Research Article

"Comparative Analysis of Machine Learning Models for Accurate Classification of Dry Bean Varieties Using Morphological Features"

Nan Xiao Fei 南晓斐 , Syed Zakaria Mehmood 资恺

School of Computer Science & Artificial Intelligence.

Zhengzhou University, Zhengzhou city, Henan.

1. Abstract

In agriculture, precise crop variety classification is essential for improving quality control and resource management. In order to categorize dry bean types, this study assesses the performance of three machine learning models: Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). We performed comprehensive data preprocessing, exploratory data analysis, and model training using the Dry Bean Dataset from the UCI Machine Learning Repository, which contains morphological properties of different types of beans. Thirty percent of the data were used for testing and seventy percent for training the models. According to the results, the SVM model had the highest accuracy (93%), followed by KNN (91.9%) and Random Forest (92.1%). These results highlight the need of choosing the right models for agricultural data by showcasing SVM's superior classification performance in this situation.

2. Introduction

Accurate crop variety classification is essential for effective resource management and quality control in modern agriculture. Dry beans are a staple crop in many parts of the world, but their physical attributes vary greatly, which makes human classification difficult and prone to error. A reliable answer is provided by machine learning (ML), which automates the categorization process with excellent consistency and precision.

This study investigates the classification of dry bean varieties using three machine learning algorithms: Random Forest[2], K-Nearest Neighbors (KNN)[1][2][3], and Support Vector Machine (SVM)[4]. The project seeks to assess and compare the accuracy performance of these models using the UCI Machine Learning Repository's Dry Bean Dataset, which contains morphological traits of numerous dry bean species.

Every algorithm offers a unique set of benefits. SVM[4] is renowned for its efficiency in high-dimensional spaces, KNN [1][2][3]is appreciated for its readability and simplicity, and Random Forest[2]is commended for its capacity to manage huge datasets and reduce overfitting. We intend to ascertain which technique offers the most dependable classification accuracy for this application by training these models on the dry bean dataset. The study's findings will add to the expanding corpus of knowledge on machine learning applications in agriculture by highlighting the potential of these tools to improve agricultural quality control and productivity.

3. Experimentation

The thorough method used to categorize dry bean varieties using Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) machine learning models is described in this part. The Dry Bean Dataset, which covers a variety of dry bean morphological characteristics, was acquired from the UCI Machine Learning Repository and used for this investigation. The procedures include importing data, cleaning data, performing exploratory data analysis (EDA), and training and assessing machine learning models.

Importing the Dry Bean Dataset into the workspace was the first step. Several features that are necessary for the classification task are present in this dataset. Not only was the dataset imported, but also the required libraries, including scikit-learn for machine learning, pandas, numpy, matplotlib, seaborn for EDA, and seaborn for EDA.

To gain a deeper understanding of the dataset, EDA was used. This involved looking at correlations between the different features and showing their distributions. Table 1 a list of the features of the dataset:

Table 1 : Features of dataset

Features	Description
Area(A)	The area of a bean zone and the number of pixels within its boundaries.
Perimeter (P)	Bean circumference is defined as the length of its border.
Major axis length (L)	The distance between the ends of the longest line that can be drawn from a bean.
Minor axis length (I)	The longest line that can be drawn from the bean while standing perpendicular to the main axis.
Aspect ratio (K)	Defines the relationship between L and l.
Eccentricity (Ec)	Eccentricity of the ellipse having the same moments as the region.
Convex area (C)	Number of pixels in the smallest convex polygon that can contain the area of a bean seed.
Equivalent diameter	The diameter of a circle having the same area as a bean seed area.
Extent (Ex)	The ratio of the pixels in the bounding box to the bean area.
Solidity (S)	Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.
Roundness (R)	Calculated with the following formula: (4piA)/(P^2).
Compactness (CO)	Measures the roundness of an object: Ed/L.
ShapeFactor1 (SF1)	First shape factor.
ShapeFactor2 (SF2)	Second shape factor.
ShapeFactor3 (SF3)	Third shape factor.
ShapeFactor4 (SF4)	Fourth shape factor.
Class	Bean variety (Seker, Barbunya, Bombay, Cali, Dermosan, Horoz, and Sira).

Data wrangling and cleaning were required after the EDA was finished to make sure the dataset was appropriate for machine learning applications. This required looking for duplicates and missing values. Thankfully, there were no missing values in the dataset. To keep the models from being redundant and from having any biases, duplicate rows and columns were found and eliminated. After the dataset had been cleaned, it was divided into training and testing sets. To make sure the models had enough data to learn from and were properly tested on unseen data, a ratio of 70% training data to 30% testing data was used. The scikit-learn train_test_split function was used to implement this split.

Training and assessing the three machine learning models—KNN, Random Forest, and SVM—formed the central focus of the investigation.

4. Results

Based on how well the three machine learning models—Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM)—classified the dry bean types, their performance was assessed. The difference in their respective accuracies is shown in Figure 1. The dataset offered a solid foundation for this assessment, with 70% of the data being training and 30% being testing.

With an accuracy of 91.9%, the KNN model demonstrated a high degree of precision in its classification of the various dry bean varieties. The simplicity and efficacy of this model's use of the nearest neighbors for categorization accounts for its performance. Its accuracy was marginally less than that of the SVM model, nevertheless.

With an accuracy of 92.1%, the Random Forest model—an ensemble learning technique—was produced. By aggregating the predictions of numerous decision trees, the model is able to tolerate fluctuations in the dataset, as evidenced by its somewhat superior performance when compared to KNN.

Compared to the other two models, the SVM model fared better, attaining the maximum accuracy of 93%. SVM performed better because it could locate the best hyperplane for class separation even in high-dimensional spaces.

In conclusion, all three models showed good classification abilities; however, for this particular dataset, the SVM model worked best shown in Figure 2, closely followed by Random Forest and KNN. These results highlight how crucial model selection is to getting the best possible classification accuracy for datasets related to agriculture.

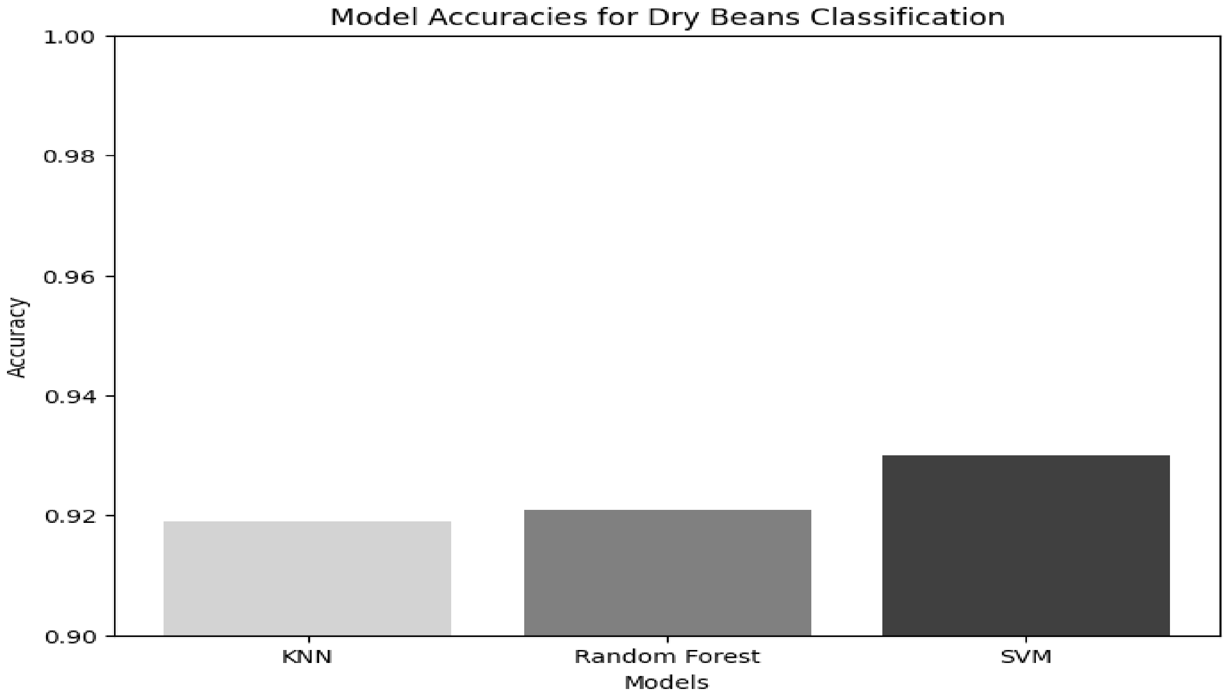

Figure 1: Model accuracies for dry beans classification

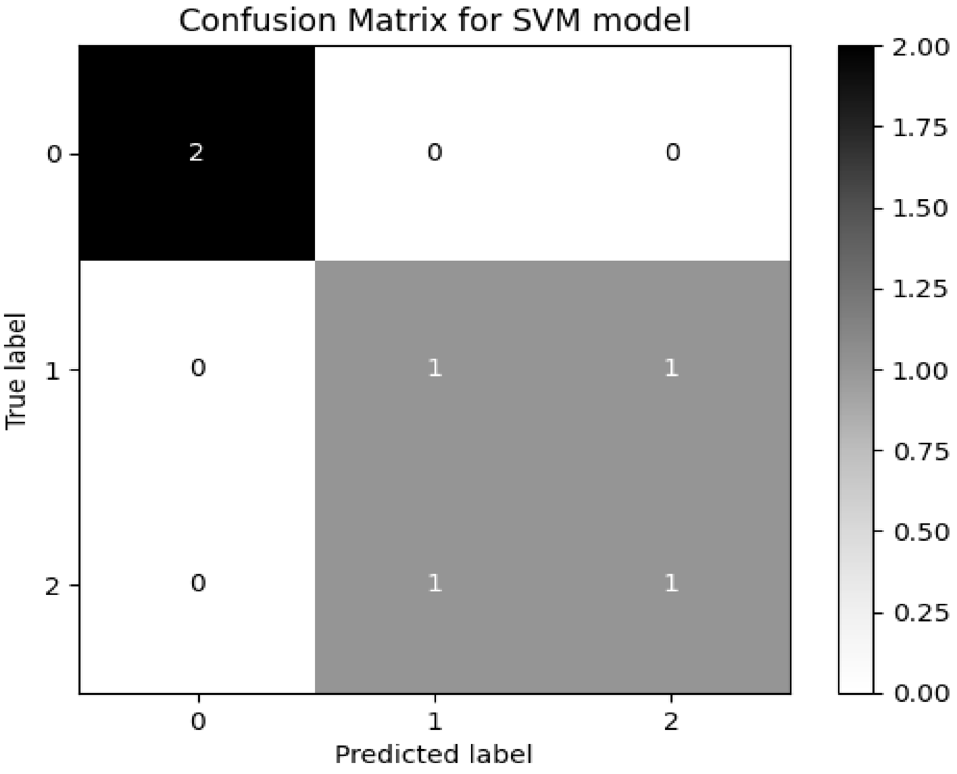

Figure 2: Confusion matrix for SVM model

Acknowledgements

I the author, would like to express my sincere gratitude to my teacher, Nan Xiao Fei (南晓斐), for providing me with the invaluable guidance and support throughout this project. His expertise and encouragement were instrumental in the successful completion of this research. I am also thankful for the opportunity to explore and apply machine learning techniques in the field of agricultural classification.

References

[1] M. V. Subbarao, J. T. S. Sindhu, Y. C. A. Padmanabha Reddy, V. Ravuri, K. P. Vasavi and G. C. Ram, "Performance Analysis of Feature Selection Algorithms in the Classification of Dry Beans using KNN and Neural Networks," 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2023.

[2] Jaime Carlos Macuácua, Jorge António Silva Centeno, Caísse Amisse, Data mining approach for dry bean seeds classification, Smart Agricultural Technology, Volume 5, 2023, 100240, ISSN 2772-3755

[3] Murat Koklu, Ilker Ali Ozkan, Multiclass classification of dry beans using computer vision and machine learning techniques,Computers and Electronics in Agriculture, Volume 174, 2020, 105507, ISSN 0168-1699

[4] Taspinar, Y.S., Dogan, M., Cinar, I. et al. Computer vision classification of dry beans (Phaseolus vulgaris L.) based on deep transfer learning techniques. Eur Food Res Technol 248,2707-2725 (2022).