Literature Review, Data Description, and Approach

Tara Shingadia

500847789

Ceni Babaoglu, Ph.D.

June 13th, 2022

GitHub link: [***https://github.com/Ta2299/CIND-820#cind-820***](https://github.com/Ta2299/CIND-820#cind-820)

Table of Contents

[Project Abstract 3](#_Toc104808296)

[Literature Review 5](#_Toc104808296)

[Introduction 5](file:///C:\Users\Tara\Downloads\Ryerson%20report%20template%20(1).docx#_Toc452457338)

[Related Works 6](file:///C:\Users\Tara\Downloads\Ryerson%20report%20template%20(1).docx#_Toc452457338)

[Methodology 1](file:///C:\Users\Tara\Downloads\Ryerson%20report%20template%20(1).docx#_Toc452457338)1

[Data Description 12](#_Toc104808296)

[Univariate Analysis 1](file:///C:\Users\Tara\Downloads\Ryerson%20report%20template%20(1).docx#_Toc452457338)2

[Data Dictionary 12](#_Toc452457339)

[Data Distribution 13](#_Toc452457339)

[Bivariate Analysis 1](file:///C:\Users\Tara\Downloads\Ryerson%20report%20template%20(1).docx#_Toc452457338)3

[Correlation Analysis 13](#_Toc452457339)

[Pairwise relations 15](#_Toc452457339)

[References 17](#_Toc104808296)

# Project Abstract

There are many different types of rice grown globally, and determining which type of rice a grain falls under is vital as there are many different rice varieties. The dataset chosen will look at rice variants grown in Turkey, specifically the Osmancik and Cammeo variants. Classifying grains manually would not be feasible, so being able to do this using imagery and models would be much more efficient. These images of rice grains were processed, resulting in a dataset with features for each grain of rice and a class variable that classified the grain as either Cammeo or Osmancik varieties. Assuming we have a program that detects the values for the seven attributes, we will be using classification and regression modeling to predict if the grain is Cammeo or Osmancik.

Some of the issues that will be addressed are: which variables impact whether a grain is Cammeo or Osmancik and which ones have the most significant impact? Given the conditions, how likely is a grain to be Cammeo instead of Osmancik?

The data being used is the Rice Dataset (Cammeo and Osmancik) found in the UCI Machine Learning Repository. This dataset looks at the independent variables: Area, Perimeter, Major Axis Length, Minor Axis Length, Eccentricity, Convex Area, and the Extent of each grain of rice. The dataset contains 3810 rows and eight columns, including the class attribute column, the dependent variable. The class attribute is ‘Class’, classifying each instance as either ‘Cammeo’ or ‘Osmancik.’

The techniques that will be used for this analysis include K-fold cross-validation to improve the effectiveness of the models, and given that the class variable contains two classes, multiple logistic regression modeling, as well as Naïve Bayes classification and decision tree classifier, will also be performed. Multiple logistic regression modeling and Naïve Bayes classification will give us the probability of the grain variations given the variables, which can then be used on a new instance. The decision tree classifier will classify grains based on the conditions of the various attributes. After these models are developed, they will be compared using a confusion matrix. The accuracy, precision, sensitivity (recall), specificity, and error rate will be calculated to determine how effective and reliable the models are.

# Literature Review

**Introduction**

Rice is one of the most widely eaten foods worldwide, with global consumption of 509.87 million metric tons in the 2021/2022 crop year (Shahbandeh 2022). With more than 120,000 different varieties of rice in the world, the classification of grains is significant. Being able to distinguish between types is important but manually doing so is inefficient and costly.

With the growth in technology, people and industries adapt to make processes more efficient. Using image processing, grains can be processed and have their features extracted from the images, which can then be used to classify the types of grains based on the features. Using machine learning algorithms on the extracted data would allow large amounts of data to be analyzed quickly and efficiently without having manual labourers identifying grains individually, which can be prone to error and time-consuming. The extracted features of grains of rice are typically geometric parameters such as length, width, perimeter, area, etc.; after pulling the features, classification algorithms are performed. Many different types of classification algorithms have been used to classify agricultural produce, such as LR (Koklu et al., 2021), KNN (Ozkan et al., 2021), CNN (Dheir et al., 2019; Singh et al., 2022), NB (Cinar & Koklu, 2019), ANN (Koklu et al., 2021), MLP (Cinar & Koklu, 2019), DT (Cinar & Koklu, 2019) and more.

**Related works**

Many researchers have completed research on classifying agricultural products based on the variants. Below we look at some research related to the classification of rice, pistachios, nuts, dates, wheat, and apples. All works below have processed images and extracted features to develop a dataset for classification models or used the processed images directly for classification.

In the study by Cinar & Koklu (2019), they obtained 3810 images of Cammeo and Osmancik rice grains and processed them to get seven features. They performed cross-validation using a k-value of 4 with the data divided into a 25% test set and 75% train set. They developed the following models to classify grains as either Cammeo or Osmancik with the following accuracy levels: Logistic Regression (LR) 93.02%, Multiplayer Perceptron (MLP) 92.86%, Support Vector Machine (SVM) 92.83%, Decision Tree (DT) 92.49%, Random Forest (RF) 92.39%, Naïve Bayes (NB) 91.71%, and K-Nearest Neighbor (KNN) 88.58%. This study found that LR had the highest accuracy at 93.02%, and all of the classifiers had an accuracy of 90%+ besides the KNN classifier.

Another study on rice was done by Koklu, Cinar, and Taspinar (2021), where five rice varieties were classified using deep learning methods. The varieties included Arborio, Basmati, Ipsala, Jasmine, and Karacadag. The dataset had 75k images with 15k of each rice type. After processing the photos, the dataset had 12 morphological, four shape, and 90 colour features. K-fold cross-validation was used with k equal to 10. The deep learning methods used were ANN, CNN, and DNN. The average accuracy achieved for each of the methods was 99.87% (ANN), 99.95% (DNN), and 100% (CNN). Thus, concluding the most accurate classifier was the CNN model.

Ibrahim et al. (2019) suggest using a multi-class support vector machine (SVM) for rice grain classification. They classified Basmati, Ponni, and Brown rice using 90 images. Using a multi-class SVM, they could classify more classes than a normal SVM. The One-against-One (OAO) technique was used and achieved an accuracy of 90% for basmati, 93.33% for Ponni and Brown rice, and overall accuracy of 92.22%.

A contrastive study was done by Ibrahim et al. (2020) where rice grains were classified using a multi-class support vector machine (SVM) vs. artificial neural network (ANN). The same data was used in this study as in Ibrahim et al. (2019) with Basmati, Ponni, and Brown rice. The data was partitioned into a train, test, and validation set twice, with the values being 80%, 10%, 10%, and 70%, 15%, and 15%, respectively. For ANN, the first data partitioned set (80/10/10) had an overall correct classification rate (OCCR) of 92.25%, and the second partitioned set (70/15/15) had an OCCR of 93.34%. For the multi-class SVM, the OCCR was 92.22%, as observed in the study by Ibrahim et al. (2019). This study determined that using an ANN model can provide more accurate and precise results.

Singh et al. (2022) completed a study on classifying pistachios where they aimed to classify pistachios using a total of 2148 images, 1232 of Kirmizi type and 916 of Siirt type. They created three different CNN models using the transfer learning method, with AlexNet, VGG16, and VGG19. The three models had a classification success of 94.42%, 98.84%, and 98.14%, respectively.

A study on pistachio classification by Ozkan et al. (2021) used an improved KNN classifier. They performed principal component analysis (PCA) on the image data to reduce the feature vector size. Dimensionality reduction was used to prevent overfitting by eliminating the high dimensional data disadvantage. They developed a weighted and non-weighted KNN model using Euclidean distance, and the accuracy was 83.38% and 87.38% for the weighted model. Implementing PCA, they achieved better results; for the PCA weighted KNN model, they reached an accuracy of 94.18%. This study showed that PCA reduces the amount of noise in data and provides a higher accuracy model showing an increase in 10% accuracy for the weighted KNN model.

Dheir et al. (2019) conducted a study that classified five nuts: hazelnut, chestnut, nut forest, nut pecan, and walnut. The dataset contained 2868 images, 1390 training images, 550 validation images, and 910 test images. The features extraction had four convolutional layers: Relu activation function followed by max-pooling layer, two dense layers, and a layer with Softmax activation. After training the data and achieving an accuracy of 81.28%, it was found that the data was being overfitted, so they applied an augmentation to rescale the data and reached a validation accuracy of 94%. The overall accuracy achieved from the model was 98%.

Koklu et al. (2021) classified date fruits by variety after obtaining images, processing the photos, and extracting features from the pictures. There were seven types of dates: Barhee, Deglet Nour, Sukkary, Rotab Mozafati, Ruthana, Sawafi, and Sagai. The three classification models were LR, ANN, and stacking LR and ANN. Cross-validation was performed using k folds where k was equal to 10. KNN had an accuracy of 92.2%, LR had an accuracy of 91%, and stacking the two had an accuracy of 92.8%. From the confusion matrix developed, it was evident that the classification success of the Safawi variant was the highest, and the success of the Barhee variant was the lowest.

In a classification study on apple varieties, Sabanci et al. (2016) used KNN and MLP algorithms to classify 90 apples: Golden Delicious (30), Granny Smith (30), and Starking Delicious (30) apples. Images were captured of these apples, and the properties used for image processing were the diameter, area, perimeter, and fullness, as well as the colour properties red, green, and blue. The image processing results extracted seven features: radius, perimeter, area, volume, mass, and eccentricity. An open-source software, WEKA, was used to perform the classification. The best results for KNN were achieved using KNN, where k was 3 with a success of 97.77%. The highest accuracy for MLP was using MLP with five neurons in the hidden layer to achieve an accuracy of 98.88%. Therefore, classification using the MLP model provided better results.

6400 images of 40 different wheat grains were classified by Olgun et al. (2016) using dense SIFT features with an SVM classifier. After obtaining the DSIFT vector size, the dimensions were reduced using a k-means clustering algorithm to get k clusters where k equals 1000. A Bag of Words (BoW) model was created to reduce the features, and then 10-fold cross-validation was performed on the data before an SVM model was created. The resulting SVM model classified wheat grains with an overall accuracy of 88.33%.

Although very similar research has already been done in Classification of Rice Varieties Using Artificial Intelligence Methods (Cinar & Koklu, 2019), where they compared LR, MLP, SVM, DT, RF, NB, and KNN on the same set of data (Cammeo and Osmancik), this research will replicate and apply k-fold cross-validation using a different k-value to see if it helps improve the classifiers chosen. This classification results of DT, NB, and LR models after applying the k-fold cross-validation will be compared with the results achieved by Cinar and Koklu (2019), who used a k value of 4.

A lot of the classification studies that have been done used KNN, MLP, SVM, ANN, and CNN but hardly any have used Naïve Bayes, Decision Tree, and Logistic Regression. It would be interesting to compare the accuracy of the SVM and ANN classifiers by Ibrahim et al. (2020) as their study was also on the classification of grains of rice. Although they were classifying different variants from Cammeo and Osmancik, the process would be similar.

Regarding similar studies, Koklu et al. (2021) used LR to classify dates that are different from rice. Still, the study is similar to the research that will be completed as they also used k-fold cross-validation with k being ten, and the research that will be done will have a k-value other than 4.

Koklu, Cinar, and Taspinar (2021) did a related study on classifying five types of rice with CNN, ANN, and DNN, which are three different methods from the ones chosen for this research. So, the accuracy results will be compared with those achieved from DT, NB, and LR models.

It’s also important to note that the image collection, processing, and feature extraction have already been done to create the Cammeo and Osmancik dataset, which will be used in this research. The dataset was created based on the collection, processing, and feature extraction done by Cinar & Koklu (2019).

**Methodology**

The following methodology chart outlines the steps that will be taken in this research.

Diagram

Description automatically generated

# Data Description

**Univariate Analysis**

**Data Dictionary**

**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 region formed by the rice grain to the bounding box pixels

8. Class: Cammeo and Osmancik.

*Data Dictionary obtained from Cinar and Koklu(2019) pp.188-194.*

**Data Distribution**

The dependent variable is the binary variable ‘Class’, which is not too imbalanced. The data for the independent variables are distributed normally, as seen by the histograms below.

Chart, histogram

Description automatically generated

**Bivariate Analysis**

**Correlation Analysis**

The correlation between the independent variables is shown in the graph below. As we can see, the Area, Convex Area, Perimeter, and Major Axis Length are all very highly correlated. The Area and the Convex Area have the exact correlation with the Perimeter and Major Axis Length, so we can safely remove the feature with the lower variance, Area.

There is also a high correlation between Perimeter, Major Axis Length, and the Convex Area. After checking the variance of these three attributes, the conclusion was made to remove the Perimeter and Major Axis Length variables as they had the lower variances.

Table

Description automatically generated with low confidence

After removing the very highly correlated variables, we are left with the following correlations.

Table

Description automatically generated with medium confidence

**Pairwise Relations**

The pairwise relations chart shows the relations between the independent variables. The minor axis length and convex area have a positive relationship, and the minor axis length and eccentricity have a negative relationship.

A picture containing company name

Description automatically generated

Strip plots were used to show the relationship between the chosen quantitative independent variables and the categorical dependent variable.

These graphs show that the minor axis length, eccentricity, and the convex area of the Osmancik species tend to have a lower minimum and maximum compared to the Cammeo variety. It’s also evident that the variety does not affect the extent.

Chart, box and whisker chart

Description automatically generated

# References

Rice (Cammeo and Osmancik) Data Set. UCI Machine Learning Repository: Rice (Cammeo and Osmancik) data set. (n.d.). Retrieved May 24, 2022, from https://archive.ics.uci.edu/ml/datasets/Rice+%28Cammeo+and+Osmancik%29

Cinar, I. and Koklu, M. (2019). Classification of Rice Varieties Using Artificial Intelligence Methods. International Journal of Intelligent Systems and Applications in Engineering, vol.7, no.3 (Sep. 2019), pp.188-194. https://doi.org/10.18201/ijisae.2019355381.

Dheir, I. M., Mettleq, A. S. A., &amp; Elsharif, A. A. (2019). Nuts Types Classification Using Deep Learning. International Journal of Academic Information Systems Research (IJAISR), 3(12), 12–17.

Ibrahim, S., Amirah Zulkifli, N., Sabri, N., Amilah Shari, A., &amp; Noordin, M. R. (2019). Rice grain classification using multi-class support Vector Machine (SVM). IAES International Journal of Artificial Intelligence (IJ-AI), 8(3), 215. https://doi.org/10.11591/ijai.v8.i3.pp215-220

Ibrahim, S., Kamaruddin, S. B., Zabidi, A., &amp; Md. Ghani, N. A. (2020). Contrastive Analysis of rice grain classification techniques: Multi-class support Vector Machine vs artificial neural network. IAES International Journal of Artificial Intelligence (IJ-AI), 9(4), 616. https://doi.org/10.11591/ijai.v9.i4.pp616-622

Koklu, M., Cinar, I., & Taspinar, Y. S. (2021, June 23). Classification of rice varieties with deep learning methods. Computers and Electronics in Agriculture. Retrieved June 12, 2022, from https://www.sciencedirect.com/science/article/pii/S0168169921003021

Koklu, Murat & Kursun, Ramazan & Taspinar, Yavuz & Cinar, Ilkay. (2021). Classification of Date Fruits into Genetic Varieties Using Image Analysis. Mathematical Problems in Engineering. 2021. 10.1155/2021/4793293.

Olgun, M., Onarcan, A. O., Özkan, K., Işik, Ş., Sezer, O., Özgişi, K., Ayter, N. G., Başçiftçi, Z. B., Ardiç, M., &amp; Koyuncu, O. (2016, February 15). Wheat grain classification by using dense sift features with an SVM classifier. Computers and Electronics in Agriculture. Retrieved June 12, 2022, from https://www.sciencedirect.com/science/article/pii/S0168169916300060

Ozkan, Ilker Ali & Koklu, Murat & Saraçoğlu, Rıdvan. (2021). Classification of Pistachio Species Using Improved k-NN Classifier. Progress in Nutrition. 23. e2021044. 10.23751/pn.v23i2.9686.

Sabanci, Kadir & Ünlerşen, Muhammed. (2016). Different Apple Varieties Classification Using kNN and MLP Algorithms. International Journal of Intelligent Systems and Applications in Engineering. 4. 166-169. 10.18201/ijisae.2016Special%20Issue-146967.

Sabanci, K., Kayabasi, A., & Toktas, A. (2017). Computer vision‐based method for classification of wheat grains using artificial neural network. Journal of the Science of Food and Agriculture, 97(8), 2588–2593. https://doi.org/10.1002/jsfa.8080

Shahbandeh, M. (2022, January 31). Total global rice consumption 2021/22. Statista. Retrieved June 11, 2022, from https://www.statista.com/statistics/255977/total-global-rice-consumption/

Singh, Dilbag & Taspinar, Yavuz & Kursun, Ramazan & Cinar, Ilkay & Koklu, Murat & Ozkan, Ilker Ali & Lee, Heung-No. (2022). Classification and Analysis of Pistachio Species with Pre-Trained Deep Learning Models. Electronics. 11. 1-14. 10.3390/electronics11070981.