**Assignment Two**

Semester 1 2022

Student Names: Ruiyang Xu; Jiali Lu

Student IDs: 18017865;21135614

## PAPER NAME: Foundations of Data Science

PAPER CODE: COMP615

Due Date: 5th Jun 2022 (midnight NZ time)

TOTAL MARKS: 100

Instructions:

1. The following actions may be deemed to constitute a breach of the General Academic Regulations Part 7: Academic Discipline,

* Communicating with or collaborating with another person regarding the Assignment
* Copying from any other student work for your Assignment
* Copying from any third-party websites unless it is an open book Assignment
* Uses any other unfair means

1. Please email [DCT.EXAM@AUT.AC.NZ](mailto:DCT.EXAM@AUT.AC.NZ) if you have any technical issues with your Assessment/Assignment/Test submission on Canvas immediately
2. Attach your code for all the datasets in the appendix section.

# Acknowledgement

In writing this paper, we have received help, support, and inspiration from COMP615 teaching team. I would like to express our deepest gratitude to Dr Akbar, for his kind clarification and support. Our special thanks are given to Dr Sara and Dr Nawal, for their inspiring suggestions and support in all lab sessions, which have enabled us to produce this paper.

# Abstract

The technological explosion has paved the way for agriculture to flourish exponentially thus contributing to better yield of crops through the aid of machine learning, and mechanical systems in agriculture. In this paper, we have investigated various types of dry beans followed by a deep neural network-based approach to classifying the beans automatically. The results show that our approach had an accuracy of 93.33% with the dataset that consisted of 7 varieties of dry beans.

[PAPER NAME: Foundations of Data Science 1](#_Toc105341092)

[Acknowledgement 2](#_Toc105341093)

[Abstract 3](#_Toc105341094)

[1. Introduction 7](#_Toc105341095)

[1.1 General Background 7](#_Toc105341096)

[1.2 Purpose of this study 7](#_Toc105341097)

[2. Data Exploration 8](#_Toc105341098)

[2.1 Features, Instances and Data types of Dry Bean Dataset 8](#_Toc105341099)

[2.2 Dataset Inputs and the Output (Class) 9](#_Toc105341100)

[2.3 Summary Statistics of the Continuous Numerical Features. 11](#_Toc105341101)

[3.Decision Tree Classifier 14](#_Toc105341102)

[3.1 An optimal Decision Tree 15](#_Toc105341103)

[3.2 The Role of the Two Parameters in Building a Decision Tree 19](#_Toc105341104)

[3.3 Confusion Matrix 21](#_Toc105341105)

[3.4 Feature Importance 22](#_Toc105341106)

[4 Artificial Neural Network (ANN) 22](#_Toc105341107)

[4.1 Five Most Significant Features 23](#_Toc105341108)

[4.2 Single hidden layer clarification Accuracy 24](#_Toc105341109)

[4.3 Function for Loss values and Test Segments 25](#_Toc105341110)

[4.4 Two hidden Layers Classification Accuracy. 26](#_Toc105341111)

[4.5 Accuracy Variation with Two layers 27](#_Toc105341112)

[5 Performance Evaluation 27](#_Toc105341113)

[Reference: 28](#_Toc105341114)

List of Tables

[Table 1 Feature, instances and data types of Dry bean dataset 8](#_Toc105341083)

[Table 2 basic statistics of the dataset 9](#_Toc105341084)

[Table 3 summary statistics of the continuous numerical features of dry bean dataset. 9](#_Toc105341085)

[Table 4 the total number of null values in the dataset 10](#_Toc105341086)

[Table 5 Correlation between Different features 12](#_Toc105341087)

[Table 6 16 different features of the dataset 16](#_Toc105341088)

[Table 7 Feature importance 22](#_Toc105341089)

[Table 8 a single hidden layer with various k neurons (k<=25) 25](#_Toc105341090)

[Table 9 **Two hidden Layers Classification Accuracy** 27](#_Toc105341091)

List of Figures

[Figure 1 7 different bean types 10](#_Toc105341071)

[Figure 2 Continuous Numerical Features 11](#_Toc105341072)

[Figure 3 Correlation between Different features 13](#_Toc105341073)

[Figure 4 The selected features of dry beans (pair plot). 14](#_Toc105341074)

[Figure 5 dry bean types 15](#_Toc105341075)

[Figure 6 Decision Tree 18](#_Toc105341076)

[Figure 7 Accuracy via max depth 19](#_Toc105341077)

[Figure 8 Accuracy via max leaf\_nodes 20](#_Toc105341078)

[Figure 9 confusion matrix 21](#_Toc105341079)

[Figure 10 Significant Features 23](#_Toc105341080)

[Figure 11 Function for Loss values 25](#_Toc105341081)

[Figure 12 Function for Test Segments 26](#_Toc105341082)

# 1. Introduction

## 1.1 General Background

Dry beans are one of the largest edible legume crops in the world, and their genetic diversity is wide. As seed quality has a decisive impact on crop production, the classification of dry beans is crucial not only because of economic importance but also because manual classification is labour intensive.

A picture containing food, vegetable, variety, containing

Description automatically generated

## 1.2 Purpose of this study

As dried beans are the most important leguminous legumes with the largest yield in the world, and seed quality has a decisive impact on crop production, this study sheds light on classifying dry bean varieties obtained from production. This has been analysed via several machine learning and artificial neural network techniques.

# 2. Data Exploration

## 2.1 Features, Instances and Data types of Dry Bean Dataset

This data set mainly recorded different types of dry beans. A total of 13611 samples were recorded, including 16 different main features (12 dimensions and 4 shape forms). A sample amounts to 16 geometrical features and a label to classify various types of beans. The species of beans are: Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sira and the rest of features are Area, Perimeter, Major Axis Length, Minor Axis Length, Aspect Ratio, Eccentricity, Convex Area, Equivalent Diameter, Extent, Solidity, Roundness, Compactness, Shape Factor 1 to 4...

The following table reveals the feature, instance, and data types of dry beans over 13 k samples of dry beans of 7 various species and their geometry.

Text

Description automatically generatedTable

Description automatically generated

Table 1 Feature, instances and data types of Dry bean dataset

## 2.2 Dataset Inputs and the Output (Class)

Table

Description automatically generated

Table 2 basic statistics of the dataset

Graphical user interface, table

Description automatically generated

Table 3 summary statistics of the continuous numerical features of dry bean dataset.

Table 2 shows the basic information about the dataset, it shows the number of rows and columns, column index, number of columns of null values, type of column and memory occupation. Table 3 shows summary statistics of the continuous numerical features of the dry bean dataset.

Table

Description automatically generated

Graphical user interface, text

Description automatically generated with medium confidence

A picture containing graphical user interface

Description automatically generated

Table 4 the total number of null values in the dataset

Table 4 has counted the total number of null values in the dataset, as the result, there is no null value in the research dataset. It also shows there are 68 duplicate data and dropped these data to avoid interference with the research and test results.

Chart, bar chart

Description automatically generated

Figure 1 7 different bean types

From Figure 1, there are 7 different beans that have been researched and classified. This plot shows the total number of different types of dry beans. According to the plot, class distribution is not balanced. The DERMASON has the largest total number, while Bombay has the least number of dry beans.

## 2.3 Summary Statistics of the Continuous Numerical Features.

A picture containing text, dog

Description automatically generated

Figure 2 Continuous Numerical Features

According to Figure 2, there are some plots that are very similar, because they have similar distribution, such as Area, Perimeter, Major axis length and Minor axis length, the roundness and extend. But the four Shape factor both have different distributions.

Table

Description automatically generated

Table 5 Correlation between Different features

Table

Description automatically generated

Figure 3 Correlation between Different features

As we can see from the above heat map, some Features have a very strong correlation with each other, such as area and convex area, the correlation number is 1, and the Equiv Diameter has a very strong correlation with convex area (0.99) and perimeter (0.99). Besides, there are some very strong negative correlations in this plot. Compactness has -0.99 and -0.97 with Aspect Ratio and Eccentricity. There are some positive or negative correlations between features and shape factors. But it’s now quite clear how shape factors could affect or are affected by other features.

A picture containing text

Description automatically generated

Figure 4 The selected features of dry beans (pair plot).

In this plot, we could clearly observe that there is a scatter area is very different from the other main scatter area. It is speculated that the main feature of a special type of dry bean are obviously different from the other type of dry bean.

# 3.Decision Tree Classifier

A decision tree (DT) belongs to the class of so called non-parametric algorithms. In fact, a decision tree has parameters, but their number is not constant. A decision tree tries to find the best questions by partitioning the dataset to reduce information impurity. The great advantage of decision trees is that they are extremely intuitive. On the other hand, a decision tree has no limited degrees of freedom, so it is easy to overfit. The splits made by a decision tree are always orthogonal (made on one feature at a time), so the decision tree is very sensitive to data rotation.

## 3.1 An optimal Decision Tree

A picture containing diagram

Description automatically generatedDiagram

Description automatically generated

Figure 5 dry bean types

According to Figure5, the plot of the boxplot has shown there are many outliers in the dataset, which means that one dry bean type is much larger than the other dry bean types. As we can see from Figure 6, the Bombay classes are different from the other classes. Especially the Bombay type. This type of dry bean is always bigger than other types.

Table

Description automatically generated

Table 6 16 different features of the dataset

This plot displays the 16 different features of the dataset and rename the types of dry bean to different integers.

Table

Description automatically generated

Table

Description automatically generated

A picture containing calendar

Description automatically generated

A picture containing text

Description automatically generated

A picture containing graphical user interface

Description automatically generated

According to the table above, the whole dataset has been split into 2 parts (test and train) and the size is 0.3. The above plots are the accuracy of the 2 sets, the train set and the test set.

Diagram, engineering drawing

Description automatically generated

Figure 6 Decision Tree

According to this decision tree, the set has been classified by 16 different features, the max depth is 6 and there are 19 nodes. Total samples are 9480. The first condition is No.2 feature is less than 280.397 or large than that and gini is 0.828. This set has 930(SEKER), 388(BARBUNYA), 1122(BOMBAY), 2474(CALI), 1296(DERMASON), 1449(HOROZ), 1821(SIRA). After the above condition, the set has been split to 2 parts. the condition for left part is No.12 feature less than 0.007 or bigger than that, gini is 0.54, and there are 3(SEKER), 0(BARBUNYA), 0(BOMBAY), 2351(CALI), 9(DERMASON), 1340(HOROZ), 314(SIRA). Total for this part is 4017. The right one in the next depth is one of result for this decision tree, it has 2566 samples. Main type is CALI (2268), and 8(DERMASON), 73(HOROZ), 217(SIRA). The left part will be continued to classify.

There are 1451 samples, condition is No.14 feature is less than 0.727 or bigger than that. 3(SEKER), 0(BARBUNYA), 0(BOMBAY), 83(CALI), 1(DERMASON), 1267(HOROZ) and 97(SIRA) in this part. We will get 2 results here, the left result is 0(SEKER), 0(BARBUNYA), 0(BOMBAY), 56(CALI), 1(DERMASON), 26(HOROZ), 82(SIRA). The right result is 3(SEKER), 0(BARBUNYA), 0(BOMBAY), 27(CALI), 0(DERMASON), 1241(HOROZ), 15(SIRA). The rest of this decision tree will be classified by same condition.

## 3.2 The Role of the Two Parameters in Building a Decision Tree

In this paper, one parameter is limiting the depth of the decision tree, and another is limiting the max number of nodes to improve the accuracy.

Chart, line chart

Description automatically generated

Figure 7 Accuracy via max depth

Chart, line chart

Description automatically generated

Figure 8 Accuracy via max leaf\_nodes

Figure 6 has trying to find the accuracy via max depth and the accuracy for this is not very high. The highest number is 0.763 when the max depth is equal to 6. To improve the accuracy of this dataset, we are trying to use the Standard Scaler to get higher accuracy. However, the dataset still gets the same value based on the above figures.

## 3.3 Confusion Matrix

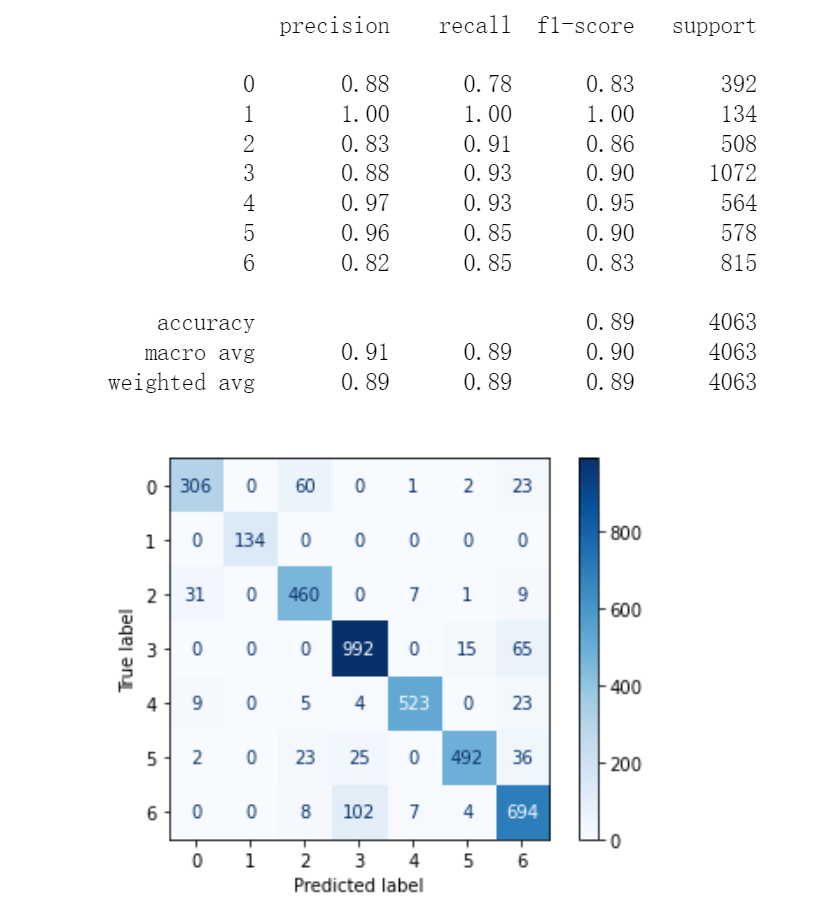


Figure 9 confusion matrix

This confusion matrix and summary report has displayed the total number of predict. For the type 0(SEKER) dry bean, there are 348 true samples, 392 predict samples, in this set, 306 samples are correct. Therefore, precision = 306 / (306 + 31 + 9 + 2) = 306 / 348 = 0.88, recall = 306 / (306 + 60 +1 + 2 + 23) = 306 / 392 = 0.78, f1-score = 2 \* precision \* recall / (precision + recall) = 1.3728 / 1.66 = 0.83. This type 1(BARBUNYA) dry bean has the highest precision, recall and f1-score, both get 1. Presumably, it has the least total number of samples in this test set. Type 3(CALI) has the largest number of samples in this test set, it has 1072 samples. 992 samples are correct predict. Precision = 992 / (992 + 4 + 25 + 102) = 992 / 1123 = 0.88, recall = 992 / (992 + 15+ 65) = 992 / 1072 = 0.93, f1-score = 2 \* precision \* recall / (precision + recall) = 1.6368 / 1.81 = 0.90. The rest of type is same methods to calculate.

## 3.4 Feature Importance

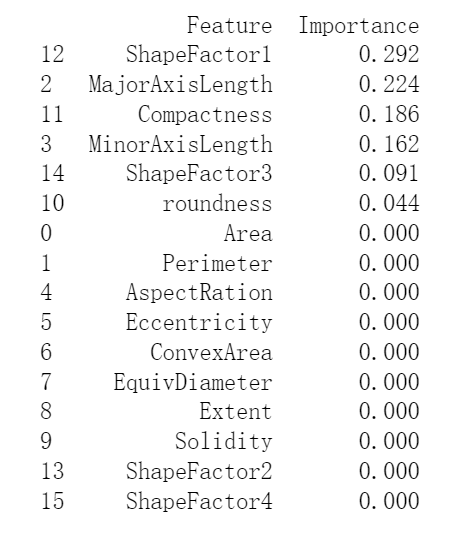


Table 7 Feature importance

According to this plot, the feature that most affects dry bean classification is the shape factor 1, which has 0.29, then is the major axis length, which has 0.22.

# 4 Artificial Neural Network (ANN)

Besides the machine learning presented above, a deep learning technique, the artificial neural network has also been tried in this paper. For an artificial neural network, the data needs additional treatment.

## 4.1 Five Most Significant Features

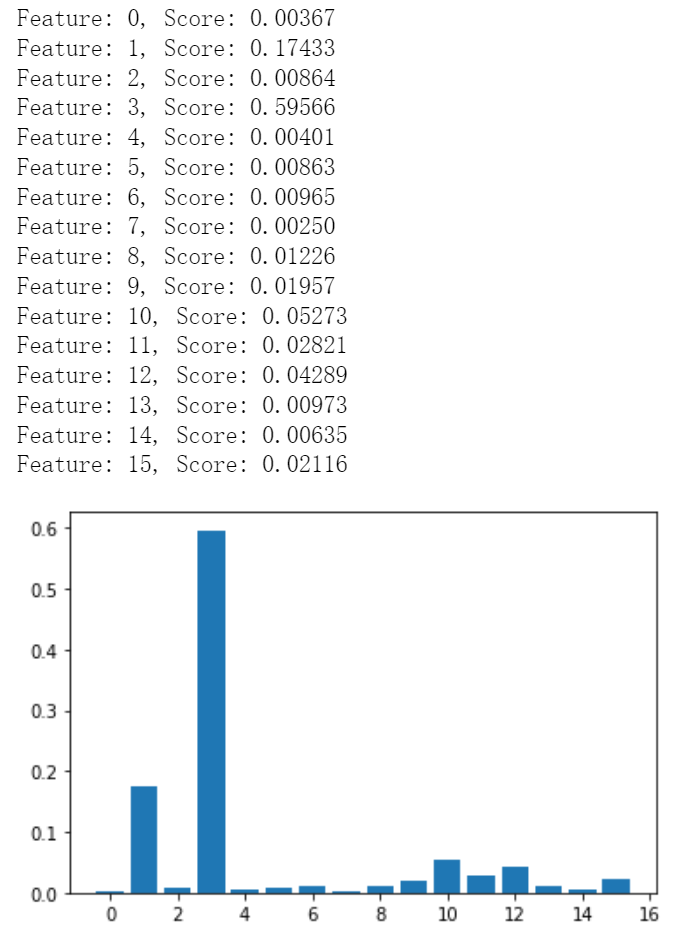


Figure 10 Significant Features

We use the Classification and expression trees (CART) provide importance scores based on Gini coefficient or entropy reduction. In this plot, we use the Decision Tree Regressor method to get the importance score. According to this method, the highest score is the feature 3(Minor Axis Length), then is the feature 1(Perimeter). Next 3 features are feature 10 (roundness), feature 12(shape factor 1) and feature 11(compactness). Thus, the top five most significant features are: feature 3, feature 1, feature 10, feature 12, feature 14.

## 4.2 Single hidden layer clarification Accuracy

Below is the table for a single hidden layer with various k neurons (k<=25). We can clearly see with the increase of the k the accuracy increases as well to its max value 93.23% when k = 22. Besides, the number of iterations has reached limit 500 but hasn’t reached the best accuracies, but after k > =3, we can see all numbers of iterations are round 200-300.

|  |  |  |
| --- | --- | --- |
| hidden layer k | Number of iterations | classification accuracy (%) |
| 1 | 500 | 60.52178193453114 |
| 2 | 500 | 90.22889490524243 |
| 3 | 317 | 91.26261383214373 |
| 4 | 235 | 91.8040856509968 |
| 5 | 296 | 91.90253507260645 |
| 6 | 246 | 92.1978833374354 |
| 7 | 198 | 92.37016982525228 |
| 8 | 212 | 92.14865862663056 |
| 9 | 216 | 92.4193945360571 |
| 10 | 287 | 92.76396751169086 |
| 11 | 257 | 92.71474280088604 |
| 12 | 300 | 92.73935515628847 |
| 13 | 255 | 92.96086635491017 |
| 14 | 258 | 92.81319222249569 |
| 15 | 236 | 92.51784395766674 |
| 16 | 291 | 93.10854048732465 |
| 17 | 329 | 93.01009106571499 |
| 18 | 220 | 92.86241693330052 |
| 19 | 349 | 93.13315284272706 |
| 20 | 313 | 93.05931577651981 |
| 21 | 317 | 93.05931577651981 |
| 22 | 245 | 93.2316022643367 |
| 23 | 265 | 93.08392813192222 |
| 24 | 261 | 93.15776519812945 |
| 25 | 307 | 93.20698990893429 |

Table 8 a single hidden layer with various k neurons (k<=25)

## 4.3 Function for Loss values and Test Segments

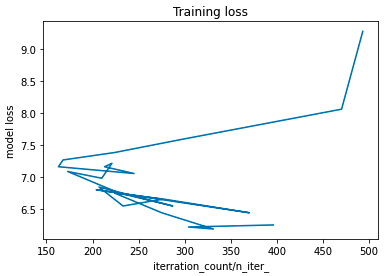


Figure 11 Function for Loss values

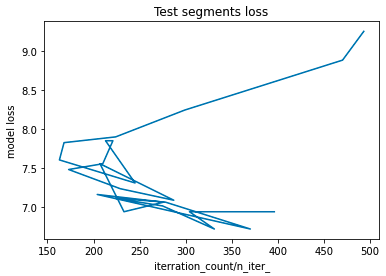


Figure 12 Function for Test Segments

From the plots above, we can see when the iteration count/n\_iter\_ is high (close to 500) model loss for both training and test segments are high (>9%). In contrast, when the iteration count/n\_iter\_ is around 300~400, the model loss for the Training is lowest (< 6.5%); when the iteration count/n\_iter\_ is around 325~375, the model loss for the test segments is lowest (< 6.8%).

## 4.4 Two hidden Layers Classification Accuracy.

|  |  |
| --- | --- |
| 24, 1 | 93.30543933054393 |
| 23, 2 | 93.2316022643367 |
| 22, 3 | 92.86241693330052 |
| 21, 4 | 93.10854048732465 |
| 20, 5 | 93.10854048732465 |
| 19, 6 | 93.33005168594634 |
| 18, 7 | 93.2316022643367 |
| 17, 8 | 93.0347034211174 |
| 16, 9 | 92.98547871031258 |
| 15, 10 | 92.96086635491017 |
| 14, 11 | 93.13315284272706 |
| 13, 12 | 93.01009106571499 |
| 12, 13 | 93.20698990893429 |
| 11, 14 | 92.98547871031258 |
| 10, 15 | 92.78857986709328 |
| 9, 16 | 93.13315284272706 |
| 8, 17 | 93.30543933054393 |
| 7, 18 | 93.08392813192222 |
| 6, 19 | 92.93625399950776 |
| 5, 20 | 92.96086635491017 |
| 4, 21 | 92.91164164410533 |
| 3, 22 | 93.2316022643367 |
| 2, 23 | 92.93625399950776 |
| 1, 24 | 92.81319222249569 |

Table 9 **Two hidden Layers Classification Accuracy**

## 4.5 Accuracy Variation with Two layers

From the table we can see when the first hidden layer is 16 and the second hidden layer is 9, MLP Testing set score is highest, 93.33%; Compared to the best result in 4 b) (93.23%), we can see transferring neurons from the first hidden layer to the second layer, can increase the classification accuracy.

One of the reasons could be splitting the neurons into different hidden layers could benefit the accuracy. Based on our current dataset, adding another hidden layer can optimize our classification accuracy.

# 5 Performance Evaluation

According to the 2 different classifier methods, the MLP classifier the better method to classify the dry beans. The accuracy of MLP classification method is more than 93% and Decision tree only have 81%.

For the classification of dry beans, the method with higher accuracy could more accurately obtain the classification results of corresponding kinds of dry beans from a very large number of dry bean samples. For this test, we have achieved a high accuracy rate via MLP classifier method, completed the set goal of the test, and the test results are successful.

# Reference:

1. J. Arun Pandian, G. Geetharamani and B. Annette, "Data Augmentation on Plant Leaf Disease Image Dataset Using Image Manipulation and Deep Learning Techniques,"

2. P. Bir, R. Kumar and G. Singh, "Transfer Learning based Tomato Leaf Disease Detection for mobile applications,"

3. S. Bock and M. Weiß, "A Proof of Local Convergence for the Adam Optimizer," 2019 International Joint Conference on Neural Networks (IJCNN)

4. N. E. West and T. O'Shea, "Deep architectures for modulation recognition," 2017

5. M. A. Mercioni and S. Holban, "The Most Used Activation Functions: Classic Versus Current," 2020