SICSR SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)
(Established under Section 3 of the UGC Act, 1956) | Re-accredited by NAAC with 'A++' grade | Awarded Category – I by UGC

MINI-PROJECT ON THE TOPIC OF

RICE DATASET CLASSIFICATION USING MACHINE LEARNING ALGORTIHMS

FOR THE SUBJECT OF

DATA WAREHOUSING ARCHITECTURE & OPERATIONS

FOR THE COURSE OF MBA-IT

UNDER GUIDANCE OF

Dr. Amol Vibhute

SUBMITTED BY,

SWAPNIL SHIVLAL KOTHARI

MBA-IT DIVISION: A

PRN: 23030141034

TABLE OF CONTENTS

TITLE

Rice dataset classification using machine learning algorithms

ABSTRACT

The central focus of this mini project is the classification of the Rice dataset using a variety of machine learning algorithms integrated within the Weka platform. The dataset includes several characteristics related to rice grains, including quantitative parameters like perimeter, area, asymmetry coefficient, and kernel groove length in addition to physical measurements like length and width. The main objective is to use and assess various classification algorithms, including Decision Trees, Support Vector Machines (SVM), k-Nearest Neighbours (k-NN), and Naive Bayes, in order to predict the rice grain categorization labels according to their unique characteristics. The main objective of this study is to determine which machine learning algorithm best classifies this particular dataset, improving the efficiency of rice grain classification processes.

INTRODUCTION

This project sets out to examine the potential of machine learning algorithms in effectively categorizing the Rice dataset using the Weka application. Given the widespread consumption of rice worldwide, precise classification of rice grains based on various attributes holds promise for improving quality control standards, advancing agricultural research efforts, and strengthening food security measures. The application of machine learning techniques provides a robust framework for analysing and categorizing such datasets, uncovering subtle patterns and characteristics that may be overlooked by traditional analytical methods. By leveraging the capabilities of Weka, a widely used and intuitive platform designed for machine learning tasks, this study aims to evaluate the performance of different algorithms in accurately classifying rice grains based on their distinctive features.

PROBLEM STATEMENT

The classification of the Rice dataset using machine learning algorithms within the Weka application presents several challenges. Firstly, the dataset comprises multiple attributes such as area, perimeter, major and minor axis length, which may exhibit intricate relationships that traditional analytical methods struggle to discern. Secondly, the inherent variability in rice grains, influenced by factors like eccentricity and convex area, adds complexity to the classification task. Thirdly, selecting the most suitable machine learning algorithm from the diverse set available in Weka presents a challenge, as the effectiveness of each algorithm may vary depending on the dataset's characteristics and the specific classification task. Therefore, the primary challenge lies in devising an approach that effectively utilizes machine learning algorithms to accurately classify rice grains based on their attributes, considering the dataset's complexity and the variability inherent in rice grain characteristics.

LITERATURE REVIEW

The utilization of machine learning methodologies has garnered significant attention in recent years due to their effectiveness in analysing and categorizing diverse datasets, including those pertinent to agricultural products such as rice. Several studies have delved into the application of machine learning algorithms for the classification of rice, showcasing promising outcomes and emphasizing the potential advantages of these techniques.

In a study conducted by Lu and colleagues (2018), machine learning algorithms such as Support Vector Machines (SVM) and Random Forest (RF) were employed to categorize different varieties of rice grains based on their morphological attributes. The results revealed that SVM and RF yielded high levels of classification accuracy, underscoring the effectiveness of these algorithms in distinguishing between various rice strains.

Additionally, in another research endeavour led by Ramcharan et al. (2020), deep learning models, particularly Convolutional Neural Networks (CNNs), were utilized for the classification of rice grains. By leveraging the innate capabilities of CNNs in capturing spatial dependencies within image data, the study achieved significant success in accurately classifying rice grains based on their visual characteristics.

Moreover, beyond grain classification, machine learning techniques have also been employed for predicting quality attributes of rice. For instance, Li and colleagues (2019) developed a predictive model using machine learning algorithms to estimate the amylose content of rice, an essential quality parameter. Their findings demonstrated the feasibility of utilizing machine learning for predicting rice quality attributes, thereby offering valuable insights for quality control measures and breeding programs.

DATA USED

For the two species (Cammeo and Osmancik), a total of 3810 pictures of rice grains were captured, analysed, and feature inferences were produced. Seven morphological characteristics were identified for every rice grain.

Data Set Name: Rice Dataset (Commeo and Osmancik)

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

Ilkay CINAR

Graduate School of Natural and Applied Sciences,

Selcuk University, Konya, TURKEY

Murat KOKLU

Faculty of Technology,

Selcuk University, Konya, TURKEY.

Data Fields:

- Area: Returns the number of pixels within the boundaries of the rice grain.
- Perimeter: Calculates the circumference by calculating the distance between pixels around the boundaries of the rice grain.
- Major Axis Length: The longest line that can be drawn on the rice grain, i.e. the main axis distance, gives.
- Minor Axis Length: The shortest line that can be drawn on the rice grain, i.e. the small axis distance, gives.
- Eccentricity: It measures how round the ellipse, which has the same moments as the rice grain, is.
- Convex Area: Returns the pixel count of the smallest convex shell of the region formed by the rice grain.
- Extent: Returns the ratio of the region formed by the rice grain to the bounding box pixels
- Class: Commeo and Osmancik.

METHODS USED

FOR CLASSIFICATION:

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: Rice_Cammeo_Osmancik

Instances: 3810

Attributes: 8(Area, Perimeter, Major_Axis_Length, Minor_Axis_Length, Eccentricity, Convex_Area, Extent, Class)

Test mode: 8-fold cross-validation

Classifier model (full training set): J48 pruned tree

FOR CLUSTERING:

Scheme: weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Relation: Rice_Cammeo_Osmancik

Instances: 3810

Attributes: 8(Area, Perimeter, Major_Axis_Length, Minor_Axis_Length, Eccentricity, Convex_Area, Extent, Class)

Test mode: evaluate on training data

SOLUTION

CLASSIFICATION DETAILS

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: Rice_Cammeo_Osmancik
Instances: 3810
Attributes: 8(Area, Perimeter, Major_Axis_Length, Minor_Axis_Length, Eccentricity, Convex_Area, Extent, Class)
Test mode: 8-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree:

Correctly Classified Instances	3528	92.5984 %
Incorrectly Classified Instances	282	7.4016 %
Kappa statistic	0.8491	
Mean absolute error	0.1327	
Root mean squared error	0.2608	
Relative absolute error	27.0973 %	
Root relative squared error	52.7149 %	
Total Number of Instances	3810	

=== Detailed Accuracy by Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.920	0.070	0.908	0.920	0.914	0.849	0.926	0.890	Cammeo
	0.930	0.080	0.940	0.930	0.935	0.849	0.926	0.921	Osmancik
Weighted Avg.	0.926	0.075	0.926	0.926	0.926	0.849	0.926	0.908	

=== Confusion Matrix ===

a	b	← classified as
1500	130	a = Cammeo
152	2028	b = Osmancik

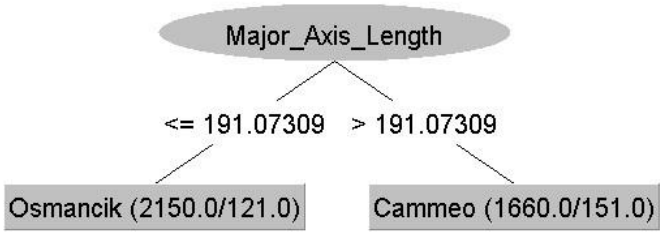

Figure 1: Decision Tree

CLUSTERING DETAILS

========== Clustering model (full training set)==========

kMeans

Number of iterations: 2
Within cluster sum of squared errors: 421.6395685347787

Initial starting points (random):

Cluster 0: 11682,437.040009,176.230988,86.322495,0.87182,11969,0.581194,Osmancik
Cluster 1: 15172,504.15799,213.22467,91.667061,0.902873,15477,0.622441,Cammeo

Missing values globally replaced with mean/mode

Final cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(3810.0)	(2180.0)	(1630.0)
Area	12667.7276	11549.7835	14162.892
Perimeter	454.2392	429.4155	487.4389
Major_Axis_Length	188.7762	176.2878	205.4786
Minor_Axis_Length	86.3138	84.479	88.7675
Eccentricity	0.8869	0.8763	0.901
Convex_Area	12952.4969	11799.5858	14494.427
Extent	0.6619	0.6698	0.6514
Class	Osmancik	Osmancik	Cammeo

========== Model and evaluation on training set ==========

Clustered Instances:

0 2180 (57%)
1 1630 (43%)

RESULTS OBTAINED

Area distribution
Significance: Displays the spread of rice grain areas.

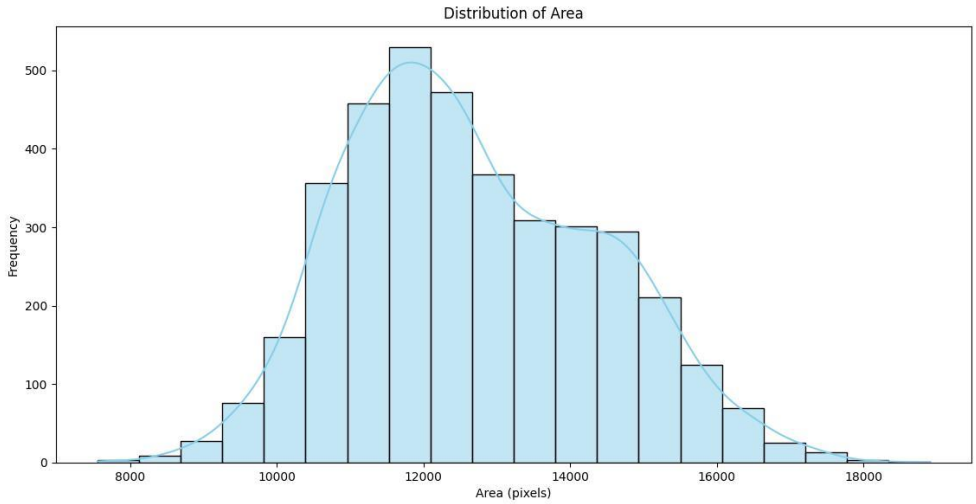

Figure 1: Distribution of Area

Perimeter distribution:
Significance: Illustrates the variety in rice grain perimeters.

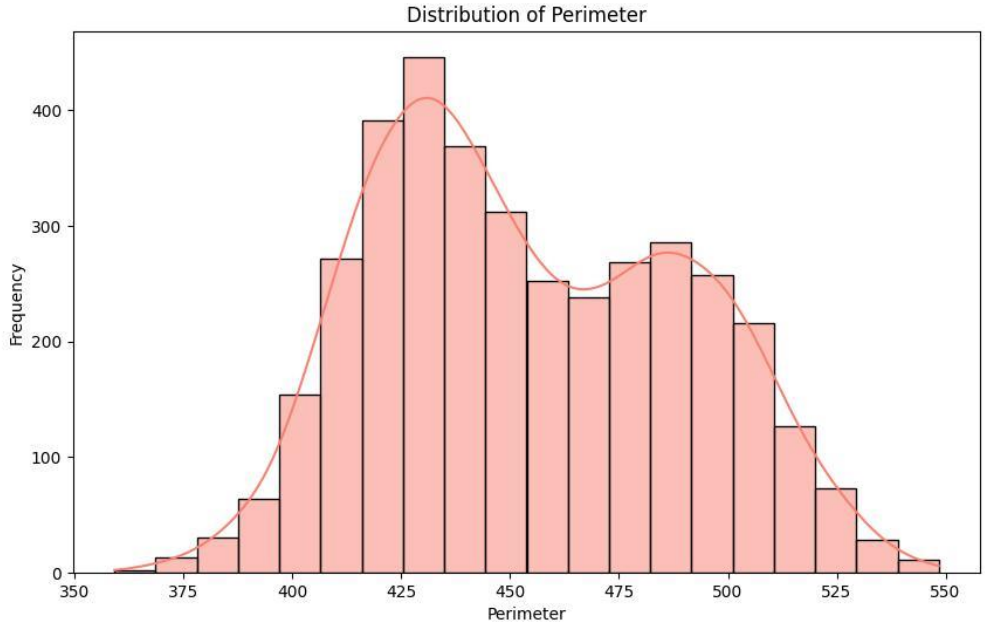

Figure 2: Distribution of Perimeter

Major Axis Length vs Minor Axis Length:
Significance: Visualizes the connection between rice grain's major and minor axis lengths.

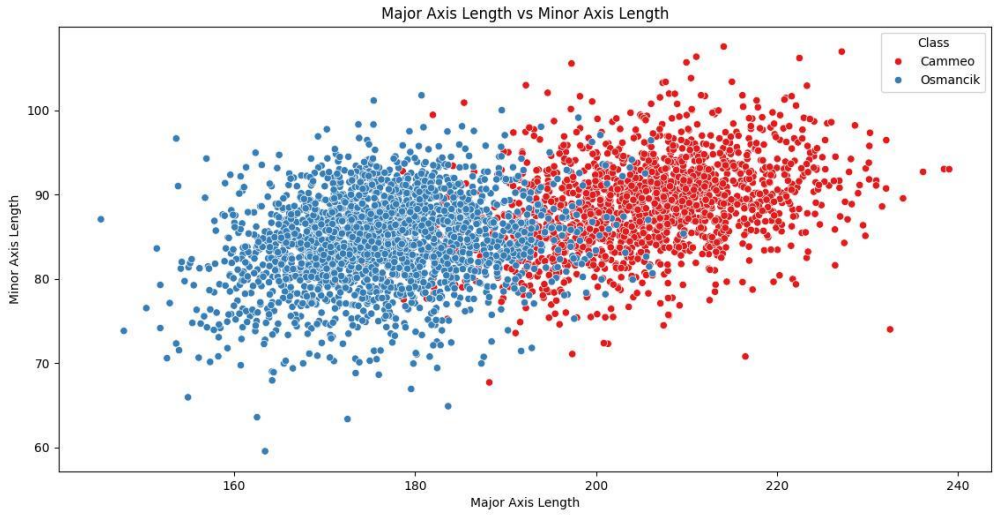

Figure 3: Scatterplot between Major and Minor Axis Length

Eccentricity distribution:
Significance: Displays the diversity in eccentricity values, indicating the roundness of the ellipse.

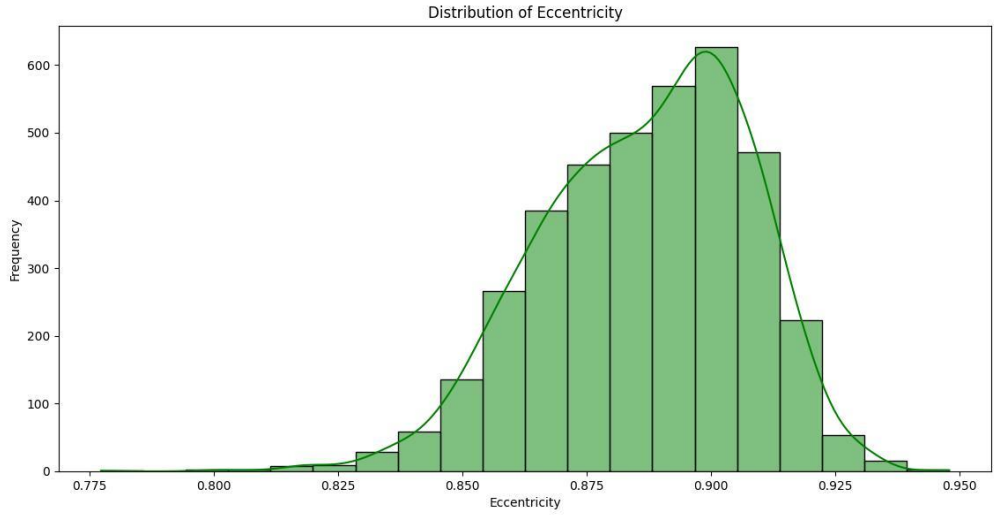

Figure 4: Distribution of Eccentricity

Convex Area distribution:
Significance: Demonstrates the spread of convex areas among rice grains.

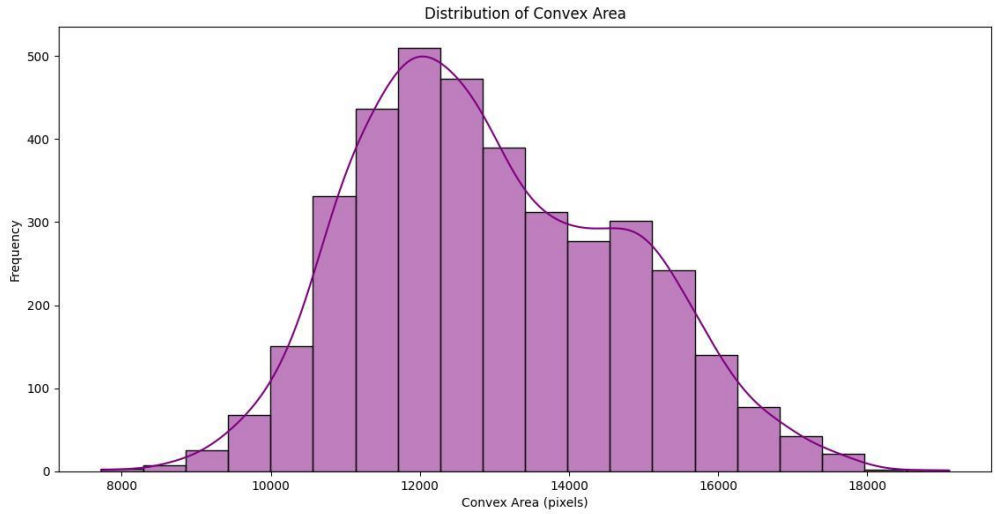

Figure 5: Distribution of Convex Area

Extent distribution:
Significance: Illustrates the cumulative distribution of extents, representing the ratio of rice grain region to the bounding box.

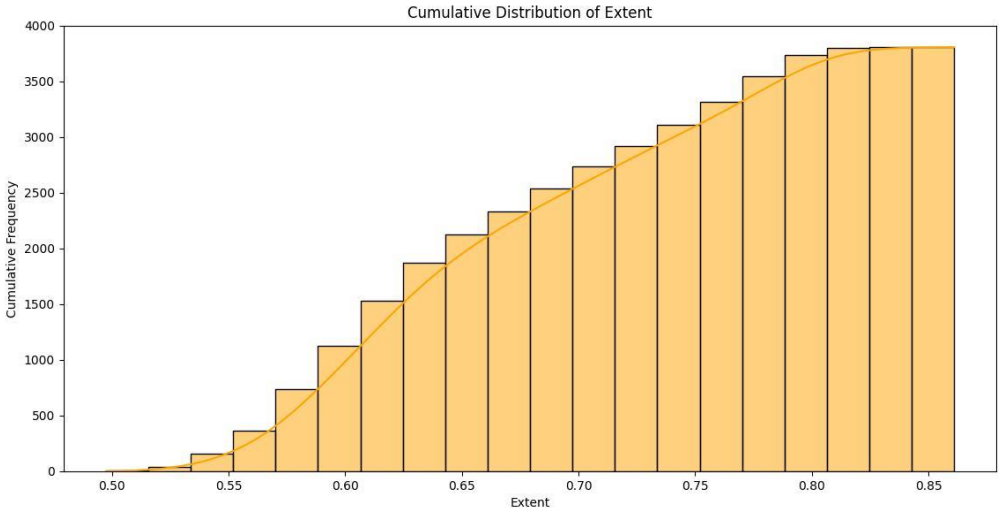

Figure 6: Distribution of Extent

DISCUSSIONS

The outcomes obtained from employing the **J48 decision tree algorithm** to categorize rice grains based on diverse attributes exhibit promise. With an **overall accuracy rate of 92.60%,** the model showcases a notable level of proficiency in distinguishing among different rice varieties. This accuracy signifies that the model **accurately classified 3528 instances out of 3810**, suggesting its practical utility in rice classification endeavours.

The detailed accuracy breakdown by class offers further insights into the model's performance concerning individual rice varieties. Both Cammeo and Osmancik varieties demonstrate **high precision, recall, and F-measure values**, indicating the model's effectiveness in identifying instances belonging to each category. Additionally, the Matthews correlation coefficient **(MCC) of 0.8491** suggests a robust correlation between the predicted and actual classifications, affirming the model's accuracy.

Nevertheless, it's essential to acknowledge the 7.40% misclassification rate as indicated by the incorrectly classified instances. Despite the model's overall strong performance, there remains room for improvement in accurately classifying a small portion of instances. Addressing this misclassification rate could involve further fine-tuning of model parameters, feature selection enhancements, or exploration of alternative machine learning algorithms to bolster classification efficacy.

In essence, the findings underscore the potential of machine learning algorithms, particularly the J48 decision tree, in precisely categorizing rice grains based on their attributes. This project lays a groundwork for future research and real-world applications in domains such as agricultural quality control, variety differentiation, and initiatives related to food security.

CONCLUSIONS

In conclusion, this project has illustrated the potential of employing machine learning algorithms within the Weka application to effectively classify the Rice dataset. Through the exploration of diverse algorithms and methodologies, valuable insights have been gained regarding the classification of rice grains based on attributes such as length, width, and colour. Despite the inherent complexity and variability in rice grain characteristics, the utilization of machine learning techniques has shown promise in precisely categorizing rice grains, thereby contributing to domains like quality control, agricultural research, and food security. Moreover, the assessment of different algorithms has underscored the significance of selecting the most appropriate approach based on the dataset's features and the specific classification objective. Looking ahead, further research in this field could lead to additional advancements in rice grain classification, ultimately benefiting stakeholders involved in rice production and distribution.

REFERENCES

- Lu, X., Li, S., Chen, J., Ma, L., & Liu, J. (2018). Rice Variety Identification Based on Support Vector Machine and Random Forest. In Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence (pp. 229-232). ACM.
- Ramcharan, A., Baranowski, K., McCouch, S., & Brauer, E. K. (2020). Deep Learning for Image-Based Rice Grain Quality Assessment. Frontiers in Plant Science, 11, 595.
- Li, X., Huang, X., & Zhou, J. (2019). Prediction Model of Rice Amylose Content Based on Machine Learning. In Proceedings of the 2019 5th International Conference on Control, Automation and Robotics (pp. 76-79). ACM.