# STA 160 - Midterm Project

---

Seed Report

**Yutian Yang (915327218)**

# I. Introduction:

## 1.1 Background

Kama, Rose, and Canadian are three common kinds of wheat that are being widely produced in the form. This dataset was obtained from the UCI machine learning repository website included many features of these three kinds of wheat. Agronomists and agriculturists are interested in researching these three kinds of wheat in order to tell the differences between their features.

## 1.2 Goal

In this research, I am interested in finding the relationship between the features of the wheats and the kind they belong to. In the meantime, I would also like to find out what are some of the outstanding aspects each wheat has shown. Different techniques would be applied under various analysis through data visualization, correlation analysis, and classification. I will implement and interpret these techniques in the following report.

# II. Dataset:

## 2.1 Data Summary

The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, randomly selected for the experiment. High quality visualization of the internal kernel structure was detected using a soft X-ray technique. It is non-destructive and considerably cheaper than other more sophisticated imaging techniques like scanning microscopy or laser technology. The images were recorded on 13x18 cm X-ray KODAK plates. Studies were conducted using combine harvested wheat grain originating from experimental fields, explored at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin. (https://archive.ics.uci.edu/ml/datasets/seeds). I will take the 7 different features of wheat as our input and choose the species of each kind of wheat as our output for our model and all of them are numerical.

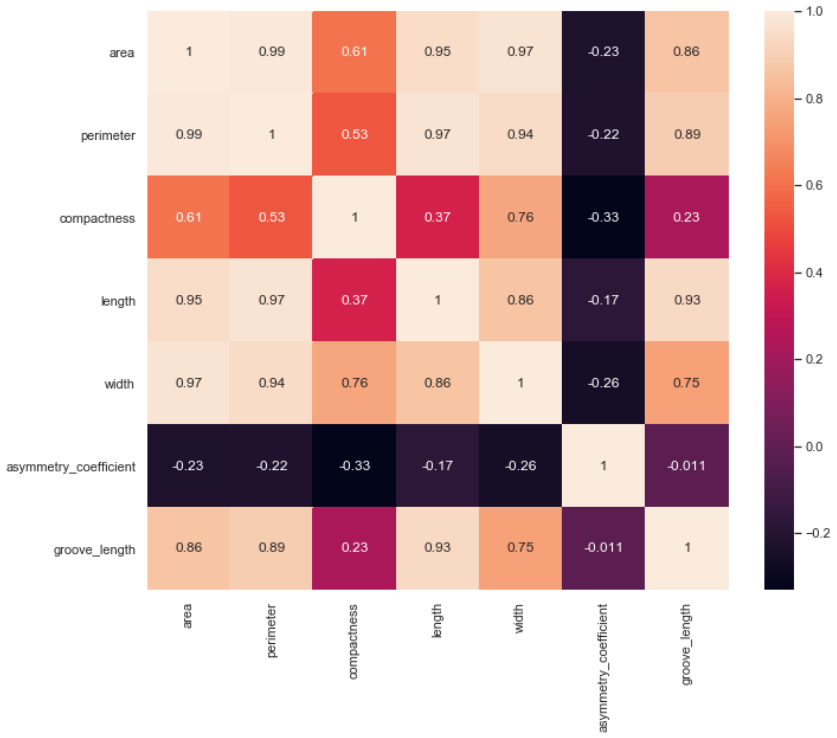
**Attributes:**

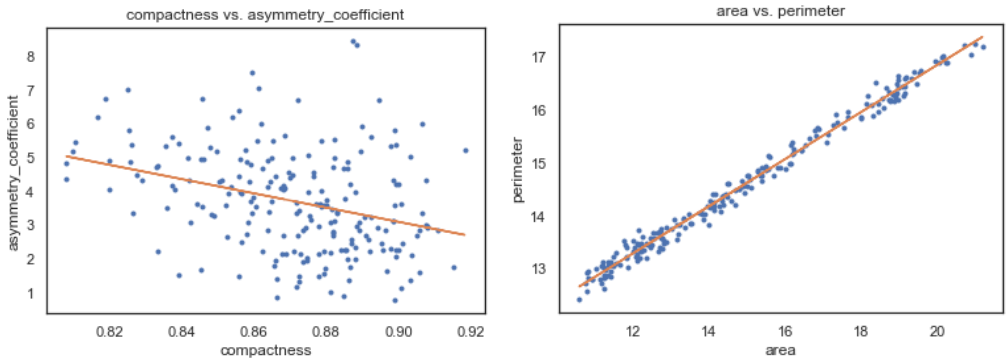
| Inputs | Descriptions |
|---|---|
| Area | Area of the wheat kernel. |
| Perimeter | Perimeter of the wheat kernel. |
| Compactness | C = 4*pi*A/P^2, Compactness of the wheat kernel. |
| Length | Length of wheat kernel. |
| Width | Width of wheat kernel. |
| Asymmetry Coefficient | The skewness of the wheat kernel. |
| Groove Length | Length of kernel groove. |

Output:

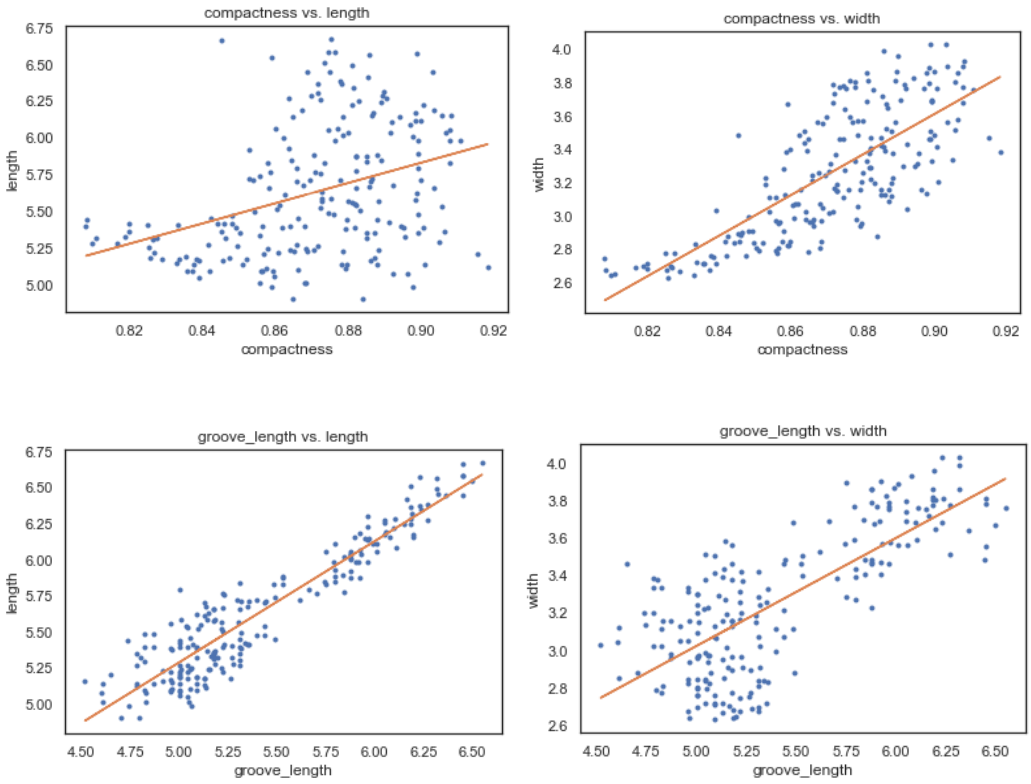
| | |
|---|---|
| Grain Variety | The species of the different wheat kernel. |

## 2.2 Data Visualization




In this step, I was interested in finding the relationship between each feature. By creating the Correlation plot for 7 observation variables, I was able to identify their correlation value via opacity of the color. Thus, I found that all correlation except asymmetry coefficient among these variables are positive. The area, perimeter, width, and length have the largest average correlation among all other variables. Moreover, the correlation between area and perimeter is 0.99 which indicates a very high positive correlation.

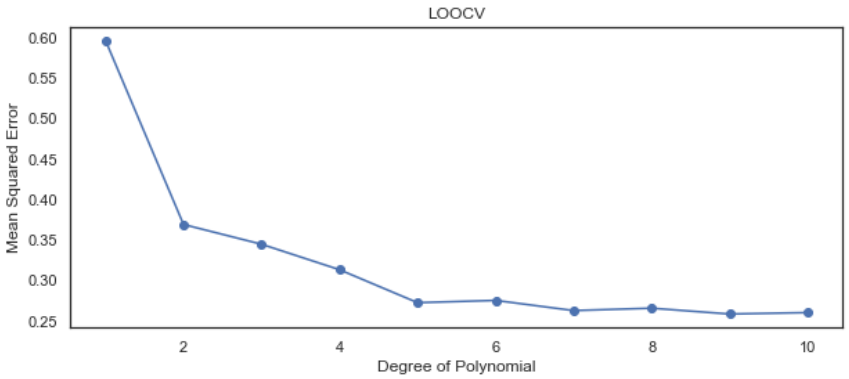



An interesting noticing is that compactness does not have strong correlation with all other variables except for asymmetry coefficient. Groove length is highly correlated with kernel length but not seed width, however, compactness is the opposite.

# III. Main Analysis:

## 3.1 Leave One Out Cross Validation

In this section, I conduct the LOOCV to analyze the data and the significance portance of the predicting variable as a 1-dimension example for analysis. Since the variable “area” has the average largest overall correlation with other variables, I picked it as the predicting variable for the polynomial model for cross validation.




In this figure, it shows that the elbow point of this polynomial function is degree 5, which means the model will be most efficient at polynomial degree equals to 5. Under this model, the R square equals to 0.608 which means 60.8 percent of the data is explained in this model, and all variables under this model are significant since the p-value of the parameter are equals to 0.

## 3.2 Grid Search with Random Forest

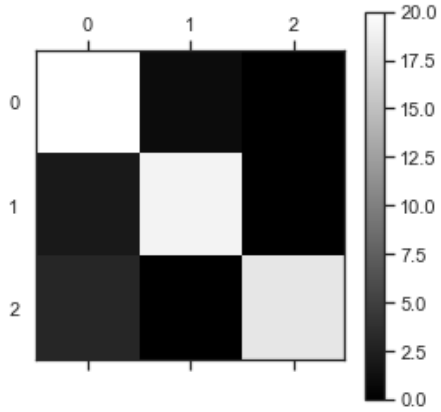
Considering this dataset contains multiple labels. Grid search with random forest method was implemented in order to perform better classification. I set limitation of max features to be 1 to 7, the number of estimators either 3, 10, or 30, and the model can either use bootstrap or not when classifying.

| | bootstrap | max_features | n_estimators | accuracy |
|---|---|---|---|---|
| **36** | False | 6 | 3 | 0.952381 |
| **31** | False | 4 | 10 | 0.952381 |
| **20** | True | 7 | 30 | 0.952381 |

Above table shows the grid search result under the random forest method. It indicates that when bootstrap is False, and we limited the max features of the model to be 6 and choose the number of the estimators to be 3, the classification model will give the final prediction accuracy of 0. 952381 on the training dataset, which is a relatively high accuracy. I then implemented this

model to run through the test dataset, I gathered the result and arranged them into the following confusion matrix.

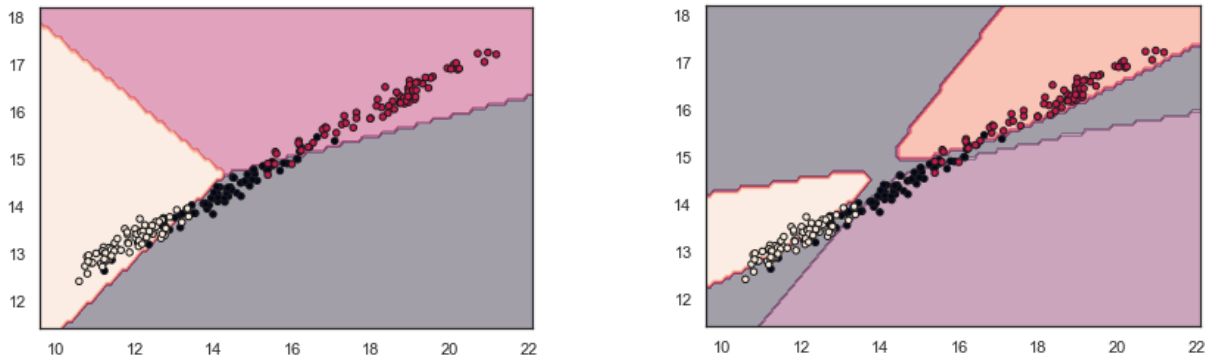
| Predict<br>True | 0<br>(Kama) | 1<br>(Rosa) | 2<br>(Canadian) |
|---|---|---|---|
| 0 (Kama) | 20 | 1 | 0 |
| 1 (Rosa) | 2 | 19 | 0 |
| 2<br>(Canadian) | 3 | 0 | 18 |



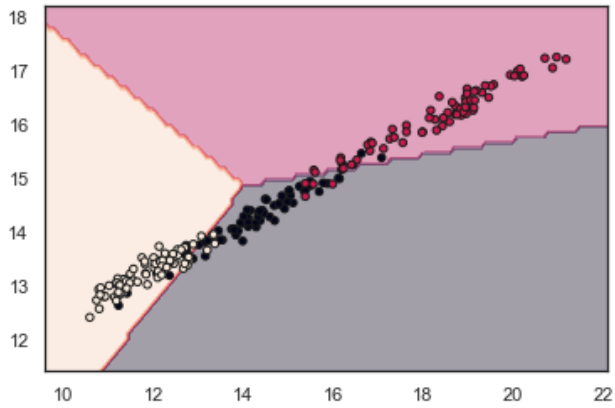

The above figure also shows the classification result. If the color of the block is closer to white on the left diagonal blocks, it means the model predicts the data correctly. Calculated from the table and figure indicates, the accuracy of this model which grid search suggests is optimal turns out to be 0.9048, which is about 5 percent lower than predicting the training dataset.

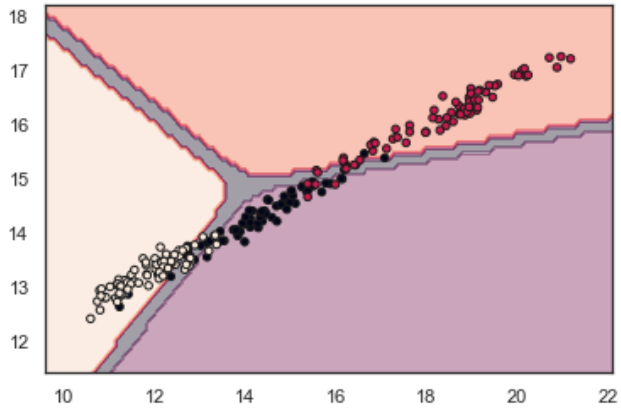
## 3.3 Multiclass Classification

Thinking of some of the variables within the dataset contains multiple dimension, the multiclass classification was performed to achieve more visualized result. Since variables “area” and “perimeter” are highly associated and have high overall correlation with other variables, these two variables were chosen as the predicting variables in this classification.




The two figures above show the result of multiclass classification under One vs. Rest method. The three different colored regions indicate the different species each wheat kernel belongs to. In the right figure, I added confidence level into the figure; thus, only points that are in the yellow, green, and red regions are confidently classified.




I also tried to classify the dataset through the multinomial logistic regression method with the same predicting variables. The result without confidence level are similar, however, when adding confidence level, the multinomial regression does a much better job on confidently classify the data points.

# IV. Conclusion:

This study provides multiple ways to predict and analysis the data under different circumstances. For 1-dimensional analysis, I implemented the LOOCV with polynomial model to predict the kernel species. The result suggests when polynomial degree equals to 5, the model will perform most efficiently. For K-dimensional analysis, I conducted grid search with random forest method and multiclass classification. The grid search with random forest provides relatively high accuracy with training dataset. The accuracy is about 5% lower when passing the model through the testing dataset, but it is acceptable. I then implemented the MCC under One vs. Rest and multinomial logistic regression method. Without confidence level, both methods

provide similar results, but when adding the confidence level, multinomial logistic classification performs a better job than the One vs. Rest method.