Machine Learning Project

Rice type classification: Jasmine and Gonen

Cerabolini Aurora, Capano Kevin, Pirola Federico, Strada Corinna

University of Milano-Bicocca
MSc Data Science
Academic Year 2021/2022

ABSTRACT

In order to ensure an effective and fruitful rice cultivation process, it is essential that the seeds belonging to a specific species are provided to farmers. In this regard, Machine Learning models were built with the aim of distinguishing between two species of rice (Jasmine and Gonen) in the most effective way possible. In addition, it was researched which of the two was the simplest type of rice to identify. The results may be useful, in the future, to ensure a more effective and faster selection of seeds compared to the selection method used today.

Contents

Introduction 1
Dataset 2
Classifiers and Performance Measures 2
  Classifiers 2
  Performance Measures 3
Pre-Processing, Hyperparameter Tuning e Accuracy Analysis 4
  Feature Selection 4
  Cross Validation 4
  Hyper-parameter Tuning 4
  Accuracy comparison 5
  Validation and confidence intervals 5
First research question 6
Second research question 7
Conclusion 7
References 8

Introduction

Rice is one of the most consumed products in the world[1]. Its cultivation begins with the selection of seeds, which are planted after suitable soil preparation.

For the entire life cycle of the plant, it is essential to control the management of water levels and nutrient absorption, determining, if necessary, the use of fertilizers. At the right time, we move on to the harvest and any post-harvest treatments[2].

The presence of problems relating to the seed selection step could delay or stop the entire production cycle: in particular, a fundamental step in this case is to ensure that there is no contamination or mix between different categories of rice.

In order for only one type of rice to be supplied to the growers, selection processes are carried out. However, these processes are manual[3] and are based on a small number of samples. The results, therefore, are often extremely unreliable. For this reason, the use of automated rice classification methods is recommended to ensure fast and reliable categorization.

In the dataset used for this project, two different categories of rice are considered: "Jasmine" and "Gonen".

Jasmine rice originates in Thailand and is characterized by a very high level of appearance, quality and aroma[4]. Gonen rice, on the other hand, comes from Turkey and is the third rice in the world in terms of productivity and, like all other

Turkish rice seed varieties, is characterized by high germination rates.

The primary goal, in this case, is to determine with the greatest possible success the type of rice grain based on the variables identified in the dataset, thus identifying the best classification model.

The secondary goal, on the other hand, is to establish which of the two varieties of rice is more easily identifiable.

The initial part of this article is dedicated to the description of the dataset and a preliminary analysis of the information contained within it. After that, there is a presentation of the developed classification models, the best hyper-parameters chosen, and the performance measures used.

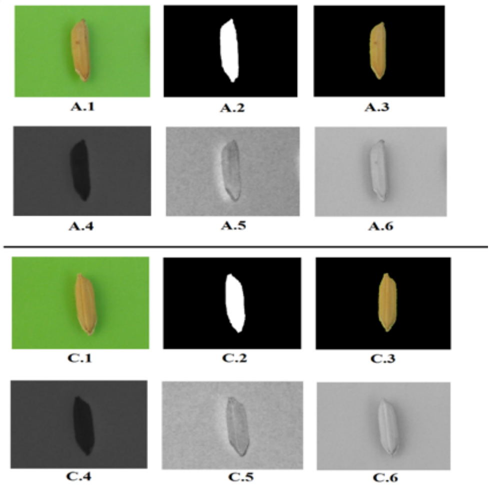
The final part of the work is therefore dedicated to the description of the analyses carried out and the results achieved, based on the two different research goals.

Dataset

There are no missing values in the dataset and all the variables, with the exception of the id and the target, are numeric.

The variables considered are 12 and were extracted from high-definition images of the beans:

- Id: identification code of each instance. This variable was not taken into account for the analyzes.
- Area: number of pixels in the region generated by the grain.
- Eccentricity: this measure presupposes the consideration of the grain of rice as an ellipse. In particular, it considers the relationship between the foci of the ellipse having the same second moments of the region generated by the grain and the length of its major axis.



- MajorAxisLength: length of the major axis of the ellipse which has the same second moments as the grain region.
- MinorAxisLength: length of the minor axis of the ellipse which has the same second moments as the grain region.
- ConvexArea: consists of the number of pixels that belong to the convex image, that is a binary image that contains all the pixels within the convex envelope. It is generated by the intersection of all subsets that contain a certain vector space.
- EquivDiameter: is the diameter of a circle that has the same area as that of the region of the grain of rice. In formula:

$$Equiv_{diameter} = \sqrt{\frac{4\ x\ Area}{\pi}}$$

- Extent: returns the ratio between the number of pixels relating to the grain region and those of the bounding box[5], i.e. the box with the smallest size in which the element can be contained.

$$Extent = \frac{Area}{Bounding\ Box\ Area}$$

- Perimeter: count of the number of pixels present on the outline of the grain of rice.
- Roundness: it is relative to the roundness of the grain of rice.
- AspectRation: is the ratio between the major axis and the minor axis of the ellipse that represents the grain.
- Class: is the target variable. It is binary, where 0 represents Gonen rice (45.1%), while 1 represents Jasmine (54.9%): the distribution is equal between the two classes, so the dataset is not unbalanced.

Classifiers and Performance Measures

Classifiers

In order to achieve the goals of this research, eight classification models have been implemented:

- Decision Tree: it consists of a succession of binary splits depending on the variable which guarantees maximum homogeneity in the nodes and maximum heterogeneity between the nodes in the target.
- Random Forest: it is obtained from the aggregation through bagging of the results of different trees built on subsets of randomly selected variables.

- Gradient Boosting: this type of ensemble model is also obtained from the aggregation of trees; however, while Random Forest builds each model independently, Gradient Boosting builds one at a time by iterating over the gradient of its optimization function.
- Logistic Regression: a non-linear regression model that can be used when the target is dichotomous.
- Support-Vectors Machines: often abbreviated to SVM, it leads to classification through the construction of hyperplanes.
- Multi Layer Perceptron: it is a particular type of neural network that has a single hidden layer.
- Stacking: A model consisting of a logistic regression, a Decision Tree (Cart) and a K-nearest neighbor, generated using the Stacking technique.
- Nearest Neighbor: it classifies the object on the basis of the characteristics of its neighbors.

Performance Measures

Having defined the Jasmine class (1) as a 'positive' class and the Gonen class (0) as a 'negative' class, the confusion matrix is reported:

		Observed	
		Positive	Negative
Predicted	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

- Accuracy: indicates the percentage of units correctly classified in their class of origin. This measure is used as a metric to be maximized when optimizing the model parameters.

$$\frac{TP + TN}{TP + FP + FN + TN}$$

- Specificity: also known as True Negative Rate, it represents the fraction of negative records correctly identified by the model. This metric is evaluated on the validation data to evaluate in percentage terms the correct classification of the Gonen class rice grains by the various models.

$$\frac{TN}{TN + FP}$$

- Sensitivity: Also known as True Positive Rate, it is the fraction of positive records correctly predicted by the classification model. As for Specificity, this metric is obtained in order to quantify the proportion of correctly classified Jasmine grains.

$$\frac{TP}{TP + FN}$$

- ROC: is the acronym for "Receiver Operating Characteristic Curve", and it is a graphic technique for evaluating classification models. The classification error for positives is represented on the abscissa axis (ie the complementary to 1 of the Specificity), while the Sensitivity value is represented on the ordinate axis. This shows how the percentages of correct classification of positives and negatives vary simultaneously. The greater the upward rectangularization of the curve, the better the performance of the model.

- AUC: area below the ROC curve, used as an assessment measure in the comparison of competitive models. The higher the AUC the better the performance of the model.

- LIFT: like ROC, it is a graphical technique for comparing binary classification models. The deciles (or percentiles) of the distribution are shown on the abscissa axis, while the Lift scores are shown on the ordinate axis. For each model the statistical units are ordered based on the probability of classification in the target interest class and divided into 10 intervals (hence the deciles). The lift values corresponding to the ratio between the percentage of correct classification in that decile and the Sensitivity calculated on the entire validation dataset are then calculated for each decile. The model that has the highest Lift value at the second / third decile is generally selected.

Pre-Processing, Hyperparameter Tuning e Accuracy Analysis

Feature Selection

The feature selection procedure, applied on the Training dataset, has the purpose of identifying which features are useful for solving the classification problem.

To identify the best performing attributes, the CfsSubsetEval approach was used: multivariate filter that selects jointly irrelevant and redundant attributes; therefore, the attributes included in the subset are uncorrelated with each other and are instead strongly associated with the target variable.

The model used to select the variables is a classification tree (J48), which has selected the five best performing variables: *Area, MinorAxisLength, Eccentricity, Extent* and *Roundness.*

The models are built both on the complete dataset and on the dataset obtained following the selection of the features in order to appreciate the differences in terms of classification performance.

Cross Validation

Before training the models on the Training data (67% of the total records) and passing them to the validation phase (33% of the total records), we proceed to calculate robust accuracy measures on the Training data by means of a Cross-Validation process in order to correctly classify the varieties of rice grains.

For some models, the optimization of the parameters of interest is carried out during the internal validation process, in order to train the models in the next phase once the combination of parameters that leads to maximizing accuracy (metric of interest) has been defined.

The number of partitions and therefore the number of iterations of the Cross - Validation cycle is fixed at 10. A stratified sampling is chosen in order to maintain the same frequency distribution of the classes for each partition. We then calculate the average of the 10 accuracy values obtained to obtain a robust measure of correct classification of the rice grains and to be able to make a first comparison between the classification models.

Hyper-parameter Tuning

While the Logistic Regression model and the Stacking criterion do not have parameters for which optimization is useful, for the other models we opt to select a combination of parameters that maximize the accuracy of the classifier.

Gradient Boosting is optimized as a function of the L1 (alpha) and L2 (lambda) penalty coefficients of the cost function, which help prevent overfitting by disadvantaging overly complex models.

For the Random Forest, the performance obtained by aggregating a different number of models (decision trees) is compared, while for the Decision Tree the minimum number of observations per final node is selected as a parameter to be optimized.

For the SVM we opt for a comparison between different Kernel functions in order to select the one that leads to a higher accuracy in the model, while for the k-Nearest Neighbor we select the number of neighbors that maximizes the performance of the model.

Finally, as regards the Multi Layer Perceptron, the combination of Momentum Rate and Learning Rate is selected in the Back Propagation process that leads to a higher level of accuracy.

For each model, the information relating to the ranges in which the optimization of the parameters takes place, the parameters selected during the optimization phase itself and the average accuracy measured in correspondence with the best combination of parameters are shown below.

Optimization results following the feature selection:

Model	Hyper - Parameter	Values	Step	Best Parms	Acc.
XG-Boost	Alfa Lambda	[0 : 1] [0 : 1]	0.1 0.1	0 0	0.974
Random Forest	Number of Models	[100: 500]	200	200	0.988
MLP	Learning R Momentum	[0 : 1] [0 : 1]	0.4 0.8	1 0.8	0.989
SVM	Kernel	{linear, polyn., radial, sigm.}	/	Linear	0.974
Decision Tree	Min. Obs. Final Nodes	[3 : 20]	13	15	0.987
k-NN	K - Neighbor	[2 : 10]	1	9	0.989

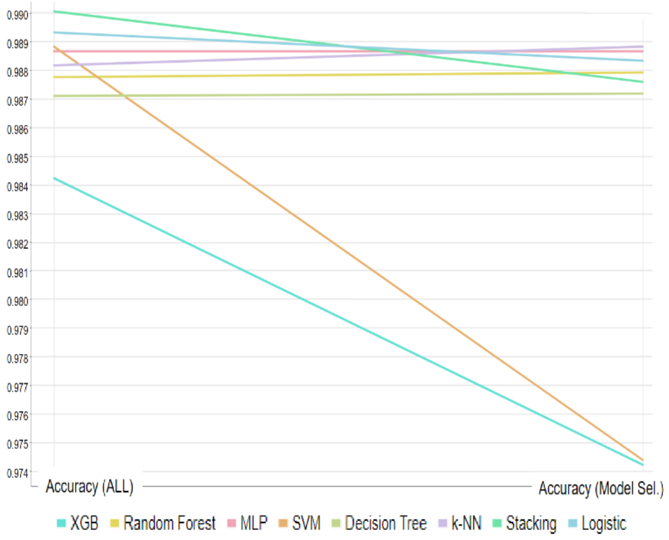
Optimization results of the models built on the complete training dataset:

Model	Hyper - Parameter	Values	Step	Best Parms	Acc.
XG-Boost	Alfa Lambda	[0 : 1] [0 : 1]	0.1 0.1	0 0	0.984
Random Forest	Number of Models	[100: 500]	100	200	0.988
MLP	Learning R Momentum	[0 : 1] [0 : 1]	0.2 0.2	1 0.8	0.989
SVM	Kernel	{linear, polyn., radial, sigm.}	/	Linear	0.989
Decision Tree	Min. Obs. Final Nodes	[3 : 20]	1	15	0.987
k-NN	K - Neighbor	[2 : 10]	1	10	0.988

Accuracy comparison

Below are the average accuracy values of the models obtained during the Cross-Validation process, comparing the averages obtained for the models built on the totality of the variables and those trained, following the feature selection, on the selection of variables adopted by the tree classification.

In general, it is noted that the average accuracy values obtained as a result of the feature selection are very close to those obtained from the models trained on the entire dataset. The only pairs that make an exception are the XG-Boost models and the Support-Vectors Machines. In particular, the Support-Vector Machines on the complete dataset and on the selection of variables have an average accuracy value of 0.989 and 0.974 respectively.

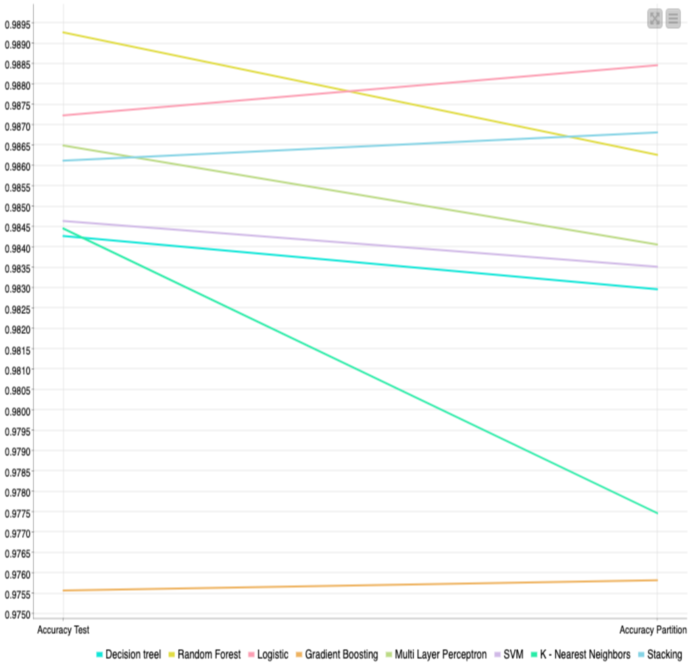


More generally, the XG-Boost is on average less performing than the other models in the classification of rice grains, while the Stacking method on complete data leads to a higher average accuracy value of 0.99.

Validation and confidence intervals

In addition to the accuracy values of the individual models, we are also interested in estimating the uncertainty of these values. To achieve this, the initial dataset was divided into Partition A (90%) and Partition B (10%) through the stratified sampling procedure in which the stratification variable is Class. Subsequently, partition A was further divided into A_Training (67% of the records of Partition A) and A_Test (33% of the records of Partition A).

The eight classifiers were trained on partition A_Training and tested on both A_Test and Partition B. The different Accuracy measures obtained were then compared by implementing a *line plot* that depicts the two levels of accuracy on the same graph:



It is immediately evident that the kNN model is the one that records the greatest differences between the Accuracy values depending on the partition. There are also minor differences in the other models, with the exception of the Gradient Boosting model.

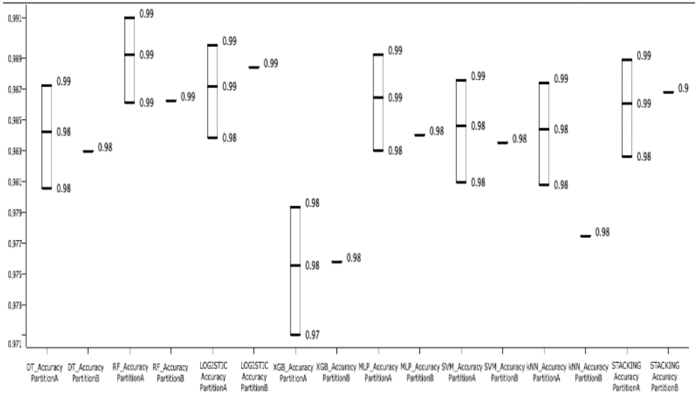
The Random Forest, Multi Layer Perceptron, SVM, Decision Tree and kNN models have a higher accuracy value in the A_Test partition while the opposite occurs for the Logistic and Stacking models.

To better highlight the differences between the two accuracies obtained for each model, the confidence intervals on the Accuracy obtained in the first partition were calculated.

The Wilson Score Interval was used for the confidence intervals at the 95% confidence level:

$$\left(\frac{acc + \frac{z^2_{1-\frac{\alpha}{2}}}{2N} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{acc}{N} - \frac{acc^2}{N} + \frac{z^2_{1-\frac{\alpha}{2}}}{4N}}}{1 + \frac{z^2_{1-\alpha/2}}{N}}, \frac{acc + \frac{z^2_{1-\frac{\alpha}{2}}}{2N} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{acc}{N} - \frac{acc^2}{N} + \frac{z^2_{1-\frac{\alpha}{2}}}{4N}}}{1 + \frac{z^2_{1-\alpha/2}}{N}} \right)$$

The Accuracy confidence intervals on the first partition and the Accuracy value obtained on Partition B were represented using the boxplots, in order to check whether the accuracy value on Partition B falls within the interval or not:



As you can see, for the k-NN classifier the Accuracy value obtained on Partition B does not fall within the confidence interval built for the accuracy tested on the A_Test partition. For this model, in fact, we had already seen how the line plot returned the line with the greatest slope.

For all the remaining models, the Accuracy value in Partition B is included in the confidence interval even if there are major variations in the Accuracy between partition A_Test and Partition B in the Random Forest, Logistic and Multi Layer Perceptron models.

In conclusion, taking into account Accuracy, the best model is Gradient Boosting.

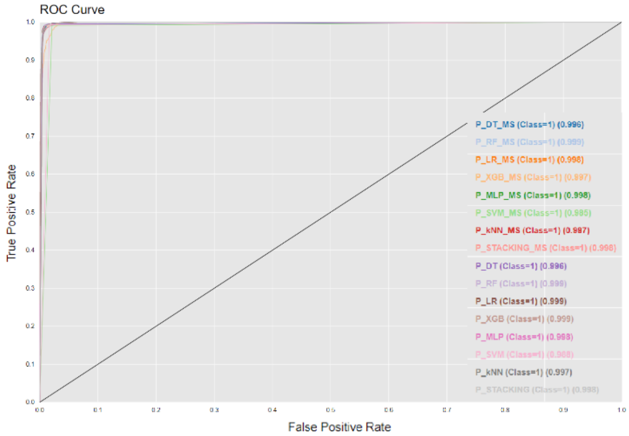
In order to obtain more performing models, in the subsequent analyses the models were trained with the combinations of parameters selected during the optimization phase described above.

First research question

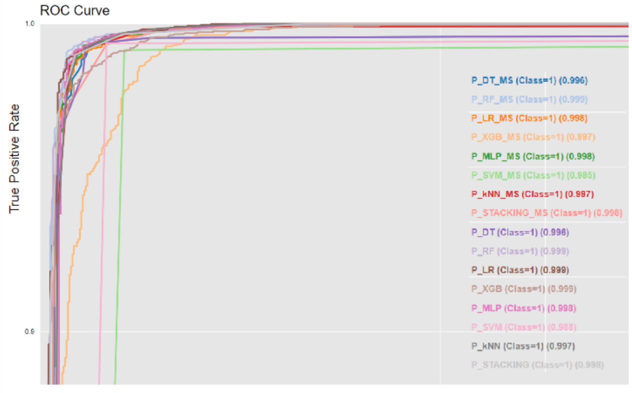
The best classification model is identified with the help of the ROC and LIFT metrics.

From the graph generated by the ROC we can observe that all the models have very good classification abilities; however, many curves overlap and therefore it is not possible to identify a classifier that performs better in absolute terms.

In the image, the models with the wording "MS" are the models trained following the feature selection.

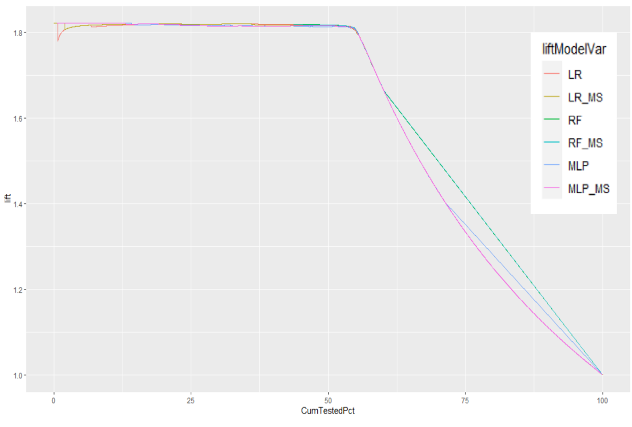


Since the curves are very overlapping and squashed together, below we can take a closer look at how the classifiers behave: what you immediately notice is that among the worst models there are Stacking and Gradient Boosting, while among the best ones there are the Random Forest, the Neural Network, the Logistics and the tree.



Based on these observations, it was decided to compare the models on the basis of their LIFT curves by referring to the LIFT metric: after defining a percentage of observations with higher posterior probability, this metric allows us to understand how much our model is performing in the correct classification of a specific class on that specific subset of observations. For simplicity we consider the classification capacity of the models considering the first 15% of the observations, ordered with respect to the posterior probabilities, with reference to the Jasmine variety (class = 1).

In the graph below, for clarity, only the lift curves of the best performing models have been displayed.



The resulting table is the following:

Modello	Lift Value (15′ percentile)
Tree	2.732
Tree MS	2.732
Random Forest	2.732
Random Forest MS	2.732
Logistic	2.726
Logistic MS	2.729
XGB	2.732
XGB MS	2.732
MLP	2.726
MLP MS	2.729
SVM	2.732
SVM MS	2.732
k-NN	2.732
k-NNMS	2.732
Stacking	2.732
Stacking MS	2.732

As can be seen, the classifications obtained with the different models lead almost all to the same Lift value, except the logistic model and the neural network.
Based on the joint observation of ROC curves and Lift values, also taking into account the very different computational times between the various algorithms, the classifier considered the best is the Random Forest.
In general, however, all classifiers lead to excellent results in the classification of rice grains.

Second research question

To identify the rice variety that is classified better on average, or with an average lower error, we compare the average Sensitivity and Specificity values obtained starting from the predictions of the trained models following the feature selection. These values are obtained on test data, independent from the training data on which the model was trained, in order to provide a more robust measure of accuracy.

As can be seen in the table below, both varieties of rice are classified very well (as was deducible from the accuracy values previously observed), however the Jasmine variety is able to be identified slightly better. This slight difference could be due to the higher number of observations tagged as "Jasmine" in the dataset compared to the number of "Gonen" class records, thus providing the models with more instances to learn to classify correctly.

Sensitivity (class = 1, “Jasmine”)	0.99
Specificity (class = 0, “Gonen”)	0.974

Conclusion

Ultimately, the variables obtained by analyzing the photos of the rice grains guarantee an almost perfect classification of the rice grains of the two varieties.
Furthermore, even considering a subset of these variables, the calculated performance metrics do not lead to results that are particularly distant from those obtained by taking into account the entire set of variables. In particular, the subset consisting of the area, the minor axis of the ellipse inscribed in the bean, the value of eccentricity, extent and roundness are already excellent in classifying the two varieties of rice.
Both varieties are correctly classified for more than 95% of cases. However, the 'Jasmine' variety has a higher percentage of correctly classified records, an indication that the variables selected through a dimensionality reduction technique are able to better identify the 'Jasmine' variety grains than those of the 'Gonen' class.
However, the idea that this slight difference in identifying capacity between classes is exclusively due to a greater presence of records labeled 'Jasmine', thus giving the classifiers the opportunity to learn better, remains valid.
In general, regardless of the classification criterion, a subset of variables is sufficient in order to predict the variety of rice with very high accuracy.

References

[1]https://www.cotecna.com/en/media/articles/world-rice-trade-in-brief#:~:text=According%20to%20the%20USDA3,South%2DEast%20Asia%20in%20particular.

[2]http://nipunarice.com/rice-o-pedia/cultivation-process/

[3]https://www.taste-institute.com

[4]https://scholarworks.uark.edu/cgi/viewcontent.cgi?article=1015&context=fdscuht

[5]https://medium.com/analytics-vidhya/basics-of-bounding-boxes-94e583b5e16c