# Analysis of Classification Models on the Apple Quality Dataset

## CSDS 340: Introduction to Machine Learning

**Case Study 1**

Oliver Yuan xxy485 Jack Qian cxq72

## 1.0.Functions and parameters

### *1.1. Logistic Regression*

Logistic regression is a statistical model that uses logical functions to model binary dependent variables for classification tasks.

In our study, the **parameter C**, which is the inverse of regularization intensity. It controls the penalty intensity and prevents overfitting. A lower "C" value will increase the regularization intensity, creating a simple model and increasing bias. On the contrary, a higher "C" value will reduce the regularization intensity, thereby increasing the complexity and potential variance of the model. In our design, logarithmic exploration was conducted on the variation of "C" from 0.01 to 100 to determine its impact on the model.

### *1.2. Support Vector Machine*

SVM is a classification technique that works by finding the hyperplane that best separates the classes in the feature space.

1.2.1 The **parameter C:** is the same as Logistic regression to apply regularization to controls the penalty intensity and prevents overfitting

1.2.2. **Gamma**: This parameter defines the impact of a single training example. A low value of "gamma" means that the "far" point contributes to the decision boundary, while a high value means the opposite. This hyperparameter adjustment focuses on identifying the optimal point, where the model can be well generalized without overfitting.

### *1.3. Decision Trees*

Decision Tree is the model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

1.3.1 **Max Depth:**The maximum depth of a tree can limit the complexity of the learning model and reduce the risk of overfitting. Considered various depths from unrestricted to a maximum depth of 50.

1.3.2 **Min Samples Split**: This parameter controls the minimum number of samples required to split internal nodes in the tree. Higher values can reduce overfitting, but they can also prevent the model from learning complex patterns

## 2.0.Data preprocessing algorithms

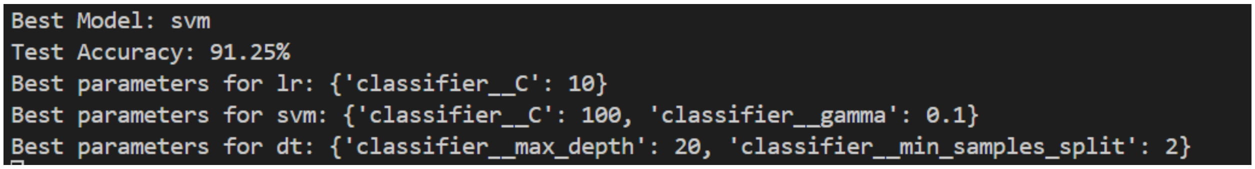
In our experiment, we applied standardScaler from scikit-learn to standardize the dataset's features by removing the mean and scaling proportionally to unit variance. This standardization ensures that each feature contributes equally to the distance calculation in the model. We also try to see the difference between using it and not. By testing the standardScaler performs well in SVM. In addition, normalization is performed to rescale the features to the range of [0,1] or [-1,1].

## 3.0. Hyperparameter tuning process

### 3.1 Basic compare

We try both random search and grid search for hyperparameter tuning. The GridSearchCV performs better. By applying it, provide a detailed search in the specified parameter space. Also we use its cross-validation method not only to robust parameter adjustment but also reduces the risk of overfitting. We select 10 fold as a parameter for cross-validation that divides the data into ten groups and performs ten validations on the model. Each using a different retention set. This process ensures that the model is tested on the entire dataset to evaluate its performance.

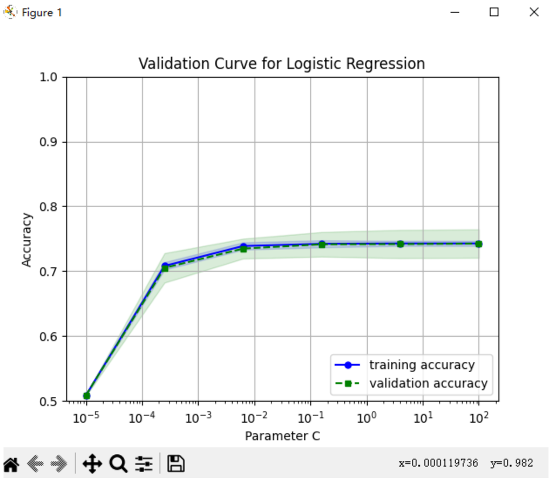
## 4.0. Final choice of classification model



```
Best Model: svm
Test Accuracy: 91.25%
Best parameters for lr: {'classifier__C': 10}
Best parameters for svm: {'classifier__C': 100, 'classifier__gamma': 0.1}
Best parameters for dt: {'classifier__max_depth': 20, 'classifier__min_samples_split': 2}
```

figure1

We first try to run all three models without normalization or standardization for data preprocessing. The SVM model shows higher accuracy over logistic regression and decision tree



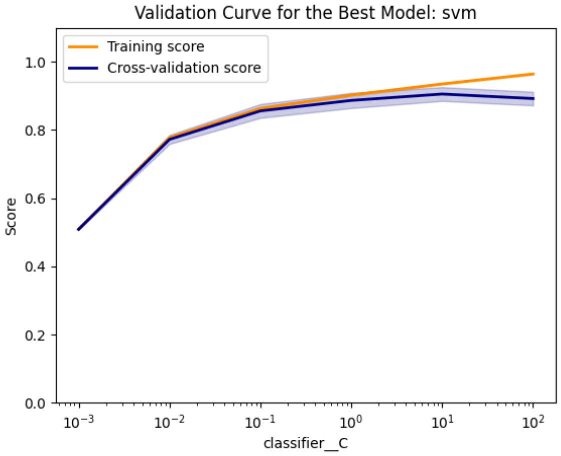


Figure 2

Figure3

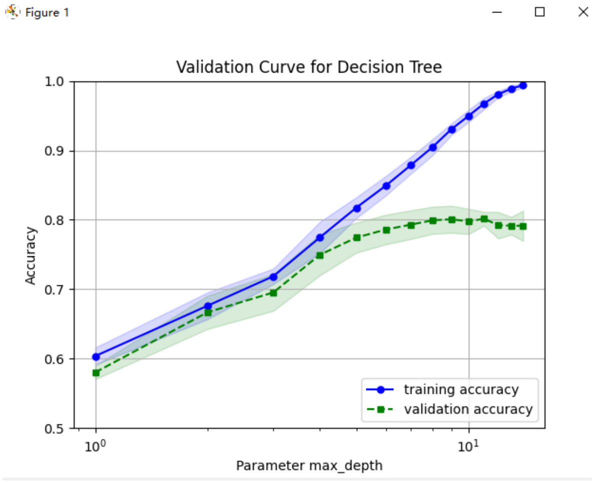


Figure4

3.2 Data preprocessing and Hyperparmeter tuning

Then we apply standardization and observe that the accuracy improved in all three models as shown in figure 2-4. Though Figure4 shows very high accuracy when increasing regulation, it is meaningless to have an overfitting problem. If we focus on the point that training and validation curves are close and high, Figure 3 performs much better than Figure 2 and Figure4. Suggest that SVM is the best Also, we found the SVM shows 10-15% higher accuracy score than Logistic Regression and decision tree. The final selection of hyperparameters, with "C" set to 100 and "gamma" set to 0.1, is the result of a careful calibration process that balances model complexity and the ability to generalize to invisible data. The radial basis function kernel of this model is adept at handling subtle differences in datasets.

5.0. Observations and discussions.

We also try to generate the histogram of raw data, such as roughness, weight, size, maturity, and juiciness, revealing different distribution ranges as shown in figure set 1 below. Most features perform normal distributions which indicate all the features are important for training. We also clear the original dataset but there is no change indicating the raw dataset is goodenough. The validation curve of Figure3 indicates that our final model SVM has been well adjusted. The training and cross-validation scores converge with the increase of the "C" parameter value, indicating that the model can be well generalized without overfitting the training data.

Also we find that Logistic Regression has low accuracy, but if the training and validating curve are convergent at the end, suggest it has good training that avoids overfitting or high bias. For decision three models, the accuracy is nearly 100% when increasing max_depth. However it means each subtree in the model only contains few data which is overfitted and meaningless.

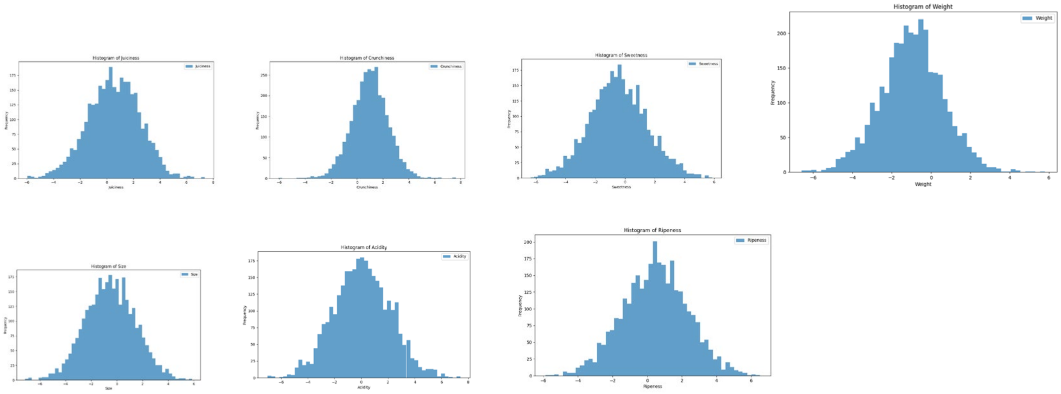


Figure Set 1:Raw Data Distribution