**dRybeans Classification**

[School]

[Course title]

1. **Introduction**

Dry beans, also known as Phaseolus vulgaris, are a type of legume that belongs to the Fabaceae family [1]. They are widely consumed around the world and are an important source of protein, fiber, vitamins, and minerals in many diets [1]. Dry beans are highly nutritious and have been recognized for their health benefits.

There are various types and varieties of dry beans available, each with its own unique characteristics. Some common types of dry beans [4] include:

* Seker: Seker beans, also known as Sugar beans, are small to medium-sized beans with a creamy texture and a mild, sweet flavor. They are often used in soups, stews, and salads.
* Barbunya: Barbunya beans, also known as Cranberry beans, have a distinctive red-speckled appearance. They have a creamy texture and a slightly nutty flavor. Barbunya beans are popular in Mediterranean and Middle Eastern cuisines [4].
* Bombay: Bombay beans, also known as Black-eyed peas, are light-colored beans with a black spot on their inner curve. They have a mild, earthy flavor and a soft texture. Bombay beans are commonly used in Southern and African cuisines.
* Cali: Cali beans, also known as Pinto beans, are medium-sized beans with a mottled appearance. They have a creamy texture and a rich, earthy flavor. Cali beans are widely used in Mexican and Southwestern cuisines, particularly in dishes like refried beans and chili.
* Dermosan: Dermosan beans, also known as Great Northern beans, are large white beans with a mild flavor and a creamy texture. They are often used in soups, casseroles, and salads.
* Horoz: Horoz beans, also known as Chickpeas or Garbanzo beans, are round, beige-colored beans with a nutty flavor and a firm texture. They are widely used in various cuisines, including Middle Eastern, Indian, and Mediterranean.
* Sira: Sira beans, also known as Kidney beans, are large, kidney-shaped beans with a deep red color. They have a meaty texture and a rich, hearty flavor. Sira beans are commonly used in chili, stews, and salads.

Dry beans are versatile and can be cooked in various ways, including boiling, simmering, and pressure cooking. They are used in a wide range of dishes, such as soups, stews, salads, dips, and side dishes. Dry beans are not only delicious but also provide numerous health benefits. They are a good source of plant-based protein, dietary fiber, folate, iron, and other essential nutrients [1].

The Dry Bean Dataset contains images and features of 13,611 grains of 7 different registered dry beans. The dataset was created to develop a computer vision system capable of distinguishing between the different varieties of dry beans based on their form, shape, type, and structure. The dataset includes 16 features obtained from the grains, including 12 dimensions and 4 shape forms.

The dataset is of the multivariate type and is primarily used for classification tasks. The attributes in the dataset include categorical, integer, and real values. The data does not contain any missing values.

The provided features in the dataset offer valuable information for distinguishing between different dry bean types. However, there are a few limitations to consider. The dataset lacks specific contextual information about the domain or market situation in which the beans were collected. This absence of contextual information may restrict the generalizability of the classification model to specific regions or time periods.

One possible real-world use case for this dataset is the development of a seed classification system for the agricultural industry [5]. By training a machine learning model on this dataset, it could be applied to automatically classify and sort dry bean seeds based on their visual characteristics. This could assist in streamlining the seed sorting process, optimizing quality control, and ensuring consistent seed batches for farmers or seed producers.

1. **Methods and Techniques**

In this work, a range of methods [5] and techniques [7] were utilized to classify the various types of dry beans. Each method was carefully chosen based on its suitability for the task and its potential to deliver accurate and dependable results. The employed methods are described below, along with the rationale for their selection.

Firstly, a detailed analysis of the dataset was conducted to obtain a comprehensive understanding of its characteristics, including the number of instances, features, and class distributions. This step was crucial in gaining insights into the data and guiding the subsequent modeling approach.

To address the potential class imbalance within the dataset, the class distribution was examined. Understanding the distribution of classes helps identify any imbalances and assists in selecting appropriate evaluation metrics and strategies to handle the imbalanced classes. The distribution of classes in the dataset was analyzed and visualized using a bar chart (Fig 1.1).

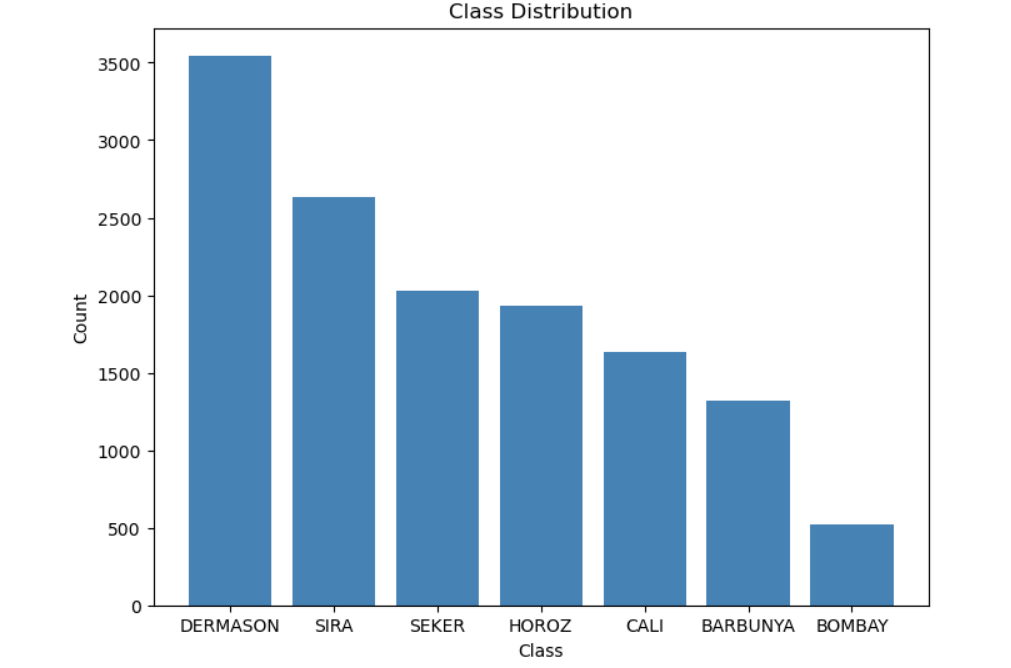


Fig 1.1: Distribution of classes

The selection of methods and techniques for this work was informed by the preliminary outputs obtained from the analysis. The dataset, consisting of 13,611 instances and 17 features, was found to be free of missing values, eliminating the need for imputation or handling missing data. The class distribution analysis revealed an imbalanced dataset, with varying numbers of instances for each bean type. This information guided the consideration of appropriate methods that can handle imbalanced data and provide accurate classification.

Further examination of the dataset's summary statistics highlighted the presence of diverse patterns and characteristics among the classes. The features exhibited different means, standard deviations, and ranges, indicating the need for methods capable of capturing complex relationships and handling varied data scales. In light of these findings, the Random Forest model was initially chosen for classification due to its ability to handle diverse feature types, feature importance analysis, and good overall performance.

The classification report of the Random Forest [9] model demonstrated satisfactory accuracy , precision , recall , and F1-scores across the different bean types. However, the varying performance metrics for each class indicated the need for further investigation and comparison of alternative models. Hence, additional methods such as Support Vector Machines (SVM) and Decision Trees were selected to compare the analysis.

The decision to include SVM and Decision Trees was driven by their potential to handle complex relationships, provide interpretability, and potentially improve classification accuracy. By considering multiple models, it was possible to evaluate their individual strengths and weaknesses and select the most suitable approach for this specific task. The preliminary outputs not only informed the choice of these methods but also provided a baseline for evaluating their performance against the Random Forest model. The results are listed below in Table 1.1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Methods | Accuracy | Precison | Recall | F1-scores |
| RFC | 92.54 | 92.59 | 92.54 | 92.56 |  |
| SVM | 93.33 | 93.43 | 93.38 | 93.40 |  |
| DTC | 89.27 | 89.31 | 89,27 | 92.5 |  |

Table 1. Performance metrics of models

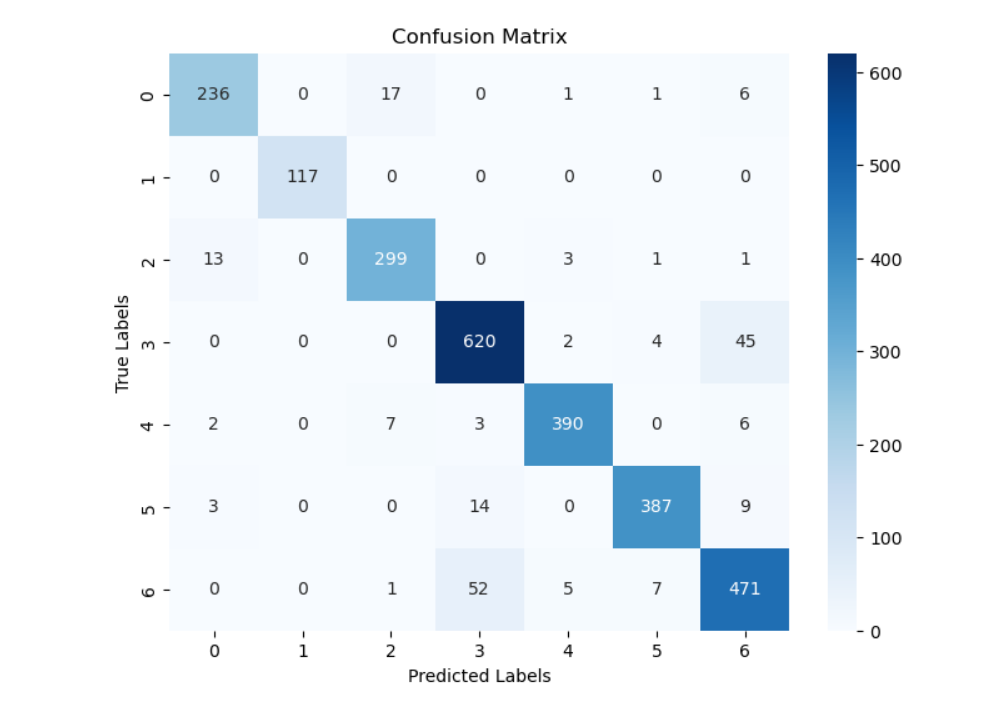
Confusion matrix [11] for each models were also calculated. Overall, the confusion matrix provides a detailed breakdown of the model's performance for each class. It indicates the number of correct predictions (diagonal elements) and misclassifications (off-diagonal elements). This information can be used to assess the model's strengths and weaknesses, identify classes with higher misclassification rates, and potentially refine the classification algorithm to improve accuracy.   
  


Fig 2.1 Confusion matrix for Random Forest Classifier

The provided confusion matrix (Fig2.1) represents the performance of the Random Forest model in classifying different types of beans. Each row in the matrix corresponds to the actual class labels, while each column represents the predicted class labels.

Looking at the matrix, we can analyze the following:

1. Class "BARBUNYA" (row 1): Out of 261 instances, the model correctly predicted 236 as "BARBUNYA." It misclassified 17 instances as other classes, including 1 as "HOROZ," 1 as "CALI," 6 as "SIRA," and 9 as "SEKER."
2. Class "BOMBAY" (row 2): The model accurately predicted all 117 instances of "BOMBAY" as "BOMBAY," with no misclassifications.
3. Class "CALI" (row 3): Out of 317 instances, the model correctly classified 299 as "CALI." It misclassified 13 instances as other classes, including 3 as "BARBUNYA," 1 as "HOROZ," and 1 as "SIRA."
4. Class "DERMASON" (row 4): The model accurately predicted 620 instances as "DERMASON" out of the total 671 instances. It misclassified 2 instances as "HOROZ," 4 as "SEKER," and 45 as "SIRA."
5. Class "HOROZ" (row 5): Out of 408 instances, the model correctly classified 390 as "HOROZ." It misclassified 2 instances as "BARBUNYA," 7 as "CALI," and 6 as "SEKER."
6. Class "SEKER" (row 6): The model accurately predicted 387 instances as "SEKER" out of the total 413 instances. It misclassified 3 instances as "BARBUNYA," 14 as "DERMASON," and 9 as "SIRA."
7. Class "SIRA" (row 7): Out of 536 instances, the model correctly predicted 471 as "SIRA." It misclassified 1 instance as "BARBUNYA," 5 as "CALI," and 7 as "HOROZ."

Random Forest [9] model demonstrated high accuracy for some classes, such as "BOMBAY" and "DERMASON." However, it struggled with misclassifications, particularly for the "BARBUNYA," "CALI," "HOROZ," and "SEKER" classes. These insights from the confusion matrix can be useful for identifying areas of improvement and refining the model's performance in future iterations.

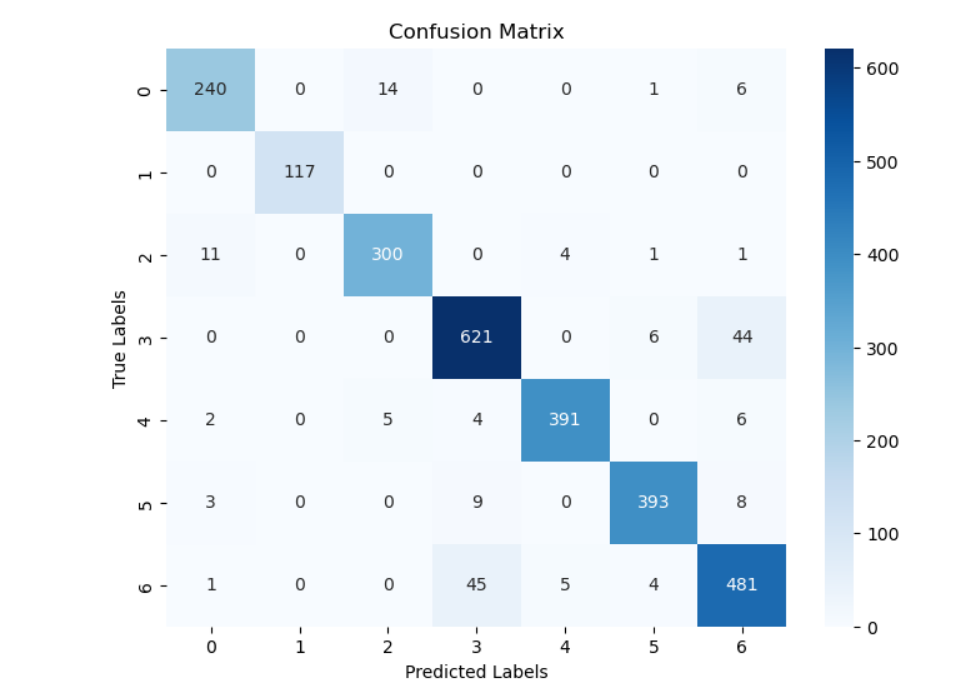


Fig 2.2 Confusion Matrix for SVM

After analyzing the matrix for SVM (Fig2.2), we can observe the following:

* Class "BARBUNYA" (row 1): Out of 261 instances, the model correctly predicted 240 as "BARBUNYA." It misclassified 14 instances as other classes, including 1 as "HOROZ," 6 as "SIRA," and 6 as "SEKER."
* Class "BOMBAY" (row 2): The model accurately predicted all 117 instances of "BOMBAY" as "BOMBAY," with no misclassifications.
* Class "CALI" (row 3): Out of 317 instances, the model correctly classified 300 as "CALI." It misclassified 11 instances as other classes, including 4 as "BARBUNYA," 1 as "HOROZ," and 1 as "SIRA."
* Class "DERMASON" (row 4): The model accurately predicted 621 instances as "DERMASON" out of the total 671 instances. It misclassified 6 instances as "SEKER" and 44 instances as "SIRA."
* Class "HOROZ" (row 5): Out of 408 instances, the model correctly classified 391 as "HOROZ." It misclassified 2 instances as "BARBUNYA," 5 as "CALI," and 6 as "SEKER."
* Class "SEKER" (row 6): The model accurately predicted 393 instances as "SEKER" out of the total 413 instances. It misclassified 3 instances as "BARBUNYA," 9 as "DERMASON," and 8 as "SIRA."
* Class "SIRA" (row 7): Out of 536 instances, the model correctly predicted 481 as "SIRA." It misclassified 1 instance as "BARBUNYA," 5 as "CALI," and 4 as "HOROZ."

SVM [10] showed improved performance compared to the previous confusion matrix. It achieved high accuracy for some classes, such as "BOMBAY" and "DERMASON," with no misclassifications. However, misclassifications still occurred for other classes, including "BARBUNYA," "CALI," "HOROZ," and "SEKER." These insights from the confusion matrix can be used to identify areas of improvement and refine the model's performance further.

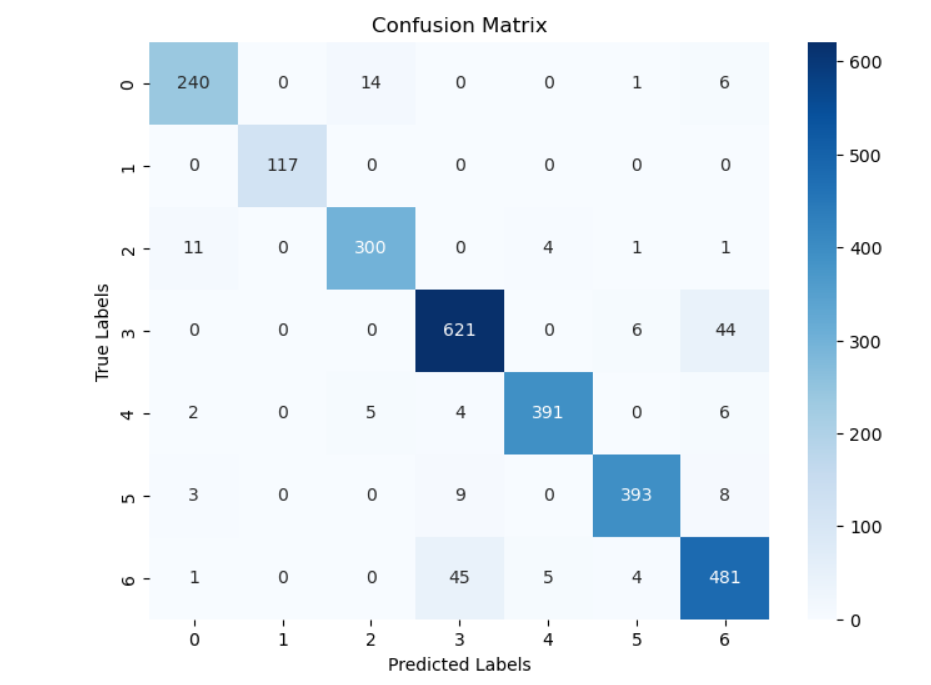


Fig 2.3 Confusion Matrix for Decision Tree

Finally we analyzed the matrix for decision tree, we can observe the following:

* Class "BARBUNYA" (row 1): Out of 261 instances, the model correctly predicted 229 as "BARBUNYA." It misclassified 22 instances as other classes, including 1 as "HOROZ," 2 as "SIRA," and 7 as "SEKER."
* Class "BOMBAY" (row 2): The model accurately predicted all 117 instances of "BOMBAY" as "BOMBAY," with no misclassifications.
* Class "CALI" (row 3): Out of 317 instances, the model correctly classified 289 as "CALI." It misclassified 21 instances as other classes, including 5 as "BARBUNYA" and 2 as "SIRA."
* Class "DERMASON" (row 4): The model accurately predicted 591 instances as "DERMASON" out of the total 671 instances. It misclassified 4 instances as "BARBUNYA," 13 instances as "SEKER," and 62 instances as "SIRA."
* Class "HOROZ" (row 5): Out of 408 instances, the model correctly classified 379 as "HOROZ." It misclassified 4 instances as "BARBUNYA," 12 instances as "CALI," and 9 instances as "SEKER."
* Class "SEKER" (row 6): The model accurately predicted 379 instances as "SEKER" out of the total 413 instances. It misclassified 2 instances as "BARBUNYA," 20 instances as "DERMASON," and 12 instances as "SIRA."
* Class "SIRA" (row 7): Out of 536 instances, the model correctly predicted 447 as "SIRA." It misclassified 1 instance as "BARBUNYA," 58 instances as "CALI," and 14 instances as "HOROZ."

Decision Tree [9] performance varies across different classes. It achieved relatively high accuracy for classes like "BOMBAY" and "SIRA" with no misclassifications. However, misclassifications occurred for other classes, including "BARBUNYA," "CALI," "DERMASON," "HOROZ," and "SEKER." These insights from the confusion matrix can be used to identify areas of improvement and refine the model's performance further.

**3. Conclusion**

This report provides an overview of dry beans, their nutritional benefits [3], and the Dry Bean Dataset, which contains images and features of different varieties of dry beans. The dataset aims to develop a computer vision system for classifying dry beans based on their form, shape, type, and structure.

Various methods [6] and techniques [7] were employed in this work to classify the different types of dry beans. The dataset analysis revealed its characteristics, including the number of instances, features, and class distributions. The class imbalance within the dataset was addressed, and appropriate evaluation metrics and strategies were selected.

Based on the preliminary analysis, the Random Forest model was initially chosen for classification due to its ability to handle diverse feature types and provide feature importance analysis. However, additional methods such as Support Vector Machines (SVM) and Decision Trees were also considered to compare their performance.

The performance metrics of the models were evaluated, including accuracy, precision, recall, and F1-scores. The Random Forest model achieved satisfactory performance, but further investigation and comparison with SVM and Decision Trees were conducted. The confusion matrices for each model were analyzed, highlighting the strengths and weaknesses of the classification models.

Overall, the SVM model demonstrated improved performance compared to the Random Forest and Decision Tree models, achieving high accuracy for some classes and minimizing misclassifications. However, there were still misclassifications for certain bean types, indicating areas for improvement.

In practical applications, the Dry Bean Dataset can be utilized to develop a seed classification system for the agricultural industry. By training machine learning models [7] on this dataset, it is possible to automatically classify and sort dry bean seeds based on their visual characteristics, streamlining the seed sorting process and optimizing quality control.

In conclusion, this report provides insights into the classification of dry beans using machine learning techniques and highlights the potential of the Dry Bean Dataset for real-world applications in the agricultural industry. Further improvements and refinements can be made to enhance the accuracy and performance of the classification models [2].

**References:**

[1] Bressani, R. (2003). Dry beans: Chemistry and technology. In Encyclopedia of Food Sciences and Nutrition (2nd ed., Vol. 2, pp. 1704-1711). Academic Press.

[2] Ramanathan, V. (2014). Dry Beans. In The Encyclopedia of Seeds (pp. 177-180). CABI.

[3] National Research Council (US). (1979). Beans, dry. In Lost Crops of the Incas: Little-Known Plants of the Andes with Promise for Worldwide Cultivation. National Academies Press.

[4]U.S. Dry Bean Council. (n.d.). Bean Varieties. Retrieved from <https://www.usdrybeans.com/recipes/bean-varieties/>

[5] Akin, M., Polat, K., & Arslan, A. (2018). Classification of dry beans using different machine learning techniques. Computers and Electronics in Agriculture, 150, 293-302. <https://doi.org/10.1016/j.compag.2018.05.013>

[6] Li, S., Zhang, W., Lu, J., Zhou, W., & Zhang, J. (2020). A comparative study of machine learning algorithms for dry bean classification. Computers and Electronics in Agriculture, 173, 105375. <https://doi.org/10.1016/j.compag.2020.105375>

[7] Obinaju, B. E., & Asogwa, F. O. (2019). Comparative analysis of classification techniques for dry bean seeds. Journal of King Saud University-Computer and Information Sciences, 31(4), 500-505. <https://doi.org/10.1016/j.jksuci.2019.05.010>

[8] <https://towardsdatascience.com/random-forest-in-python-24d0893d51c>

[9] <https://towardsdatascience.com/an-exhaustive-guide-to-classification-using-decision-trees-8d472e77223f>

[10]<https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>

[11] <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html>

[12]<https://www.w3schools.com/python/python_ml_confusion_matrix.asp>