Wheat Seeds Analysis

ASSESSMENT – 2

COMP7006 – DATA SCIENCE

Mohit Mehndiratta, 20622275
Master of Data Science

Data Exploration & Objective

- Data is gathered from UCI Machine Learning Repository.
- Data consist of 210 observation for 8 variables.
- Each variable is numeric variable except "Type" which is a factor variable.
- each variable is positively correlated except the Asymmetry Coefficient which is negatively correlated with each other.

Aims & Objectives:

- Correctly classify the "Kama" wheat type from seeds data set
- Know which geometrical properties are more important to classify the wheat type "Kama".
- To experiment with different classification methods to which methods yields the highest accuracy to resolve the problem.

Supervised Learning: Logistic Regression vs Decision trees

Logistic Regression

- Built 4 models with different combination of variables.
- Cross validation plot explained model 2 is better.
- AIC values for all models explained model 2 is better.
- Accuracy of Model 2 is 95.24% approximately.

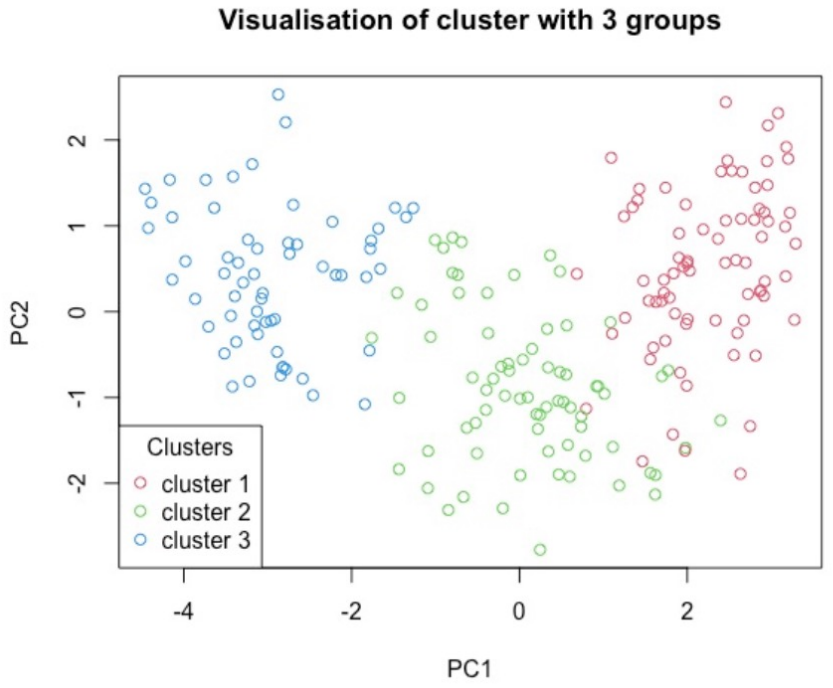
Decision Trees

- Full tree built using training data which used 4 variables with 8 terminal nodes.
- Cross validation explained that tree with size 5 is the best tree.
- Pruned tree obtained with 5 terminal nodes.
- Accuracy of pruned tree is 85.72% approximately.

Seems like model 2 of Logistic regression is better as it yields high accuracy.

Unsupervised Learning - Clustering

- Built clusters with 2, 3 and 4 centers.
- Centroid value for cluster 3 and 4 is quite similar for all the variables.
- After comparing the clusters with total withinss, cluster with center 3 is more appropriate.
- Visualization is done with the help of PCA, similar observations are clustered together



Results & Recommendations

- From supervised learning, logistic regression yields better accuracy than decision trees.
- Model 2 of logistic regression which was fitted for “is_kama” variable against all geometrical properties of seeds except “Width of Kernel” classified with approx. 95.24% accuracy.
- Classification for “Kama” seeds can be done with all of its geometrical properties except “Width of Kernel”, which can yield better result.