Predicting the Rice Yield of Sub-Saharan African Farmers Using the Performance Indicators for Sustainable Farming

MinAh Kim

I. Executive Summary

Due to the rise in the global rice price, the policymakers in Sub-Saharan Africa may consider measures to improve domestic rice production to improve food security. This study purpose that the policymakers can monitor Net Profit, labor productivity, Nitrogen Use Efficiencies, and Phosphorous Use Efficiencies to predict rice yields. When I apply K-nearest neighbor and logistic regression using these four performance indicators, the models perform enough to have larger than 0.5 f-beta score. Hence, the policymakers should pay attention to research that studies the resource allocation among these four factors, such as the 2021 Arouna et al. paper.

II. Introduction

Food security became an immediate issue in Sub-Saharan African countries as the price of stable food skyrocketed due to the high global food price and their dependency on importing (Okou et al., 2022). One way to improve food security is to ensure that the domestic market produce enough to satisfy the demand of the market. However, to make the adequate policy choice for rice production, the government need to first understand the country's farming practices.

The Sustainable Rice Platform (SRP) provides twelve Performance Indicators (PIs) to monitor the sustainability of rice cultivation practices. Recently, Arouna et al. observed the five of the twelve Performance Indicators (PIs) in the twelve Sub-Saharan African countries: Benin, Cameroon, Cote d'Ivoire, Ghana, Madagascar, Mali, Niger, Nigeria, Senegal, Sierra Leone, Tanzania, and Togo (2021). By studying the PIs, they aim to help government understand where farmers allocate their resources and make adequate policy decisions based on the knowledge (Arouna et al., 2021).

The five PIs chosen by Arouna et al. are "grain yield, net profit, labor productivity, and nitrogen (N) and phosphorus (P) use efficiencies" (p.3, Arouna et al., 2021). However, Arouna et al. does not provide any quantitative evidence to prioritize the five PIs over the remaining seven PIs. It also does not clarify what questions regarding the rice production these five PIs can answer. This project attempts to compliment Arouna et al. by studying how well the five PIs can predict the level of the grain yields of the Sub-Saharan farmers.

III. Data

To best estimate the potential yield of a farm practicing sustainable rice production methods, I need a farm-level dataset with all sustainable performance indicators and the variables that reflect the regional variance of each farm. In this research, I will use the replication data provided by the Arouna et al.

The replication data for Arouna et al. is a subset of the Africa Rice Center's data. It has 2907 observations from the rice sector production hub of twelve Sub-Saharan countries (Arouna et al., 2021). It measures five PI measures: **grain yield, net profit, labor productivity, NUE (nitrogen use efficiency), and PUE (phosphorus use efficiency)** (Arouna et al., 2021). Arouna et al. calculated the five indicators based on the survey responses from the farmers, which were retrieved by National Agricultural Research System (NARS) partners and Africa Rice Center (AfricaRice) staff. The grain yield, for instance, was estimated by dividing "the total rice production by the rice area" (p.4, 2021). They calculate the net profit by subtracting the total cost reported through the survey from the multiplication of grain yield and the market price per yield (Arouna et al., 2021). Labor productivity refers to a grain yield divided by the total labor days and hectares of land (Arouna et al., 2021). Finally, Arouna et al. calculate the elemental N and P values in the fertilizers by using the known percentage of Nitrogen and Phosphorous in them (2021). Then, they divide the elemental N and P levels into the grain yield to derive the NUE and PUE indicators (Arouna et al., 2021).

	Grain Yield	**Net Profit**	**Labor Productivity**	**NUE**	**PUE**
Attribute Type	Numerical, Ratio	Numerical, Ratio	Numerical, Ratio	Numerical, Ratio	Numerical, Ratio
Unit	Kg/Ha	$/Ha	Kg/Labor Day	Kg grain / kg elemental N	Kg grain / kg elemental P
Count	2907	2907	2907	2396	2128
NA	0	0	0	511	779
Mean	2259.0	391.3	43.7	1314.2	4132.4
SD	2046.7	595.7	69.9	25953.4	66884.5
Min	100.0	-3587.8	1.0	0.0	0.0
25%	750.0	29.5	10.0	0.0	0.0
50%	1599.0	221.0	20.0	0.0	0.0
75%	3000.0	557.5	45.0	68.0	541.5
Max	9818.0	3953	691.0	858667.0	2000000.0

Table 1 Attributes related to the Performance Indicators

This data set also records other factors determining grain yields, such as the usage of potassium fertilizer, the seeding rate, the average cost per kg of paddy rice, the number of equipment types, the rice production system, the use of herbicide, and the use of insecticide (Arouna et al., 2022). However, it does not include the seven other Performance Indicators: "Food safety", "Water use efficiency", "Pesticide use efficiency", "Greenhouse gas emissions", "Worker health & safety", "Child labor", and "Women's empowerment" (SRP, 2020a).

Performance Indicators	Reflected in the Study?
Profitability: net income from rice	O
Labor Productivity	O
Productivity: grain yield	O
Food Safety	X
Water Use efficiency	X
Nutrient-use efficiency: N	O
Nutrient-use efficiency: P	O
Pesticide use efficiency	X
Greenhouse gas emissions	X
Worker health & safety	X
Child labor	X
Women's empowerment	X

Table 2 The Five Performance Indicators Available in the Data Set

Furthermore, numerous NUE and PUE values are missing or outliers. Out of 2907 observations, 511 NUE and 779 PUE values are missing. Arouna et al. define the optimal range of NUE and PUE to be each 30-100 kg grain kg-1 and 100-400 kg grain kg-1 (pp 4-5, 2021). Indeed, many farmers apply Nitrogen and Phosphorous outside the optimal range defined by the academia based on their practices. However, some NUE and PUE values from the Ghanaian hub overly exceed the range of their peers. Hence, for an accurate analysis, this data requires preprocessing to deal with the outliers and the missing values.

Arouna et al. attempt to capture the regional variance by identifying the country, the regional hub, and the rice production system (2021). This distinction in the geographical and production system is critical because this information allows for comparing outcomes within context. For instance, comparing the product of the rainfed and irrigated system is not adequate because irrigation produces more grain yields than rainfed farms (Jaramillo et al., 2021; Peleg, 2021). Figure 1 demonstrates that the grain yield difference between the irrigated and the rainfed systems also follows this trend in the Arouna et al. dataset. Thus, we need to use the relative values based on the country and the rice production system to compare the grain yield of the farms.

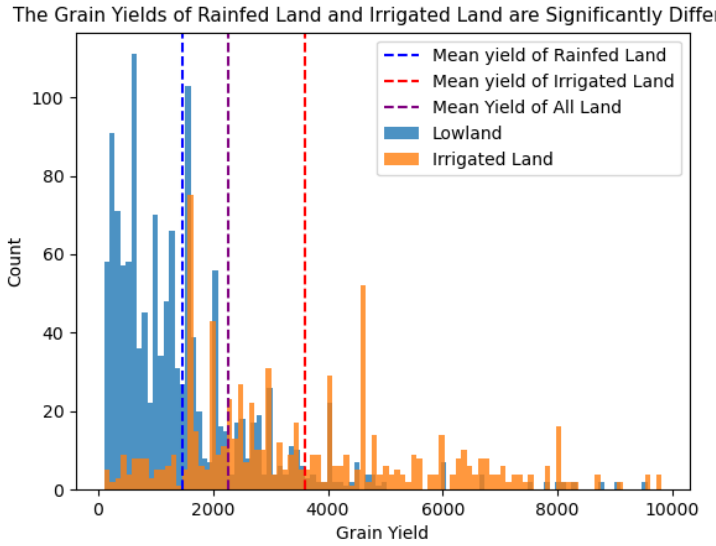


Figure 1 The Distribution of Grain Yields in Rainfed and Irrigated Land Differs Significantly

IV. Methodology

To create a model to predict the grain yield, I first need to clearly define *what to predict* and *what to use to predict*. First, I identify the target vector to clarify *what to predict*. Arouna et al. introduce three yield gap categories based on the farm's grain yield percentile relative to the other farms in the same country and the rice production system (2021). They divide the farms into the top 10%, the middle 80%, and the lowest 10% yield gap compared to their peers. Since the grain yield is not directly comparable across countries or production systems, I will adapt this classification and train the model to predict the percentile categories of grain yield of each farm in the same country and production system.

The features matrix will have the four PIs discussed in Arouna et al.: Net Profit, Labor Productivity, NUE, and PUE (2021). I exclude grain yield as it will be the subject of prediction. To address the outliers in the NUE, PUE, and profit, I will exclude data that has profit lower than -1,000, NUE over 10,000, and PUE higher than 20000. I determine the outliers based on observing the discontinuity in the data distribution.

As discussed in the previous section, NUE and PUE has a significant number of the missing values, each missing 17.5% and 26.8% of the total NUE or PUE observation. In this analysis, I will first exclude all values with missing observations. Then, I will compare the result after applying the Nearest Neighbors imputation (KNN imputation) to supplement the missing value. When an observation has a missing value, the KNN imputation finds the observation that share the most similar attributes excluding the missing value (scikit-learn, n.d.). After that, based on the distance from the observation to their neighbors, it predicts and replaces the missing value (scikit-learn, n.d.). The assumption behind using the KNN imputation is that farmers with similar farming practices are more likely to make similar choices in applying Nitrogen and Phosphorous. I hold this assumption to be valid in the context of rice production and apply it

throughout the study. Furthermore, the past literature also finds the KNN imputation to be robust even when 20% of data is missing (Troyanskaya et al., 2001). Therefore, I will use KNN imputation to replace the missing values in NUE and PUE.

Based on the target array and feature matrix, I implement the K-Nearest Neighbor (KNN) and logistic regression to build a predictive model. Similar to the KNN imputation, the KNN assumes the farmers with similar farming practices to be in a same grain yield percentile group. The KNN predict the group of an observation based on the group of the neighboring values, which are determined by calculating the distances between the values. Hence, past literature often used the K-Nearest Neighbor when they know that certain similar qualities lead to similar results. For instance, Ren et al. mention that pharmaceutical companies predict the quality of the medicine using KNN because they understand similar the raw material and manufacturing lead to similar result (2020).

On the other hand, the logistic regression calculates the log odd of being in a group, which can be translated to the probability of being in a category. Since this study divides into three groups, I will use the multinomial option provided by the scikit-learn's Logistic Regression classifier. In the past, scientists applied the multinominal logistic regression to predict the effect of the rainfall during storms to find implication applicable in larger setting (Szeląg et al., 2022). Similarly, I will use logistic regression to predict the grain yield of rice in Sub-Saharan Africa, which is a very large space of land.

Both KNN and logistic regression can be tricky in dealing with missing values. KNN often cannot calculate the distance of each observation with missing values while logistic regression involves all attributes to calculate the possibility of the observation being in each category (Tan et al. 2019). Yet, KNN has value in our case since I can rely on an easily agreeable assumption that farmers with similar farming pattern will have similar results. Logistic regression also has its unique advantage that it can handle attributes that is not relevant to the prediction (Tan et al. 2019). Thus, if some PIs are more useful in predicting than others, this discrepancy will show up from the performance differences between KNN and the logistic regression.

To evaluate the performance of each model, I utilize the precision, recall, f-beta score, and the Jaccard score for each group. To calculate these scores for each percentile group, I relabel each group as true to the rest of the values to be false. For instance, for the low 10% percentile group:

- the **precision score** divides the True Positive (the number of values that are and predicted to be in the low 10% group) to the number of all positives (the number of values that are predicted to be in the low 10% group);
- the **recall score** divides the True Positive to the number of all true values (the number of values that are in the 10% group);
- the **f-beta score** is a harmonic mean calculated based on putting the equal weight on the precision and the recall score from above.

V. Findings

First, I apply the KNN model while dropping all the missing values. Since the model performance halts constantly improving after the number of neighbors exceeds four, I choose the number of neighbors to be four. The precision and recall scores range around 0.7 throughout all three percentile categories. I can also observe that the model makes very few extreme mistakes, such as categorizing the low 10% (denoted as -1 in the confusion matrix) to be in the high 10% (1 in the confusion matrix) or vice versa.

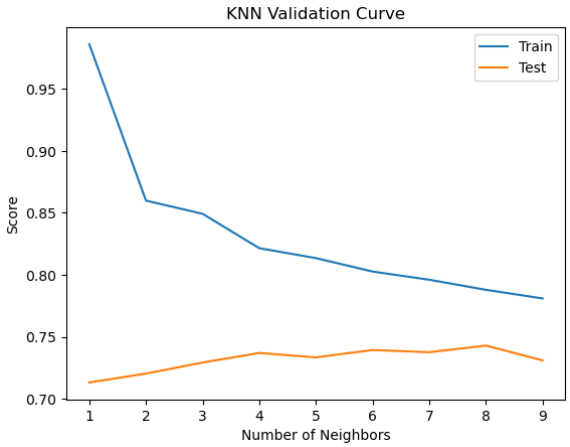


Figure 2 KNN Validation Curve

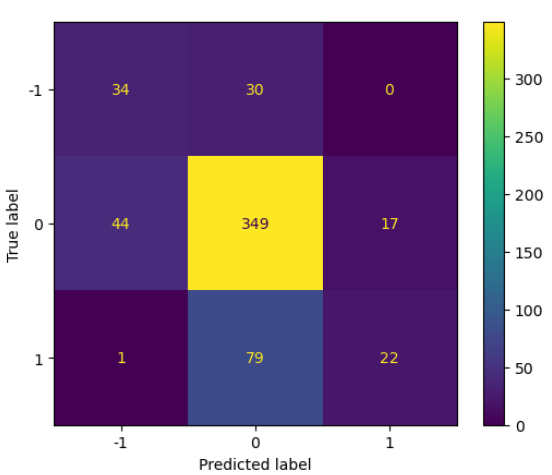
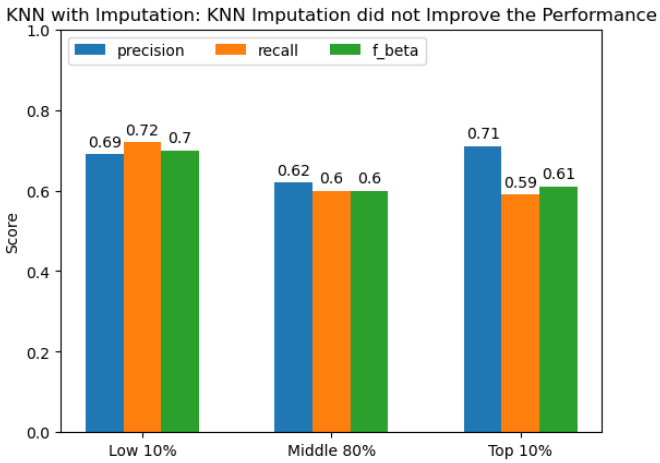


Figure 3 Confusion Matrix for KNN with KNN Imputation



Bar Chart 1 Performance Scores for Each Percentile Group for KNN with KNN Imputation

Next, I explore if the KNN imputation improves the performance of the KNN model. Since I observe the KNN model performs reasonably well in this data set when it compares a value with four neighbors, I also set up the KNN imputation algorithm to deduct the missing value by comparing four neighbors. The KNN imputation reduces the number of extreme mistakes, leaving only one case to be categorized as the lowest 10% percentile group when it was in the top 10% group. However, the KNN imputation slightly worsens the performance scores for each group. For example, the f beta scores for the middle and top 10% group move down from the 0.65 range to the near 0.6 range. Hence, filling the missing values with the KNN imputation may be useful for preventing the model from predicting completely opposite result but barely helps affects the overall performance.

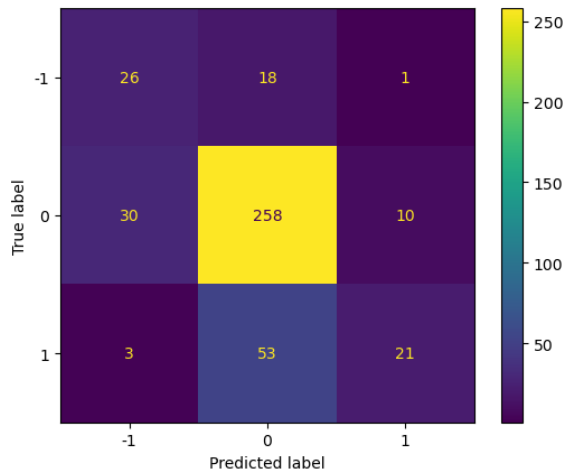
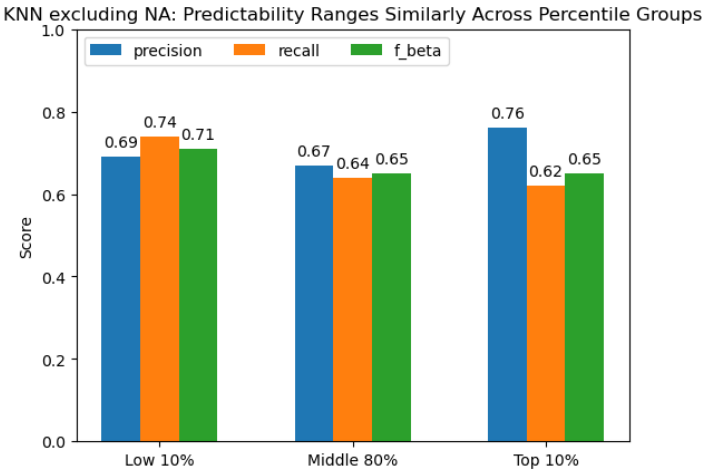


Figure 4 Confusion Matrix for KNN After Dropping Missing Values



Bar Chart 2 Performance Scores for Each Percentile Group for KNN after Dropping Missing Values

The performance of the logistic regression shows different trend from the KNN models. When I perform a logistic regression after dropping the missing value, the logistic regression estimates more values to be in the middle 80% range in comparison to the KNN models. The confusion matrix illustrates this trend through a lighter vertical line that appears when the

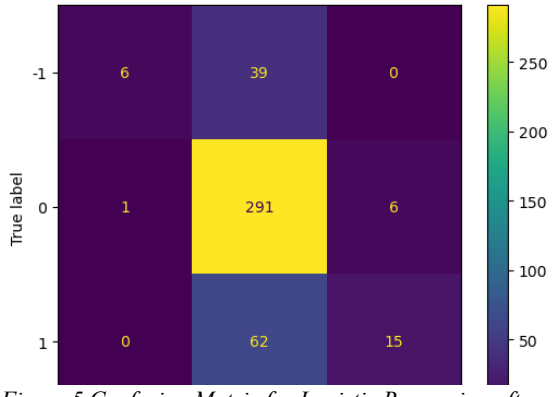
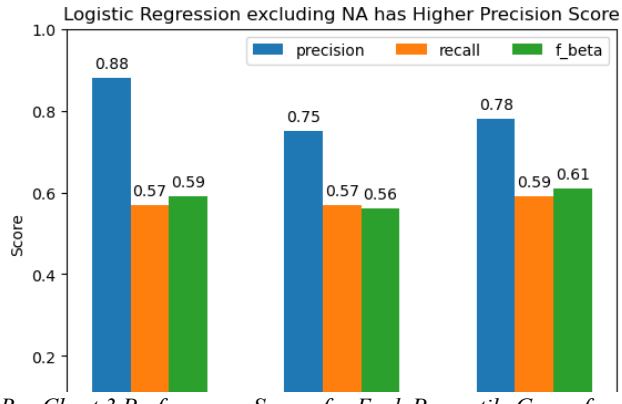


Figure 5 Confusion Matrix for Logistic Regression after Dropping Missing Values



Bar Chart 3 Performance Scores for Each Percentile Group for Logistic Regression After Dropping Missing Values

predicted label is 0 (middle 80%). Due to this tendency, the prediction scores for each group increase nearly around 0.8 or even higher as fewer values are classified into the lower and the top 10% on the first hand. On the other hand, the recall scores for the low 10% drops from 0.72 to 0.57. Hence, the f-beta scores decrease in all groups in this logistic regression compared to the KNN models.

I compare this result with a logistic regression after applying the same KNN imputation from the KNN models and discover that the KNN imputation does not change the trend of the prediction. The logistic regression does not make any extreme mistake like before, but it still tends to predict most of the values to be in the middle 80%. Consequently, the high precision score and the slight decrease in recall score persist. As a result, supplementing the estimated values using KNN imputation rarely affects the result of the logistic regression.

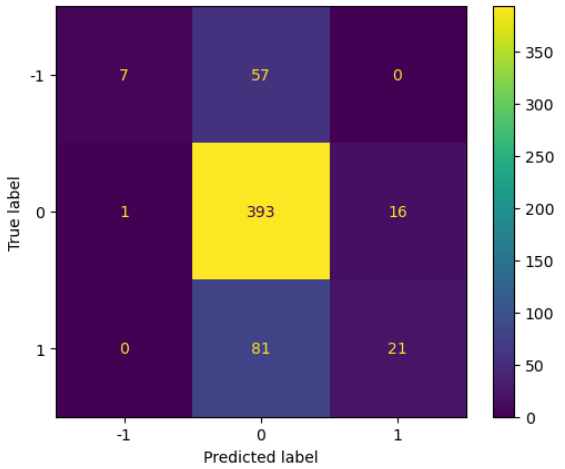
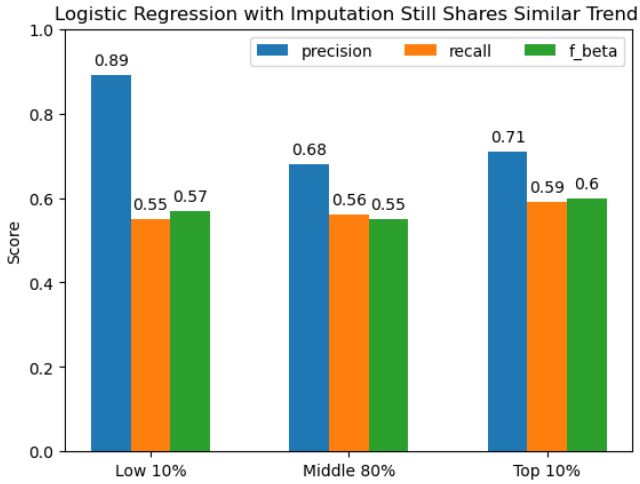


Figure 6 Confusion Matrix for Logistic Regression with KNN Imputation



Bar Chart 4 Performance Scores for Each Percentile Group for Logistic Regression with KNN Imputation

VI. Conclusion

I discover that the PIs chosen by Arouna et al. have some predictive power in estimating the grain yield. I perform a machine learning model implementing KNN and logistic regression classifier using the four PIs of interest in Arouna et al. – net profit, labor productivity, NUE PUE. The fifth PI, grain yield, becomes the subject of prediction. Although the logistic regression tends to prefer classifying observations into the middle 80% percentile, the f-beta score always ranges around 0.7 for KNN model and 0.6 for logistic regressions regardless of the missing values. Moreover, both KNN and logistic regression classifier rarely sorts the observations from the low 10% or the top 10% observation to the complete opposite group. Based on these pieces of evidence, I conclude that monitoring the PIs discussed in Arouna et al. can provide insight to the rice grain yield of the region.

Policy Implication

The policymakers can use the four PIs to estimate the domestic rice production capacity of the region. Using the estimated rice production, they can consider controlling the import or allocating the resources to support domestic farmers to stabilize the market. This study may also help finding strategy to improve the domestic rice production, but it *does not* study the social implication of increased rice production. Since the rice production in Sub-Saharan Africa is mostly the legacy of plantation during the colonial era, the policymaker might consider replacing rice with other more productive stable to improve the food security of the region. These model also do not consider what happens to the domestic rice price when the productivity increases even though net profit is deduced based on the market rice price. In fact, this study does not provide any causal inference between the PIs and the rice yield. Hence, the policymaker should decide how to enhance food security with a holistic perspective.

Limitation and Future Consideration

This study only examines the five PIs out of twelve PIs due to the limitation of information. Studying the significance of other PIs will be helpful because of two reasons. One, PIs such as greenhouse gas emissions, child labor, and women empowerment, have data pertain information on social inclusion and climate change. Hence, it can provide policy implication in the area that are often neglected in the past. Two, the other PIs may be easier to obtain. All four PIs are all deduced values from survey. Since the farmers do not measure values such as average nitrogen use or net profit in their lives, the researchers have to calculate the five PIs. In this process, the researchers divide numbers into the grain yield, which is our prediction interest but ironically also easier to find. Thus, understanding the relevance of other PIs may be convenient for the policymakers to gather information.

VII. Bibliography

Arouna, A., Devkota, K. P., Yergo, W. G., Saito, K., Frimpong, B. N., Adegbola, P. Y., Depieu, M. E., Kenyi, D. M., Ibro, G., Fall, A. A., & Usman, S. (2022). *Replication Data for: Assessing rice production sustainability performance indicators and their gaps in twelve sub-Saharan African countries*. [Data Set]. Harvard Dataverse. https://doi.org/10.7910/DVN/6K4AB4

Arouna, A., Devkota, K. P., Yergo, W. G., Saito, K., Frimpong, B. N., Adegbola, P. Y., Depieu, M. E., Kenyi, D. M., Ibro, G., Fall, A. A., & Usman, S. (2021). *Assessing rice production sustainability performance indicators and their gaps in twelve sub-Saharan African countries.* Field Crops Research, 271, 108263–108263. https://doi.org/10.1016/j.fcr.2021.108263

Brady, N.C. (1981). *Soil Factors that Influence Rice Production*. Proceedings of Symposium on Paddy Soils. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-68141-7_1

DataFlair Team. (2021, March 8). *Kernel functions-introduction to SVM Kernel & Examples*. DataFlair. Retrieved April 10, 2023, from https://data-flair.training/blogs/svm-kernel-functions/

Devkota, K. P., Pasuquin, E., Elmido-Mabilangan, A., Dikitanan, R., Singleton, G. R., Stuart, A. M., Vithoonjit, D., Vidiyangkura, L., Pustika, A. B., Afriani, R., Listyowati, C. L., Keerthisena, R. S. K., Kieu, N. T., Malabayabas, A. J., Hu, R., Pan, J., & Beebout, S. E. J. (2019). *Economic and environmental indicators of sustainable rice cultivation: A comparison across intensive irrigated rice cropping systems in six Asian countries*. Ecological Indicators, 105, 199–214. https://doi.org/10.1016/j.ecolind.2019.05.029

Devkota, K. P., Sudhir-Yadav, Khanda, C. M., Beebout, S. J., Mohapatra, B. K., Singleton, G. R., & Puskur, R. (2020). *Assessing alternative crop establishment methods with a sustainability lens in rice production systems of Eastern India*. Journal of Cleaner Production, 244, 118835–118835. https://doi.org/10.1016/j.jclepro.2019.118835

Dossou-Yovo, E. R., Vandamme, E., Dieng, I., Johnson, J.-M., & Saito, K. (2020). *Decomposing rice yield gaps into efficiency, resource and technology yield gaps in sub-Saharan Africa.* Field Crops Research, 258, 107963–. https://doi.org/10.1016/j.fcr.2020.107963

Hsu, C.-N., Huang, H.-J., & Wong, T.-T. (2003). Implications of the Dirichlet Assumption for Discretization of Continuous Variables in Naive Bayesian Classifiers. Machine Learning, 53(3), 235–263. https://doi.org/10.1023/A:1026367023636

Jaramillo, S., Graterol, E., & Pulver, E. (2020). Sustainable transformation of rainfed to irrigated agriculture through water harvesting and smart crop management

practices. *Frontiers in Sustainable Food Systems*, *4*. https://doi.org/10.3389/fsufs.2020.437086

Okou, C., Spray, J., & Unsal, D. F. (2022, September 27). *Africa food prices are soaring amid high import Reliance*. IMF Blog. Retrieved March 2, 2023, from https://www.imf.org/en/Blogs/Articles/2022/09/26/africa-food-prices-are-soaring-amid-high-import-reliance

Peleg, L. (2021, February 1). *From rainfed to irrigated agriculture: AgBlog*. Netafim. Retrieved April 9, 2023, from https://www.netafim.com/en/blog/from-rainfed-to-irrigated-agriculture/

Ren, J., Zhou, R., Farrow, M., Peiris, R., Alosi, T., Guenard, R., & Romero-Torres, S. (2020). *Application of a kNN-based similarity method to biopharmaceutical manufacturing. Biotechnology Progress*, 36(2), e2945–n/a. https://doi.org/10.1002/btpr.2945

Saito, K., Six, J., Komatsu, S., Snapp, S., Rosenstock, T., Arouna, A., Cole, S., Taulya, G., & Vanlauwe, B. (2021). *Agronomic gain: Definition, approach, and application.* Field Crops Research, 270, 108193–108193. https://doi.org/10.1016/j.fcr.2021.108193

scikit-learn. (n.d.). *Imputation of Missing Values*. scikit-learn documentation. Retrieved May 9, 2023, from https://scikit-learn.org/stable/modules/impute.html#knnimpute

Sreenivasa, S. (2020, October 12). *Radial basis function (RBF) kernel: The go-to kernel*. Medium. Retrieved April 9, 2023, from https://towardsdatascience.com/radial-basis-function-rbf-kernel-the-go-to-kernel-acf0d22c798a

Sustainable Rice Platform (SRP). (2020a). *Sustainable Rice Platform Performance Indicators for Sustainable Rice Cultivation (Version 2.1)*. https://sustainablerice.org/wp-content/uploads/2022/12/203-SRP-Performance-Indicators-Version-2.1.pdf

Sustainable Rice Platform (SRP). (2020b). *Sustainable Rice Platform Standard for Sustainable Rice Cultivation (Version 2.1)*. https://sustainablerice.org/wp-content/uploads/2022/12/103-SRP-Standard-Version-2.1.pdf

Szeląg, B., Suligowski, R., Majewski, G., Kowal, P., Bralewski, A., Bralewska, K., Anioł, E., Rogula-Kozłowska, W., & De Paola, F. (2022). *Application of Multinomial Logistic Regression to Model the Impact of Rainfall Genesis on the Performance of Storm Overflows: Case Study*. Water Resources Management, 36(10), 3699–3714. https://doi.org/10.1007/s11269-022-03223-z

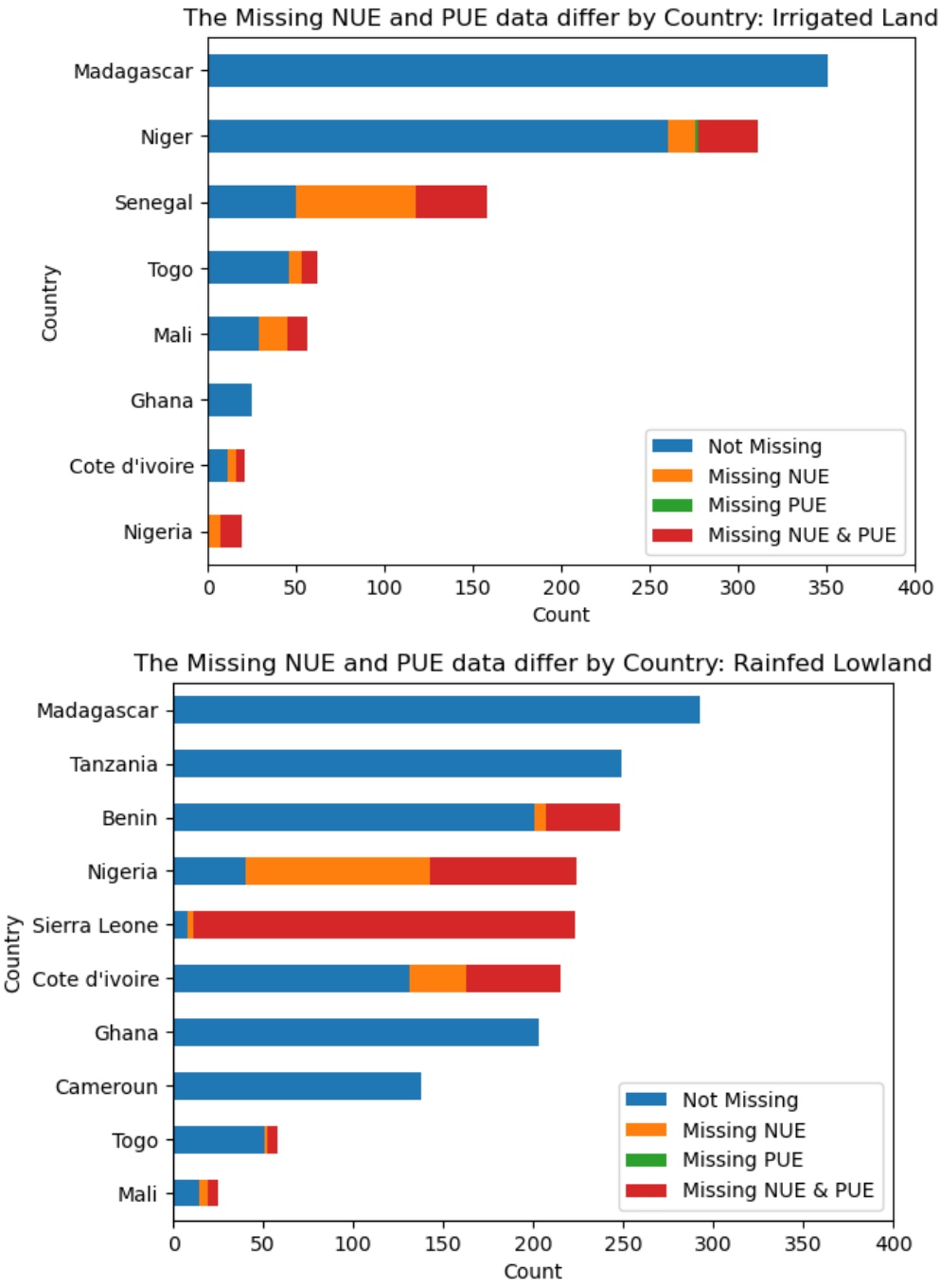
Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to data mining* (Second edition.). Pearson Education, Inc.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. BIOINFORMATICS, 17(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

van Ittersum, M. K., van Bussel, L. G. J., Wolf, J., Grassini, P., van Wart, J., Guilpart, N., Claessens, L., de Groot, H., Wiebe, K., Mason-D'Croz, D., Yang, H., Boogaard, H., van Oort, P. A. J., van Loon, M. P., Saito, K., Adimo, O., Adjei-Nsiah, S., Agali, A., Bala, A., … Cassman, K. G. (2016). *Can sub-Saharan Africa feed itself?* Proceedings of the National Academy of Sciences - PNAS, 113(52), 14964–14969. https://doi.org/10.1073/pnas.1610359113

VIII. Implementation Appendix

Appendix 1 The Distribution of Missing Data for Each Country and Rice Production System



These graphs demonstrate the distribution of missing NUE and PUE values based on the country and the rice production system. When I conduct the machine learning model before imputation, I am excluding all the red, orange, and green part of the bars. In other words, the result will reflect the countries with many missing data, such as Sierra Leone, significantly less.

On the other hand, the KNN imputation fills the non-blue part of the bar by deducing the likely value from the existing data. Therefore, the representation of countries such as Sierra Leone and Nigeria will improve in KNN imputed dataset.

As I discuss in the founding, the KNN imputation does not create noticeable difference in the result in both classifiers. However, I do not compare KNN with other imputation methods in the study. If someone can apply a more suitable imputation method, the difference in missing values in each area may lead to noteworthy differences in result.