Term End Milestone-1 (Project -1)

IDENTIFICATION OF RICE VARIETIES

Proposal & Data Selection

DSC, Bellevue University

Prashant Raghuwanshi

DSC680-T301 Applied Data Science (2225-1)

Professor Catie Williams

18/03/2022

ABSTRACT:

This term-end project-1 aims to evaluate the Students should be able to identify a business problem to address through predictive analytics. The goal is to select appropriate models and model specifications and apply the respective methods to enhance data-driven decision-making related to the business problem. Students will identify the potential use of predictive analytics, formulate the problem, identify the right sources of data, analyze data and prescribe actions to improve not only the process of decision making but also the outcome of decisions.

In this project, we are using datasets that contain, five different varieties of rice belonging to the same trademark were selected to carry out classification operations using morphological, shape, and color features. A total of 75 thousand rice grain images, including 15 thousand for each variety, were obtained. The images were pre-processed using MATLAB software and prepared for feature extraction. Using a combination of 12 morphological, 4 shape features, and 90 color features obtained from five different color spaces, a total of 106 features were extracted from the images.

For classification, models were created with algorithms using machine learning techniques of k-nearest neighbor, decision tree, logistic regression, multilayer perceptron, random forest, and support vector machines. With these models, performance measurement values were obtained for feature sets of 12, 16, 90, and 106. Among the models, the success of the algorithms with the highest average classification accuracy was achieved 97.99% with random forest for morphological features. 98.04% were obtained with random forest for morphological and shape features. It was achieved with logistic regression as 99.25% for color features. Finally, 99.91% was obtained with multilayer perceptron for morphological, shape, and color features. When

the results are examined, it is observed that with the addition of each new feature, the success of classification increases.

Based on the performance measurement values obtained, it is possible to say that the study achieved success in classifying rice varieties.

Topic :

Identification of Rice Varieties

Towards a real-time sorting system: Identification of vitreous durum rice kernels using ANN based on their morphological, color, wavelet, and gaborlet features.

Business Problem :

The modern Food Processing industry is importing grains (rice) from various international grains distributors. Their producted packed foods mostly depend on quality and varieties of imported raw gains. Even though placing the same type of rice variety order by food processing companies, most of the time Multiple gains distributors supplies the adulterated mix with the ordered rice varieties, and it results in causing the quality degradation and inconsistent taste of packed food and at last, it impacts the sales of processed food product in the competitive market place.

At present food processing companies are using random sampling and manual grain monitoring techniques to make sure the procured variety of rice is good. However, this technique is not giving consistent results, since it depends on the individual human eye for identifying the quality of the grain from a single sample. Most of the time due to lack of expertise human eyes are not able to detect the ambiguities in a sample.

Requirements, assumptions:

Requirements: This Project is trying to make use of machine learning techniques to automatically identify the variety of the rice in the given rice sample. Here I am planning to feed the data for the rice to the ML model and the ML model will detect the rice sample and send a signal to the grain sampling machine's sensors to pick out the ambiguous rice variety from the sample.

Assumption: Here I am assuming the Preprocessing operations were applied to the rice images was applied successfully and made available data for feature extraction is not having any issue.

Costs and benefits:

Costs:

- The primary cost associated with this project is the time of the people working on it.
- Computing resources for modeling
- Data collection and processing computing costs

Benefits:

- Social benefit of this model is helping the food processing companies to automatically detect the variety of rice in the provided sample of rice and helping the authority to stop low-cost mixing and adulteration practices of grain traders
- Financial benefit to the company from the ability to maintain the brand quality of processed food which results in increasing the brand loyalty for consumers and its sales.

What Questions Are We Trying To Answer? :

- Which are the main predictor's features existing in the given gain MSC sample records?
- Which Model is going to fit our use case?
- Which is the best target variable for our model?
- What is the accuracy of the model?
- Do we got any other interesting facts from datasets? Like correlations etc
- Is it possible, the ML technique can identify the grain varieties?

Datasets:

A total of 75 thousand pieces of rice grain were obtained, including 15 thousand pieces of each variety of rice (Arborio, Basmati, Ipsala, Jasmine, Karacadag). Preprocessing operations were applied to the images and made available for feature extraction. A total of 106 features were inferred from the images; 12 morphological features and 4 shape features were obtained using morphological features and 90 color features were obtained from five different color spaces (RGB, HSV, Lab*, YCbCr, XYZ).

Morphological Feature Extraction : After the processing phase of the image, therice grains on each image were treated separately and a number of features were inferred. Feature extractions have been studied from morphological point of view. In total, 7 morphological features were inferred for each grain.

Morphological feature extraction is a wide range of image processing processes that process based on the shapes found on the image. In this process, each pixel in the image is adjusted according to the value of the other pixels around it.

Describe Data.

Attribute Information:

1.) Area: Returns the number of pixels within the boundaries of the rice grain.

2.) Perimeter: Calculates the circumference by calculating the distance between pixels around the boundaries of the rice grain.

3.) Major Axis Length: The longest line that can be drawn on the rice grain, i.e. the main axis distance, gives.

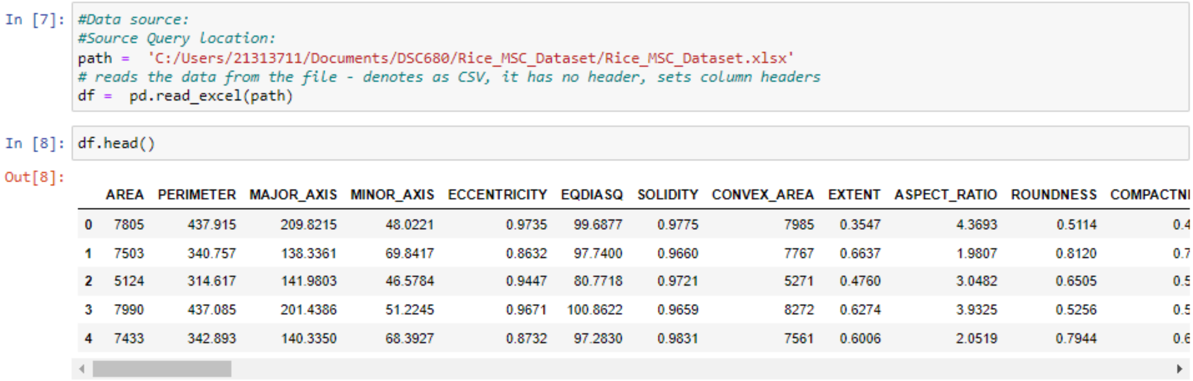
4.) Minor Axis Length: The shortest line that can be drawn on the rice grain, i.e. the small axis distance, gives.

5.) Eccentricity: It measures how round the ellipse, which has the same moments as the rice grain, is.

6.) Convex Area: Returns the pixel count of the smallest convex shell of the region formed by the rice grain.

7.) Extent: Returns the ratio of the regionformed by the rice grain to the bounding box pixels.

8.) Class: Cammeo and Osmancik rices



In [7]:

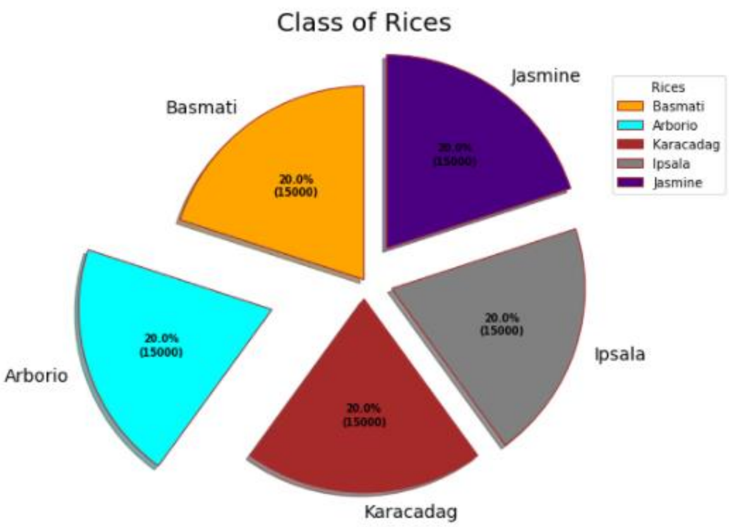
```
#Data source:
#Source Query location:
path =  'C:/Users/21313711/Documents/DSC680/Rice_MSC_Dataset/Rice_MSC_Dataset.xlsx'
# reads the data from the file - denotes as CSV, it has no header, sets column headers
df =  pd.read_excel(path)
```

In [8]:

```
df.head()
```

Out[8]:

	AREA	PERIMETER	MAJOR_AXIS	MINOR_AXIS	ECCENTRICITY	EQDIASQ	SOLIDITY	CONVEX_AREA	EXTENT	ASPECT_RATIO	ROUNDNESS	COMPACTNI
0	7805	437.915	209.8215	48.0221	0.9735	99.6877	0.9775	7985	0.3547	4.3693	0.5114	0.4
1	7503	340.757	138.3361	69.8417	0.8632	97.7400	0.9660	7767	0.6637	1.9807	0.8120	0.7
2	5124	314.617	141.9803	46.5784	0.9447	80.7718	0.9721	5271	0.4760	3.0482	0.6505	0.5
3	7990	437.085	201.4386	51.2245	0.9671	100.8622	0.9659	8272	0.6274	3.9325	0.5256	0.5
4	7433	342.893	140.3350	68.3927	0.8732	97.2830	0.9831	7561	0.6006	2.0519	0.7944	0.6



Methods:

Classification models are a method of high importance used in various fields. In class determination, classification models are used to determine which class the data belongs to. The classification model is a model that works by making predictions. The purpose of the classification is to make use of the common characteristics of the data to parse the data in question.

At part from regular classification model I am planning to use some more additional models as mentioned below.

Models were going to create by using Artificial Neural Network (ANN) and Deep Neural Network (DNN) algorithms for the feature dataset and by using the Convolutional Neural Network (CNN) algorithm for the image dataset, and classification processes were performed. Statistical results of sensitivity, specificity, prediction, F1 score, accuracy, false positive rate and false negative rate were calculated using the confusion matrix values of the models and the results of each model were going to record in a table.

Ethical Considerations:

- This Data contains processed physical images information related to multiple verities of rices and not contains any PII related information.
- Datasets and information on data was extracted from the public websites → UCL machine learning repositories.
- This data research is not going to harm any privacy.

Challenges/Issues :

Constraints: Major Constraints are related to used datasets and processed images, here the used datasets contain a total of 75 thousand rice grain images, including 15 thousand for each variety however due to the rapidly advancing Seed development process, we might not have all full collections of grain records under each gain varieties.

References :

https://archive.ics.uci.edu/ml/datasets/Rice+%28Cammeo+and+Osmancik%29

https://www.muratkoklu.com/datasets/

1: KOKLU, M., CINAR, I. and TASPINAR, Y. S. (2021). Classification of rice varieties with deep learning methods. Computers and Electronics in Agriculture, 187, 106285.

DOI: https://doi.org/10.1016/j.compag.2021.106285

2: CINAR, I. and KOKLU, M. (2021). Determination of Effective and Specific Physical Features of Rice Varieties by Computer Vision In Exterior Quality Inspection. Selcuk Journal of Agriculture and Food Sciences, 35(3), 229-243.

DOI: https://doi.org/10.15316/SJAFS.2021.252

3: CINAR, I. and KOKLU, M. (2022). Identification of Rice Varieties Using Machine Learning Algorithms. Journal of Agricultural Sciences.

DOI: https://doi.org/10.15832/ankutbd.862482

4: CINAR, I. and KOKLU, M. (2019). Classification of Rice Varieties Using Artificial Intelligence Methods. International Journal of Intelligent Systems and Applications in Engineering, 7(3), 188-194.

DOI: https://doi.org/10.18201/ijisae.2019355381