Project for CS 677

Amina Bauyrzhan

Dataset: <https://archive.ics.uci.edu/dataset/545/rice+cammeo+and+osmancik>

From the repository: Dataset Information - Among the certified rice grown in TURKEY, the Osmancik species, which has a large planting area since 1997 and the Cammeo species grown since 2014 have been selected for the study. When looking at the general characteristics of Osmancik species, they have a wide, long, glassy and dull appearance. When looking at the general characteristics of the Cammeo species, they have wide and long, glassy and dull in appearance. A total of 3810 rice grain's images were taken for the two species, processed and feature inferences were made. 7 morphological features were obtained for each grain of rice.

Table 1. Variables Description

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Role** | **Type** | **Description** | **Missing Values** |
| Area | Feature | Integer | Returns the number of pixels within the boundaries of the rice grain | no |
| Perimeter | Feature | Continuous | Calculates the circumference by calculating the distance between pixels around the boundaries of the rice grain | no |
| Major\_Axis\_Length | Feature | Continuous | The longest line that can be drawn on the rice grain, i.e. the main axis distance, gives | no |
| Minor\_Axis\_Length | Feature | Continuous | The shortest line that can be drawn on the rice grain, i.e. the small axis distance, gives | no |
| Eccentricity | Feature | Continuous | It measures how round the ellipse, which has the same moments as the rice grain, is | no |
| Convex\_Area | Feature | Integer | Returns the pixel count of the smallest convex shell of the region formed by the rice grain | no |
| Extent | Feature | Continuous | Returns the ratio of the region formed by the rice grain to the bounding box | no |
| Class | Target | Binary | Cammeo and Osmancik | no |

An exploratory data analysis was conducted, including statistical summaries of the features.

Table 2. Statistical Summary of variables

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **0** | **1** | **Overall** |
| **Area** | **mean** | 14162,89 | 11549,78 | 12667,73 |
| **std** | 1286,77 | 1041,91 | 1732,37 |
| **Perimeter** | **mean** | 487,44 | 429,42 | 454,24 |
| **std** | 22,18 | 20,15 | 35,6 |
| **Major\_Axis\_Length** | **mean** | 205,48 | 176,29 | 188,78 |
| **std** | 10,33 | 9,36 | 17,45 |
| **Minor\_Axis\_Length** | **mean** | 88,77 | 84,48 | 86,31 |
| **std** | 5,35 | 5,3 | 5,73 |
| **Eccentricity** | **mean** | 0,9 | 0,88 | 0,89 |
| **std** | 0,01 | 0,02 | 0,02 |
| **Convex\_Area** | **mean** | 14494,43 | 11799,59 | 12952,5 |
| **std** | 1309,42 | 1062,8 | 1776,97 |
| **Extent** | **mean** | 0,65 | 0,67 | 0,66 |
|  | **std** | 0,08 | 0,07 | 0,08 |

Also pairplot comparisons separately for each class was performed to understand the distributions and relationships of the features within each class of rice.

Table 3. Pairplots for Cammeo

A chart of a diagram

Description automatically generated with medium confidence

Table 4. Pairplots for Osmancik

A chart of a diagram

Description automatically generated with medium confidence

Training set size: (3048, 7)

Testing set size: (762, 7)

Model Training and Evaluation

Multiple machine learning models were employed to classify the rice varieties based on their features. The dataset was divided into training (80%) and testing (20%) sets using a random seed for reproducibility. The models trained included:

K-Nearest Neighbors (KNN): Several KNN models with different values of k (number of neighbors) were evaluated to determine the optimal configuration based on accuracy. A feature importance analysis was also conducted by iteratively removing each feature and assessing the impact on model accuracy.

Table 5. Accuracy vs k in KNN classifier

A graph with a line and a point

Description automatically generated

Scenario: Just Area is missing, Accuracy: 0.92

Scenario: Just Perimeter is missing, Accuracy: 0.88

Scenario: Just Major\_Axis\_Length is missing, Accuracy: 0.88

Scenario: Just Minor\_Axis\_Length is missing, Accuracy: 0.89

Scenario: Just Eccentricity is missing, Accuracy: 0.89

Scenario: Just Convex\_Area is missing, Accuracy: 0.91

Scenario: Just Extent is missing, Accuracy: 0.89

As a result Area was dropped

Logistic Regression: This model was applied to evaluate its performance in terms of accuracy and the ability to linearly separate the classes.

Support Vector Machine (SVM) with a Linear Kernel: Used to determine if a linear decision boundary is sufficient for class separation.

Gaussian Naive Bayes: Applied due to its efficacy with high-dimensional data

Decision Tree and Random Forest: These models were chosen for their ability to handle nonlinear relationships.

The Random Forest model was further tuned by experimenting with different numbers of trees and depths to minimize the classification error.

For some of the models, particularly those sensitive to the scale of data like Logistic Regression and SVM, feature scaling was performed using the StandardScaler to standardize the features to have zero mean and unit variance.

Table 6. Random Forest: Error rates for different values of N and d



Results.

The following table summarizes the performance metrics of each classification models used to differentiate between two rice varieties, Cammeo and Osmancik. The results include the number of true positives (TP), false positives (FP), true negatives (TN), false negatives (FN), along with the true positive rate (TPR), true negative rate (TNR), and overall accuracy of each model.

* Random Forest achieved the highest accuracy (93%) with excellent rates for correctly identifying both varieties.
* Linear SVM also performed very well, with a slightly lower accuracy (92.9%) but the best ability to distinguish non-Cammeo varieties.
* Logistic Regression was effective, with an accuracy of 92.7%.
* KNN and Naive Bayes were close in performance, with accuracies just over 91%.
* Decision Tree lagged behind the other models at 89.5% accuracy.

Random Forest and Linear SVM were the top performers, indicating more sophisticated models were better suited for this classification task

Table . Results of the models

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **TP** | **FP** | **TN** | **FN** | **TPR** | **TNR** | **Accuracy** |
| **KNN** | 397 | 39 | 301 | 25 | 0,94 | 0,89 | 0,916 |
| **Logistic Regression** | 396 | 30 | 310 | 26 | 0,94 | 0,91 | 0,927 |
| **Naive Bayes** | 394 | 37 | 303 | 28 | 0,93 | 0,89 | 0,915 |
| **Decision Tree** | 385 | 43 | 297 | 37 | 0,91 | 0,87 | 0,895 |
| **Random Forest** | 401 | 32 | 308 | 21 | 0,95 | 0,91 | 0,930 |
| **Linear SVM** | 396 | 28 | 312 | 26 | 0,94 | 0,92 | 0,929 |