ECE 657A

Assignment 1 Report

Munia Humaira & Mishaal Khan

February 7, 2025

#### 1. Assessment of Data and Applying Normalization

#### 1.1 Wine Dataset

The wine dataset has 2 csv files which have been loaded as pandas dataframe and concatenated together for convenience. It was checked for missing values. The analysis showed that the dataset didn’t contain any missing data, which allowed us to proceed without imputation.

The statistical summarization (mean, median, variance, skewness, and kurtosis) of all numerical features were computed.

* The "residual sugar" feature exhibits a particularly high variance (22.63), indicating a wide spread in values.
* The "volatile acidity" (1.49) and "chlorides" (5.39) features show substantial positive skewness.
* "Density" has a low variance (0.000009), suggesting that most samples have very similar density values.
* "Total sulfur dioxide" has a high variance (3194.72), indicating a broad range of sulfur dioxide levels.
* "Chlorides" exhibit very high kurtosis (50.89), meaning that the distribution is heavily tailed with extreme values.

A pairplot was generated to visualize relationships among features. Some attributes showed strong correlations. It reveals a strong positive correlation between alcohol content and wine quality, suggesting higher alcohol levels often indicate higher quality. Total sulfur dioxide shows a potential, weaker negative correlation with quality. Free sulfur dioxide and residual sugar show little to no clear relationship with quality. Alcohol and total sulfur dioxide are likely important features for further quality analysis. Free and total sulfur dioxide are highly correlated.

The wine dataset has a color imbalance. The average color value is 0.246 (standard deviation 0.43), indicating about 24.6% of the wines are red and 75.4% are white. To ensure better model performance, we can apply class balancing techniques, such as :

* Undersampling (Reduce Majority Class) - Randomly remove some samples from the majority class (white wines) to balance with the minority class (red wines).
* Oversampling (Increase Minority Class) - Randomly duplicate or synthetically generate more samples from the minority class (red wines).

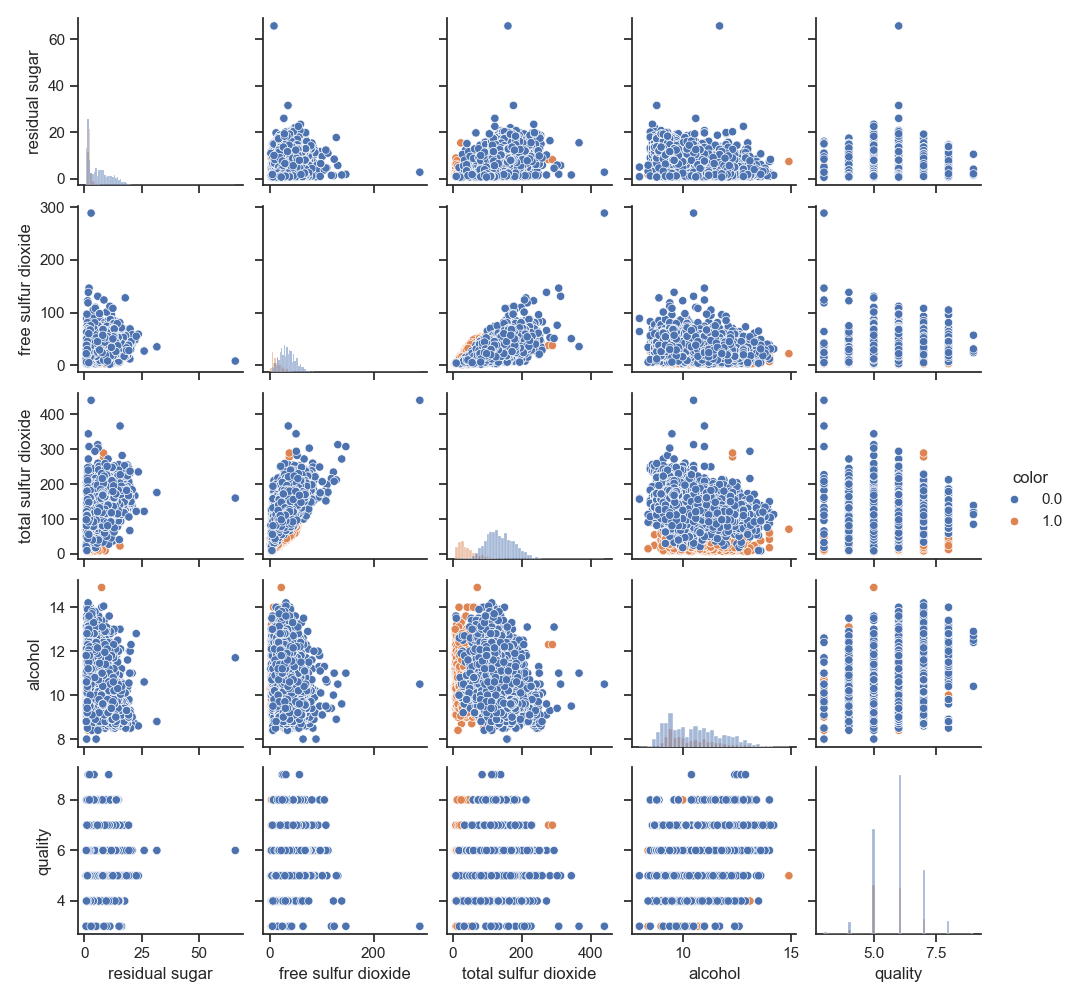
In this approach the undersampling method has been chosen.

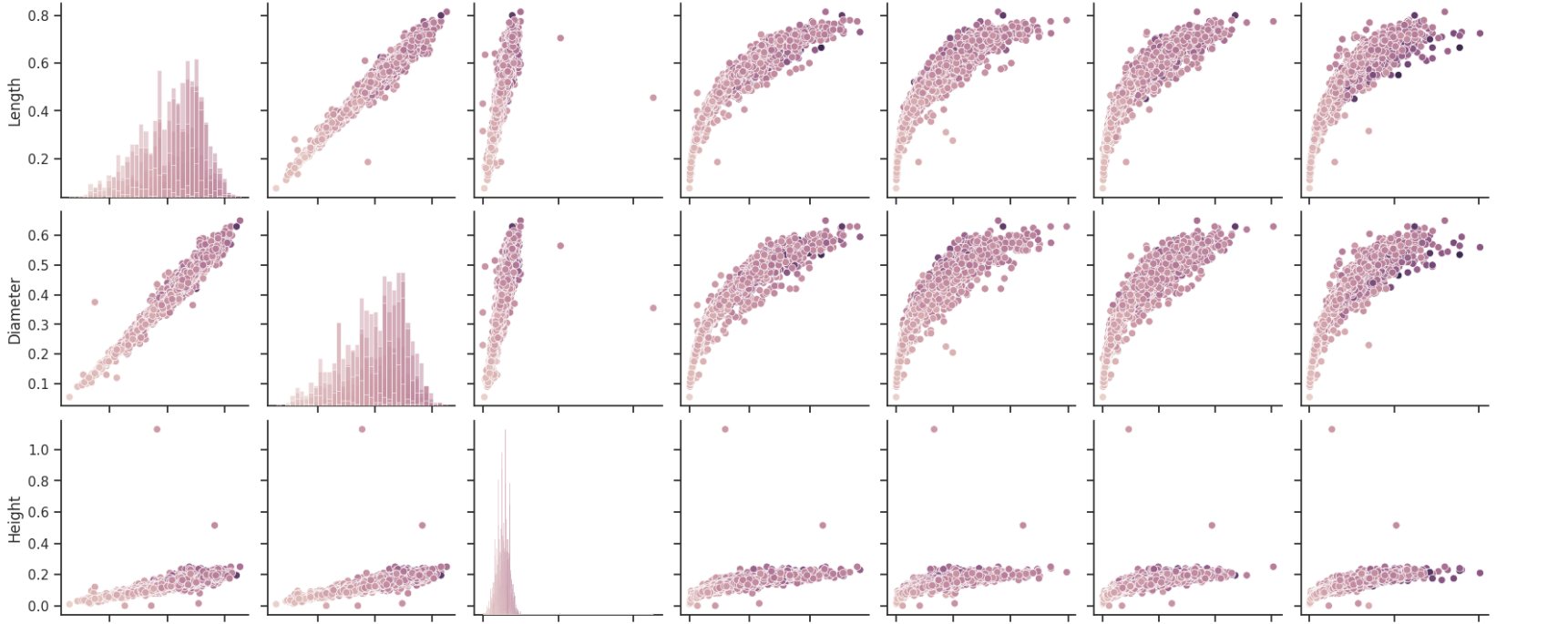
Figure 1. Pairplot of the features

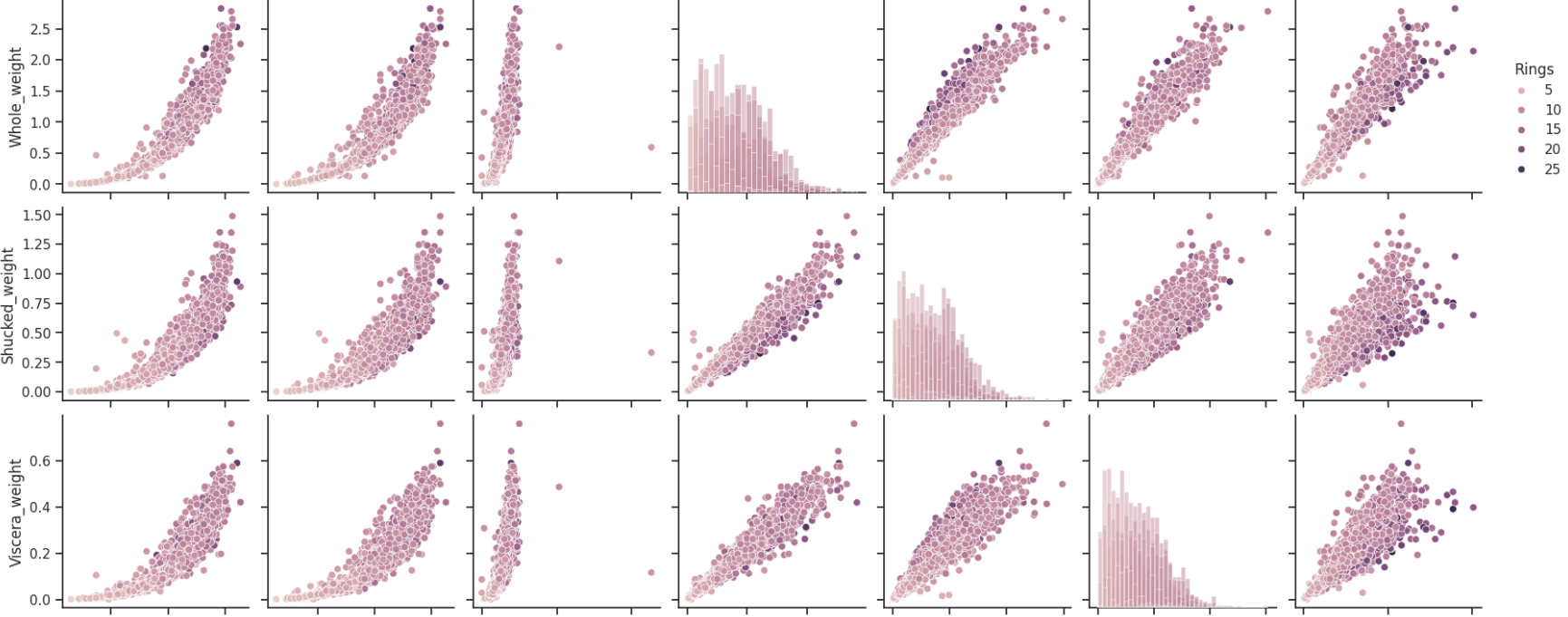
##### 1.2 Abalone Dataset

The abalone dataset has 4177 samples and 8 features. There was no missing data. The statistics on the data features were generated using Pandas’ describe function for dataframes. Additional statistics of variance, median, skew and kurtosis were observed using their respective Pandas functions. Some observations include:

* The “Height” feature has a significantly higher kurtosis than the rest of the features, at around 76.03. This means the tail of the distribution is heavy. It also has the lowest variance among the features at 0.001750, indicating height values are not spread out as much.
* “Length” and “Diameter” are slightly negatively skewed (with values of around -0.64 and -0.61 respectively), meaning that most of the length and diameter values tend to be concentrated around the higher end, while the rest of the features are positively skewed. “Height” is the most positively skewed, with a value of around 3.13.
* “Whole weight” has the highest variance, at around 0.24, indicating the most spread out values.

Figure 2 shows the pairplot for the abalone features. All the features seem to be positively correlated with one another, to varying degrees. For example, length and diameter are strongly positively correlated, implying that longer abalone have larger diameter as well, with a potentially linear trend. Furthermore, by observing the colour of the circles, the abalone with a lower amount of rings tend to have smaller lengths and diameters, whereas abalone with a larger amount of rings tend to have larger lengths and diameters.





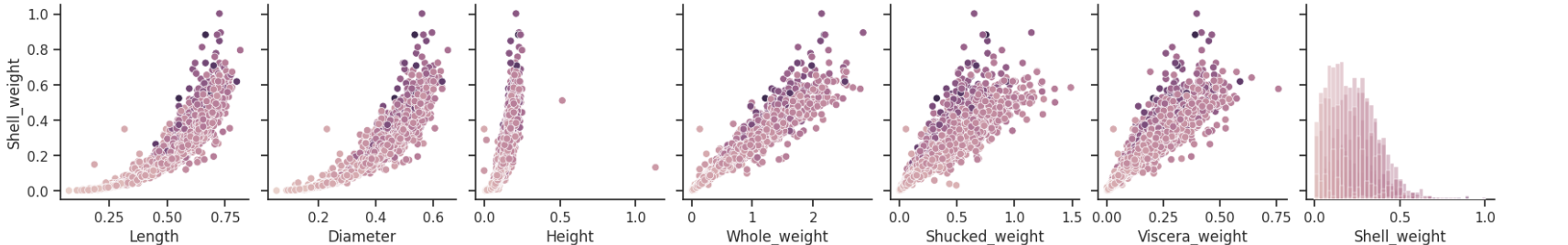


Figure 2. Pairplot for numerical abalone features

The categorical column of “Sex” was one-hot encoded using Pandas get\_dummies function as a preprocessing step so that the “Sex” feature column was replaced with three columns which encode the value.

This is not a balanced dataset because the number of samples for each class varies greatly. For example, the class with the most amount of samples is the abalone with 9 rings class, comprising 689 samples, whereas there is only 1 sample available for classes 1, 2, 25, 26, and 29 rings. Because the smallest class frequency is 1, it would not make sense to undersample and reduce the data to 1 sample per class. Instead, as mentioned in the abalone.names file included in the dataset files, in the paper "A Quantitative Comparison of Dystal and Backpropagation", classes are grouped as follows: classes 1 to 8 together, classes 9 and 10 together, and classes 11 to 29 together [1][2]. This will be the approach taken here, as this provides a better balance of samples per class. The new classes will be referred to as class 0, class 1 and class 2. The results of grouping classes (the correction) will be compared to the results from using the dataset with the original classes as-is. After applying this correction, there are 1407 samples in class 0, 1323 samples in class 1, and 1447 samples in class 2, which provides a better balance.

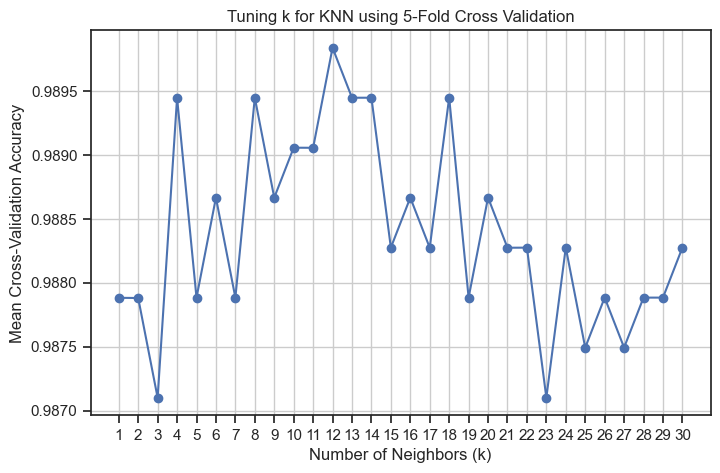
Then, standardization is applied using Scikit-learn’s StandardScaler. Categorical variables undergo one hot encoding but normalization or standardization is only applied to continuous features [3]. So the scaler is fit and applied to the numerical training data and then applied to the numerical test data.

#### 2. Classification with KNN

##### 2.1 Wine Dataset

The default parameters of the model have been used to train and test the data. An accuracy of 99% is seen.

A range of values has been chosen for the parameter to perform the 5-fold cross validation on the training data. The best value obtained from the process is 12.

Figure 3. Cross-validation to find out the best value

The best value of has been used for the training data and the accuracy calculated on the held out test set is obtained 99.38%.

To improve the result weighted KNN is used, however, no improvement from the previous accuracy has been seen.

##### 2.2 Abalone Dataset

First, Scikit-learn’s KNeighborsClassifier was used with its default parameters, which includes using as the default number of nearest neighbors. Using the unbalanced dataset, the test accuracy was around 20.69%. On the other hand, using the dataset with grouped classes results in a test accuracy of around 61.12%. The test accuracy when using the more balanced version of the abalone dataset (by grouping classes) is higher than the abalone dataset with the original classes. This is likely because after balancing there are enough samples of each class for the model to learn sufficiently. Therefore, for the remainder of the experiments, the results will be reported for the balanced dataset.

Using 5-fold cross validation on the training data, the mean validation accuracy was obtained for values ranging from 1 to 30. Figure 4 shows the mean validation accuracy versus .

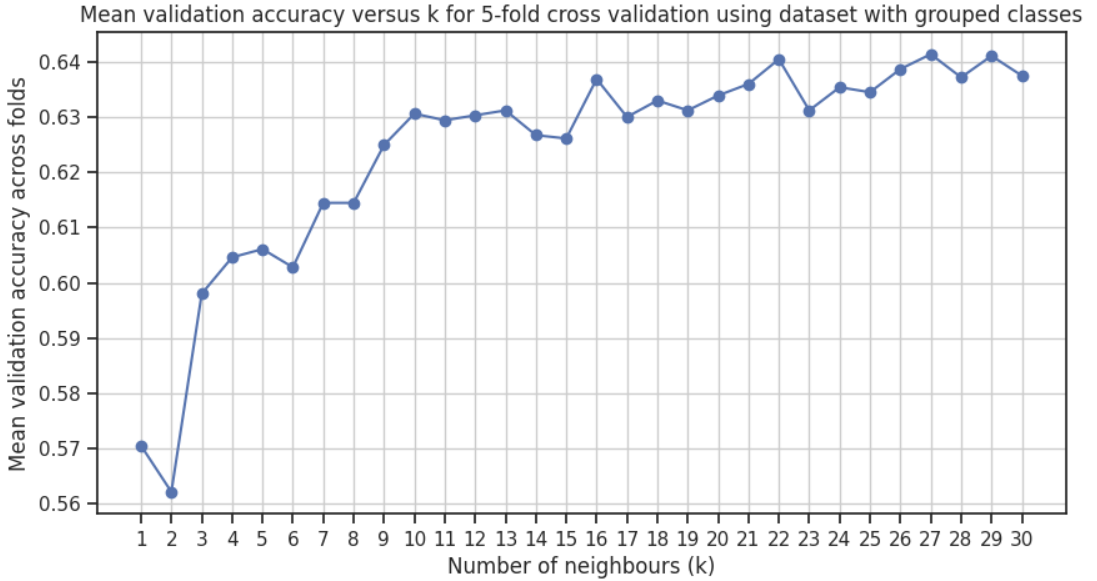


Figure 4. Mean validation accuracy versus number of nearest neighbours

The best setting found was as this resulted in the highest mean validation accuracy across the folds. Using this value of , the accuracy on the test set was around 65.43%, which is higher than the accuracy obtained with the default kNN. Using the weighted kNN with the distance metric, the accuracy increases slightly to around 66.03%.

#### 3. Decision Trees Classifier

##### 3.1 Wine Dataset

A range of maximum depth parameters for the tree has been used to extract the best max\_depth parameter for the Decision Tree classifier. The best depth has been obtained to be 11.

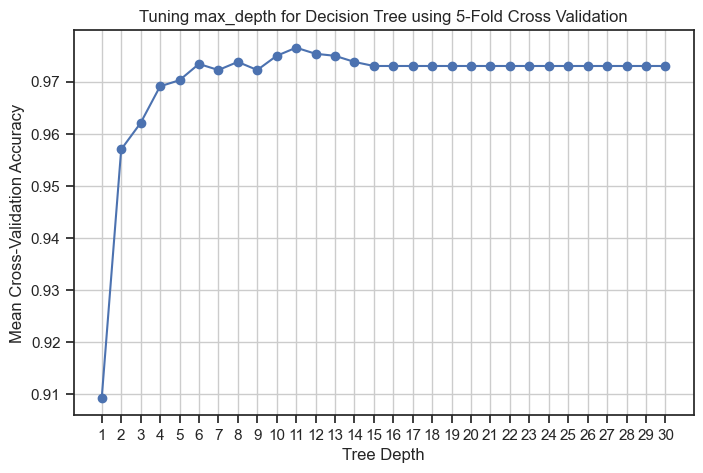


Figure 5. Mean accuracy vs relative tree depth

The classifier has been used to predict the color of the wine, an accuracy of 91.38% has been obtained. The tree has been visualized in the notebook, however it is too large to fit in the report.

The tree-splitting rules show that chlorides, sulfur dioxide, citric acid, and alcohol have been used to classify samples 0 and 1. The root decision starts with chlorides ≤ 0.06, meaning the dataset is first split based on this threshold. If chlorides ≤ 0.06, further splits occur based on total sulfur dioxide, free sulfur dioxide, and sulphates. If chlorides > 0.06, other factors like density, residual sugar, and alcohol become key decision points. Chlorides, sulfur dioxide (both total and free), citric acid, volatile acidity, alcohol, sulphates, pH, and residual sugar frequently appear in the tree, suggesting these features have a strong influence on the classification outcome. The tree has many levels, indicating a deep tree that may have a risk of overfitting (if not pruned or regularized). This depth means the model is capturing fine details but may lack generalization for new data.

##### 3.2 Abalone Dataset

The decision tree maximum depth parameter was varied from 2 to 20 levels, and the results were observed using Scikit-learn’s GridSearchCV with 5 folds. Figure 6 shows the mean validation accuracy versus the tree depth parameter used.

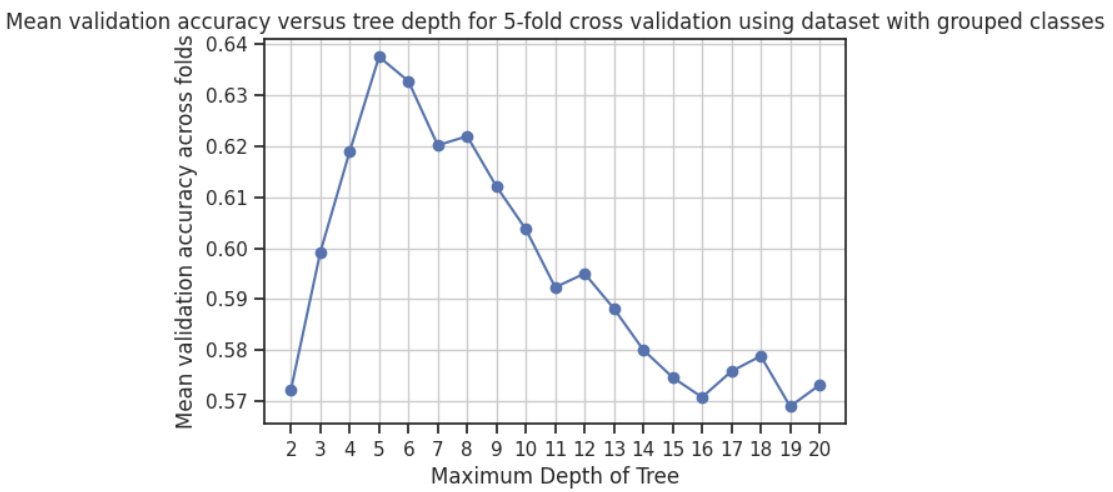


Figure 6. Mean validation accuracy versus tree depth

The best tree depth found was 5 levels, as this corresponds with the highest mean validation accuracy across the folds. With this parameter setting, the test accuracy was around 62.56%.

Using Scikit-learn’s plot\_tree function, the decision tree has been visualized in the notebook. The first splitting rule is based on the shell weight. The right side of this subtree (the side with higher shell weight) does not hold any leaves that would classify a sample as class 0. This aligns with the intuition that a higher shell weight might indicate an older specimen, with more rings. After the first splitting rule, on the left subtree, the features of “Shucked weight” and “Viscera weight” are most commonly used for splitting. On the right subtree, the features of “Shell weight” and “Shucked weight” are most commonly used for splitting. Near the top of the tree “Shell weight” is commonly used as a split decider, indicating it may play an important role in discerning between classes at a high level.

#### 4. Random Forest Classifier

##### 4.1 Wine Dataset

A Random Forest Classifier was trained using GridSearchCV with 5-fold cross-validation. The tuned hyperparameters were: Number of trees (3, 10, 50, 100, 200) and Maximum depth (2, 3, 5, 10, 20). Mean accuracy vs the above parameter settings have been plotted. From the heatmap we can see the ideal **max\_depth** = 10 and **n\_estimators** = 50. The classification with the best parameters gives an accuracy of 99.46%.

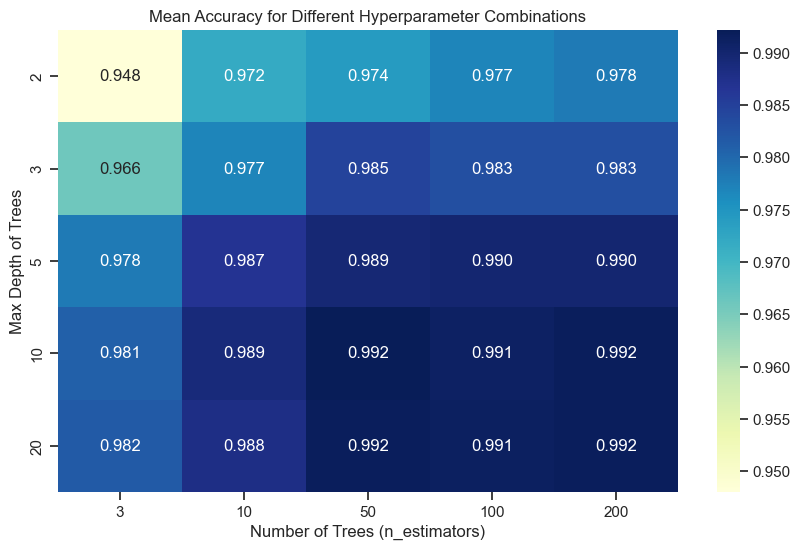


Figure 7. Hyperparameter selection

##### 4.2 Abalone Dataset

Grid search on the number of trees and depth of trees parameters was conducted. The number of trees parameter was varied from 3 to 100 in increments of 20, and the depth of trees parameter was varied from 3 to 10. Figure 8 shows the heatplot showing the mean validation accuracy for the different parameter combinations. From the heatplot it can be seen that using less trees yields lower accuracies. Using 63 trees and a tree depth of 7 levels yields the best results. The test accuracy using these parameters was around 66.99%.

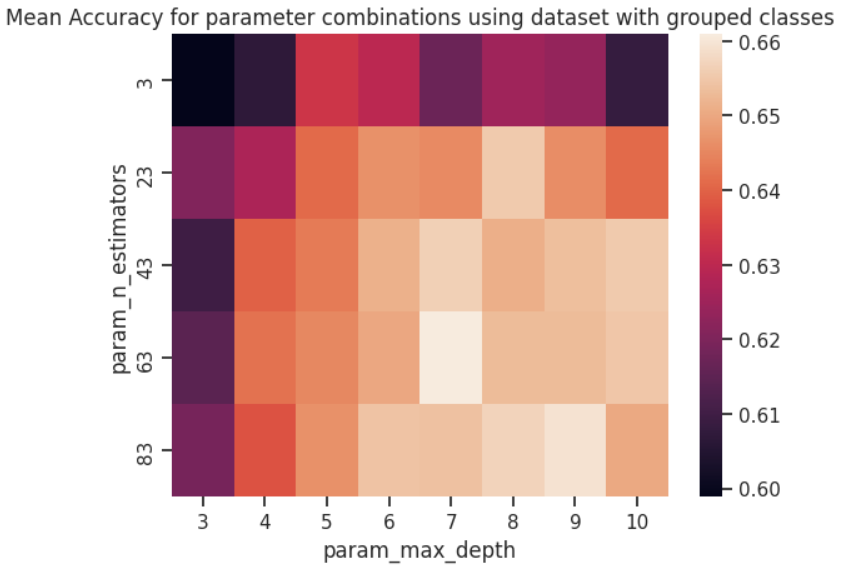


Figure 8. Heatplot for mean validation accuracy versus parameter combinations

#### 5. Results

Table 1 shows the best model settings and corresponding test accuracies rounded to 2 decimal places for both datasets.

##### 5.1 Wine Dataset

KNN has performed well suggesting that the dataset has clear clusters that kNN can classify accurately with the best value for = 12.

Decision Trees Classifier gives significantly lower accuracy than kNN even with the best setting found via the cross-validation, indicating that a single decision tree may not be capturing all the necessary patterns.

The random forest model performs slightly better than KNN and significantly better than the decision tree classifier. The improvement over a single decision tree suggests that the ensemble methods help to reduce overfitting.

##### 5.2 Abalone Dataset

In terms of comparative performance between the models, similar patterns can be observed with the abalone dataset results, with the random forest model performing the best while the decision tree performs the worst.

In terms of next steps, one can avoid data leakage by implementing a pipeline where preprocessing is done in every fold during cross validation, instead of before splitting the training and validation set, [4] and see what impact this has on the results.

Table 1. Test Accuracies for Wine and Abalone Datasets for various models

| Model | Best setting for wine dataset | Test accuracy for wine dataset | Best setting for abalone dataset | Test accuracy for abalone dataset |
| --- | --- | --- | --- | --- |
| kNN |  | 99.38% |  | 66.03% |
| Decision Tree | max\_depth=11 | 91.38% | max\_depth = 5 | 62.56% |
| Random Forest | n\_estimators=50, max\_depth=10 | 99.46% | n\_estimators = 63, max\_depth = 7 | 66.99% |

### 

**References**

[1] W. Nash, T. Sellers, S. Talbot, A. Cawthorn, and W. Ford. "Abalone," UCI Machine Learning Repository, 1994. [Online]. Available: https://doi.org/10.24432/C55C7W.

[2] D. Clark, Z. Schreter, and A. Adams, "A Quantitative Comparison of Dystal and Backpropagation," in *Australian Conference on Neural Networks, 1996*.

[3] B. Soni, “Topic 1: Standardization and Normalization,” Medium, https://medium.com/@brijesh\_soni/feature-engineering-101-7cb68d293551 (accessed Feb. 7, 2025).

[4] Silva.f.francis, “Avoiding Data Leakage in Cross-Validation,” Medium, https://medium.com/@silva.f.francis/avoiding-data-leakage-in-cross-validation-ba344d4d55c0#:~:text=A%20Proper%20Cross%2DValidation%20Pipeline,influenced%20by%20the%20validation%20data. (accessed Feb. 7, 2025).