables, indicating their usefulness for classification analysis.

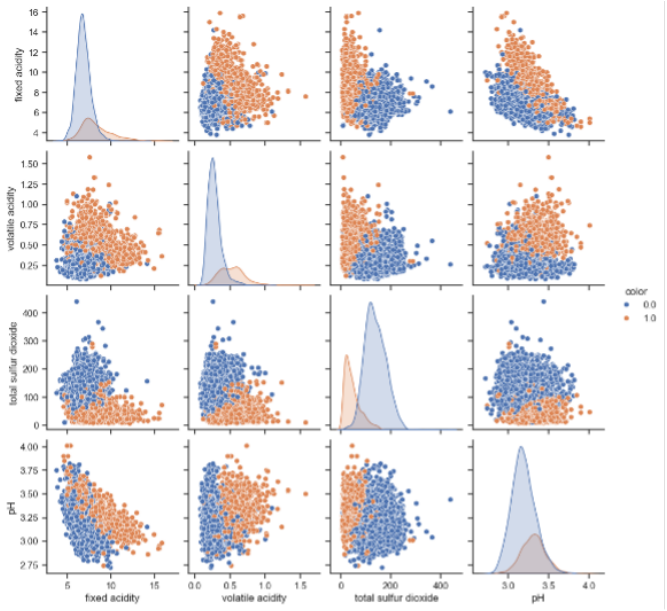


Figure 2. Pairs-plot of wind dataset only using several features

2.2. Abalone data set

To avoid duplication of descriptions, we present here directly the statistics of the abalone dataset (shown in Table 2) and pairs plot (shown in Figure 3). (note: just as discussed above, ‘Sex’ column has been replaced by three new columns: ‘Sex_F’, ‘Sex_M’ and ‘Sex_I’).

Table 2. Summary statistics of abalone dataset.

	Max	Min	Mean	Median	Variance	Skewness	Kurtosis
Length	0.8150	0.0750	0.523992	0.5450	0.014422	-0.639873	0.064621
Diameter	0.6500	0.0550	0.407881	0.4250	0.009849	-0.609198	-0.045476
Height	1.1300	0.0000	0.139516	0.1400	0.001750	3.128817	76.025509
Whole weight	2.8255	0.0020	0.828742	0.7995	0.240481	0.530959	-0.023644
Shucked weight	1.4880	0.0010	0.359367	0.3360	0.049268	0.719098	0.595124
Viscera weight	0.7600	0.0005	0.180594	0.1710	0.012015	0.591852	0.084012
Shell weight	1.0050	0.0015	0.238831	0.2340	0.019377	0.620927	0.531926
Rings	29.0000	1.0000	9.933684	9.0000	10.395266	1.114102	2.330687
Sex_F	1.0000	0.0000	0.312904	0.0000	0.215047	0.807302	-1.348909
Sex_I	1.0000	0.0000	0.321283	0.0000	0.218113	0.765708	-1.414369
Sex_M	1.0000	0.0000	0.365813	0.0000	0.232049	0.557390	-1.690126

From pairs-plot in Figure 3, we find that there exists nearly linear relationship between some features, which means that they are nearly equivalent and can provide model with very similar information. So, we can consider removing one of them, but we don’t do that here because the amount of data is not large, and we decide to try to use any different and useful information.

If they show ‘really’ rather than ‘nearly’ linear relationships, we will remove one of them. In addition, we are aware of the peaking phenomenon in feature selection[2], but it is difficult to determine the desired number of features. Due to the length requirements of the report, we do not conduct relevant experiments but use all features for the dataset.

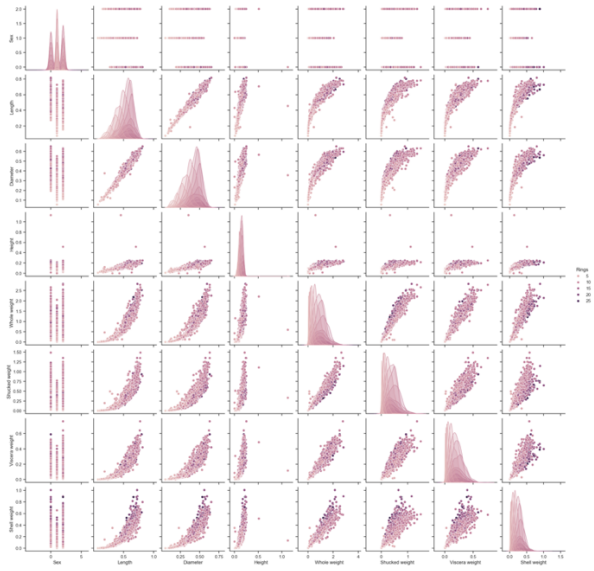


Figure 3. Pairs-plot of abalone dataset

3. Normalize Features

Here we compare two normalization methods[3], including Minmax and Z-score, with the original data without any normalization. For one, some normalizations (perhaps excluding the Minmax normalization) can correct original data with extreme skewness and kurtosis. For another, in kNN model, it allows all features to be compared and calculated on a similar scale, preventing features with too large numbers from dominating the algorithm and features with too small numbers from being ignored.

4. Check & Address Data Imbalance

In wine dataset, if one of the two labels has very little sample data, then the models cannot be fitted or trained well on that label, leading to low performance in predicting the label. In the experiment, the original ratio of positive to negative is close to 1:3, and we compare 1:1 and 1:2 with this by using the SMOTE library, which can copy some positive samples into the original data. (We do not reduce the number of negative samples because it may make our data lose some valuable information).

In contrast, the abalone dataset has over 20 labels, and some of the labels have only one sample, making them difficult, if not impossible, to balance. Thus, we do not balance them.

5. Create Train & Test & Valid Datasets

First, we divide the data into training and test sets in the ratio 80%:20%. In the training set, we then use 5-fold cross-validation to obtain the best parameter.

6. Experimental Result

6.1. Model 1: KNN Classifier

6.1.1. Wine dataset

We search k of the KNN model from 1 to 30 with a step of 1. The result of KNN is shown in Table 3. We can see that data4 where we do z-score normalization on data0 (original data) and KNN 'weights' is 'distance' has the best accuracy on test data and has k = 6.

Table 3. Experiment result of KNN under default parameters

	data0	data1	data2	data3	data4	data5
knn default parameters	best_k = 1; accuracy = 0.9508	best_k = 1; accuracy= 0.95	best_k = 1; accuracy = 0.9508	best_k = 1; accuracy = 0.9931	best_k = 6; accuracy = 0.9923	best_k = 2; accuracy = 0.9877
knn weights = 'distance'	best_k = 1 accuracy = 0.9508	best_k = 1 accuracy = 0.95	best_k = 1 accuracy = 0.9508	best_k = 1 accuracy = 0.9931	best_k = 6 accuracy = 0.9946	best_k = 2 accuracy = 0.9892
Definition of different data	data0: reserve all features. data1: remove ‘chlorides’ from data0. data2: remove ‘density’ from data0. data3: minmax normalization on data0. data4: Z-score normalization on data0. data5: Z-score normalization on data2.					

In data4, we get the best mean validation accuracy and thus only draw its plot of the mean validation accuracy vs. k across all folds is shown in Figure 4.

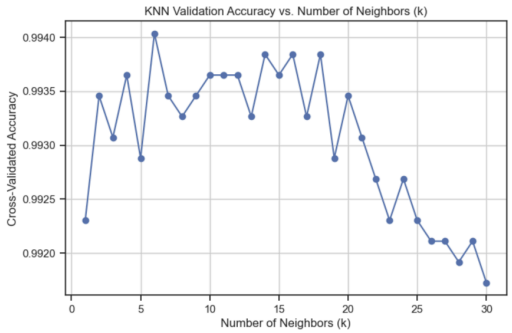


Figure 4. KNN Validation Accuracy vs. Number of Neighbors (k) of data4.

Then, based on the parameters, we experiment with balancing data where 0-label data/1-label data are 1:1 (data6) and 2:1 (data7) respectively, and its result is shown in Table 4.

Table 4. Result after balancing data

knn parameters	data4 (0/1 = 3:1)	data6 (0/1 = 1:1)	data7 (0/1 = 2:1)
weights = 'distance'	best_k = 6; accuracy = 0.9946	k = 6; accuracy = 0.9931	k = 6; accuracy =0.9931

6.1.2. Abalone dataset

In default knn parameters, best_k and the plot of 'validation accuracy vs k' are shown in the figure below.

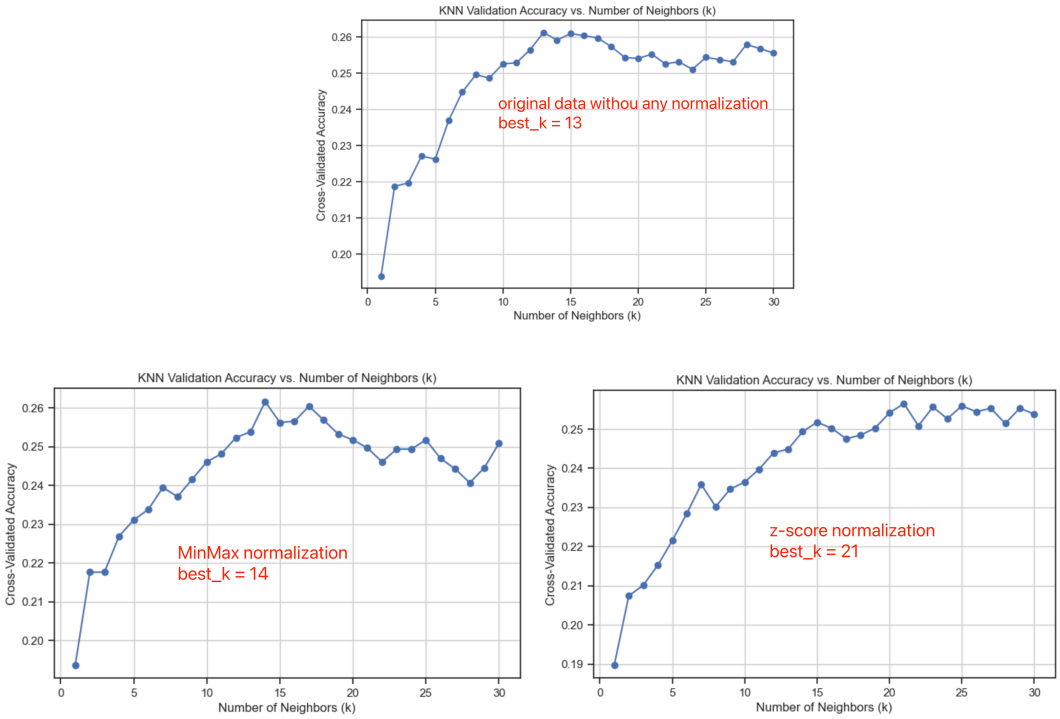


Figure 5. KNN Validation Accuracy vs. Number of Neighbors (k) of Abalone Dataset.

In default parameters and weights='distance', we compare different experiments, and their results (note: accuracy on test dataset here not on validation dataset in above Figure 5) are listed in the table below.

Table 5. KNN result of abalone test dataset

	Original data	MinMax normalization	Z-score normalization
Default settings	0.2512	0.2416	0.2751
weights='distance', other settings are default	0.262	0.2428	0.2715

6.2 Model 2: Decision Trees Classifier

6.2.1 Wine dataset

We search 'max_depth' of the Decision Trees Classifier from 1 to 20 with a step of 1. By GridSearchCV function, we get the result that best parameter is {'max_depth': 11} and best accuracy is 0.9840 and produce a plot showing the mean accuracy vs. relative to tree depth in Figure 6.

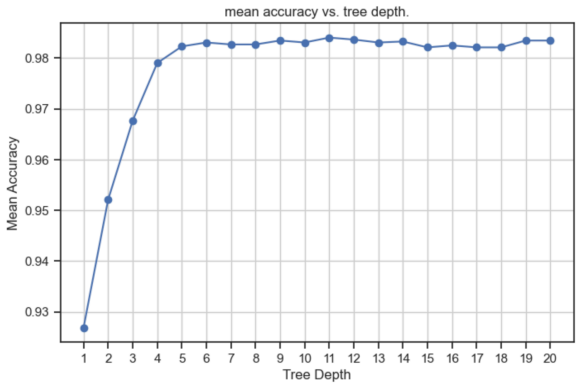


Figure 6. DT validation accuracy vs. tree depth of wine dataset.

We visualize the final decision tree, part of which is shown in Figure 7. Then, we calculate the final resulting splitting rules (features) used for the tree, sort them by times used in the rules and get the list [('total sulfur dioxide', 16), ('pH', 16), ('sulphates', 16), ('chlorides', 14), ('density', 14), ('residual sugar', 14), ('fixed acidity', 12), ('citric acid', 10), ('volatile acidity', 10)]. In ‘Select Features to Use’ part, on a pairs plot, we observe four features that look important including 'fixed acidity’, ‘volatile acidity’, ‘total sulfur dioxide' and 'pH', which are exactly included in the list.

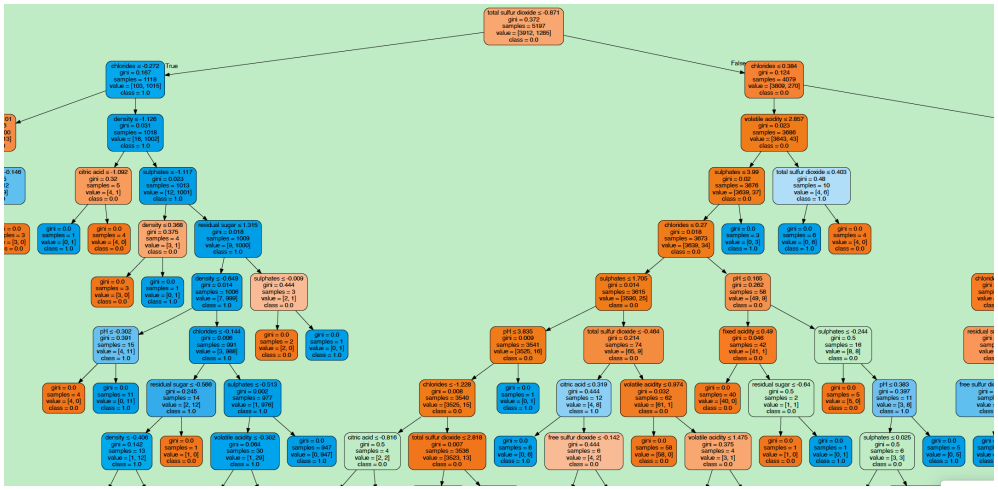


Figure 7. Part of decision tree of wine dataset.

The original raw features have clear meanings, but if we use PCA/LDA, we will not be able to explain the meaning of the new features generated by PCA/LDA and then used by the decision tree for each split, which will affect the interpretability of the decision tree model. In

addition, compared to the original raw features, PCA/LDA will lose some information, which may reduce the classification ability of the decision tree.

6.2.2 Abalone dataset

The best param is {'max_depth': 5} and result plot is shown below. We also get similar comparison between decision tree rules and the analysis about features in pairs-plot to wine dataset.

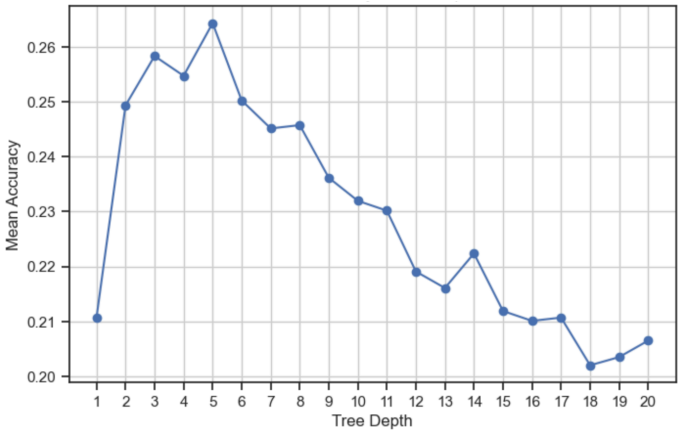


Figure 8. DT validation accuracy vs. tree depth of abalone dataset.

6.3 Model 3: Random Forest Classifier

6.3.1 Wine dataset

We set two ranges of parameters: 'max_depth': [2, 3, 5, 10, None], 'n_estimators': [10, 50, 100, 200] and produce a plot showing the mean accuracy vs. the parameter settings, shown in Figure. Run the code and then get the best parameter = {'max_depth': None, 'n_estimators': 100}, where None means that decision trees have an upper limit on how deep they can go determine by the size of the dataset.

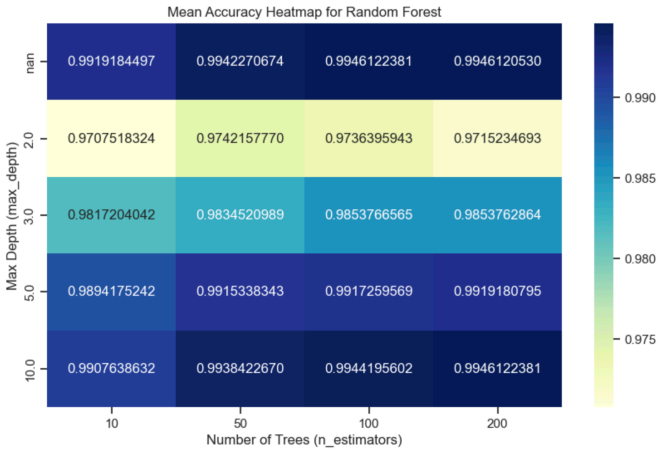


Figure 9. Heat map of random forest on wine dataset.

6.3.2 Abalone dataset

The best params are {'max_depth': 2, 'n_estimators': 50}.

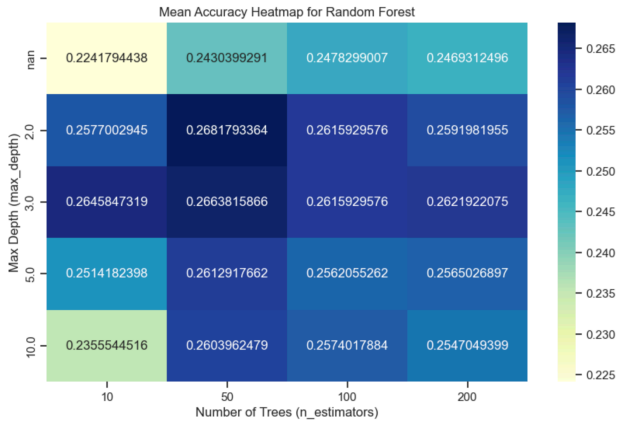


Figure 10. Heat map of random forest on abalone dataset.

7 One Hot Encoding vs Category Encoding

As we mentioned in 1.2, in the abalone dataset we encode the 'Sex' feature with two encoding methods: categorical vs. one-hot. The former is to encode 'M/F/I' with '0/1/2' and then delete the original 'Sex' column; the latter is to encode 'M/F/I' with '001/100/010', add three columns called 'Sex_M', 'Sex_F' and 'Sex_I', and then delete the original 'Sex' column, we can see the new data after one-hot encoding in Table 6.

Table 6. Statistic of abalone dataset

Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings	Sex_F	Sex_I	Sex_M
0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15	0	0	1
0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7	0	0	1
0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9	1	0	0
0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10	0	0	1
0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7	0	1	0

In the categorical encoding experiments, we get the different accuracies on test dataset, shown on Table 7. Now we use one hot encoding based on z-score & random forest, we get accuracy = 0.2990 on test dataset, which is larger than 0.2955 on categorical encoding & z-score & random forest.

	knn	decision tree	randomforest
original	0.2512	0.2632	0.2943
Minmax	0.2416	0.2632	0.2955
z-score	0.2751	0.2601	0.2955

Table 7. Accuracy of test dataset on different normalizations by three models

8 Summary Results of Best Models

We use best models to predict the test data and get accuracies, listed below.

	Wine - setting	wine	Abalone -setting	abalone
kNN	best setting: k=6, knn weights = 'distance', other knn parameters default, Z-score normalization	0.9946	best setting: k=21, knn weights = 'distance', other knn parameters default, Z-score normalization, one-hot encoding	0.2751
Decision Tree (DT)	best setting: DT max_depth=11, other DT parameters default, Z-score normalization	0.9846	best setting: DT max_depth=5, other DT parameters default, Z-score normalization, one-hot encoding	0.2823
Random Forest (RF)	best setting: RF max_depth=None, RF n_estimators=100, Other RF parameters default, Z-score normalization	0.9969	best setting: RF max_depth=5, RF n_estimators=10, Other RF parameters default, Z-score normalization, one-hot encoding	0.2990

There are some interesting points from the experiments:

a. The accuracies on the abalone dataset, 0.2+, are much lower on the wine dataset, which is consistent with what we see in their pairsplots: the wine dataset has many features with 'nearly' linear relationships (we discussed in detail in 2.2 why 'linear relationship' reduces useful features for models); the number of features is also lower than the wine dataset.
b. Categorical encoding for the 'Sex' column in the abalone dataset yields lower accuracy than one-hot encoding, likely due to false numerical relationships (e.g., $0 < 1 < 2$) introduced by categorical encoding. One-hot encoding avoids this issue, ensuring independence among categories within the 'Sex' column as their vector products are zero.
c. The performance of MinMax and Z-score normalization varies: with the KNN model's default parameters on the wine dataset, MinMax achieves higher accuracy (0.9931 vs. 0.9923), but with weights = distance, Z-score outperforms MinMax.

References:

[1]. https://www.investopedia.com/terms/k/kurtosis.asp

[2]. https://www.sciencedirect.com/science/article/abs/pii/S0167865508001426

[3]. https://www.codecademy.com/article/normalization