Development of a Python - Based Proficiency Indicator with Data Imputation Algorithm for Filipino STEAM Educators

Timothy M. Amado[1], Bryan D. Brillo[1], Jesmar R. Enerio[1], Rodnick Jose G. Prudente[1], Eloisa Mae B. Ramirez[1], Chester Keith A. Rivera[1], Ira C. Valenzuela[1], Edmon O. Fernandez[1], John Peter M. Ramos[1], Glenn C. Virrey[1]
[1]Electronics Engineering Department
Technological University of the Philippines
Manila, Philippines

***Abstract* – Lengthy questionnaires are used to determine the proficiency level of educators, but the problem with this is that, oftentimes, not all items are answered by the respondents. Incomplete data is one of the prevalent challenges in the area of statistical research since it causes the sample data to be inaccurately analyzed and drew erroneous conclusions. The goal of this research is to determine the best suited imputation method for the occurrence of missing data in the national proficiency dataset. Several of the widely used imputation methods, referred here as benchmark methods, were used in comparison to predict the missing data. These methods include K-Nearest Neighbor (KNN), Decision Tree, Random Forest, Gaussian Naïve Bayes and Linear Discriminant Analysis. From the complete data of national proficiency dataset, fixed interval percentages of missing data were removed and then generated by performing the different benchmark methods. Then, the performance of each data imputation method was evaluated through running a series of statistical experimentations. The results of analysis have shown that KNN imputation method outperforms the other methods in terms of accuracy and reliability. Moreover, it has also been found that the performance of the data imputation method is independent of the percentage of missing values in the dataset.**

***Keywords* – Recommendation System, Pedagogy, STEAM Education, K-Nearest Neighbor, Predictive Model, Lesson Exemplar, Big Data Analytics**

I. Introduction

On the Technological, Pedagogical, and Content Knowledge (TPACK) in Philippine STEAM Education 2017-2019, the development of the models was dependent on the classroom observations and interviews of random sampled educators country-wide, and the answers on the questionnaire of the educators, totaled to 1455 samples. [1] On these online survey forms, the respondents had answered the sixty (60)-item proficiency indicators completely, and not one must be left unanswered. The procedure in determining one's proficiency must require a dataset with complete information [2]. If one indicator has missing information, the formula would yield undefined result. Several feedbacks from the sampled population insisted the inconvenience in completing to answer the whole sixty (60) items, even though it was necessary to produce accurate result.

In response to the unforeseen dilemma, the concept of imputing the missing data from the respondent's dataset was considered. Several data imputation methods are used and developed in specific area coverage. One is the combination of hot deck imputation, which works when the imputed data comes from the same person who gave a missing dataset, and fractional imputation [3]**.** Also, Hadeed et.al. compared different imputation methods such as Mean Imputation, Median Imputation, Kalman Filter, Last Observation Carried Forward, and Markov chain for the missing data in monitoring air pollutants [4].

All these methods of imputation differ in the way it will be employed. The accuracy and performance must be considered in choosing the right data imputation to be used. In this case, missing data on a survey form must be completed to fully accomplish the procedure in determining the proficiency of the educator, on top of the instrument to be made for the complete dataset.

II. Methodology

This study aims to develop an instrument which clusters the proficiency of the STEAM educators for the generation of lesson exemplars. It also aims to develop a predictive model in case the respondents provide incomplete data in their self-rating proficiency indicator.

A. Creating the Instrument for Complete Data

Figure 1 shows how the proficiency of each educator was taken in every domain or dimension. Using the samples gathered, the proponents developed an instrument with the same procedure as the instrument used in the [5]. The Individual Proficiency Profile is the basis of how the respondents was clustered in a career stage suited for them. Each item in the self-rating proficiency indicators was answered using the following scales: 4 – Always true to myself, 3 – Often true to myself, 2 – Occasionally true to myself, 1 – Rarely true to myself, and 0 – Not Applicable (N/A). The individual ratio was then computed on how many times each scale was answered in a range of items comprises in a domain or dimensions.

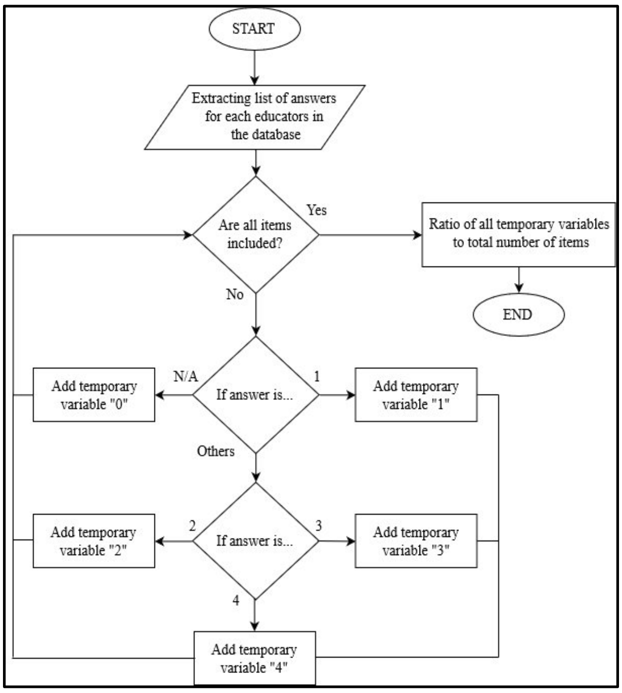
START
Extracting list of answers for each educators in the database
Are all items included?
Yes
Ratio of all temporary variables to total number of items
END
No
N/A
Add temporary variable "0"
If answer is...
1
Add temporary variable "1"
Others
2
Add temporary variable "2"
If answer is...
3
Add temporary variable "3"
4
Add temporary variable "4"

Fig. 1. Flowchart of the Development of Individual Proficiency Profile

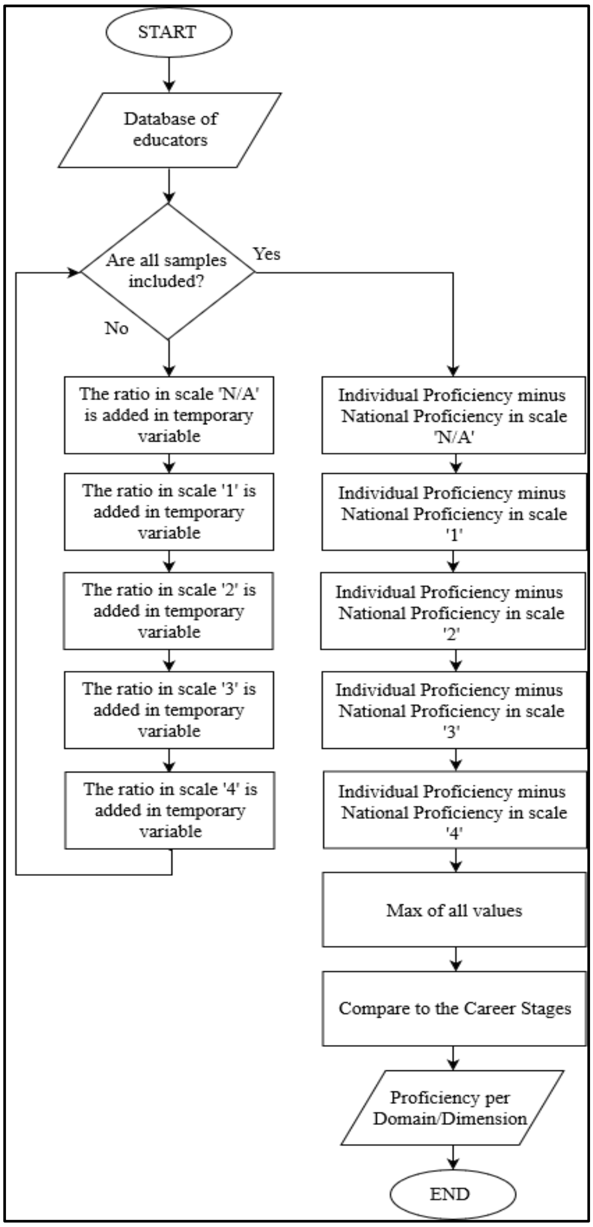
START
Database of educators
Are all samples included?
Yes
No
The ratio in scale 'N/A' is added in temporary variable
The ratio in scale '1' is added in temporary variable
The ratio in scale '2' is added in temporary variable
The ratio in scale '3' is added in temporary variable
The ratio in scale '4' is added in temporary variable
Individual Proficiency minus National Proficiency in scale 'N/A'
Individual Proficiency minus National Proficiency in scale '1'
Individual Proficiency minus National Proficiency in scale '2'
Individual Proficiency minus National Proficiency in scale '3'
Individual Proficiency minus National Proficiency in scale '4'
Max of all values
Compare to the Career Stages
Proficiency per Domain/Dimension
END

Fig. 2. Flowchart of the Development of National Proficiency Profile

A.1 Development of National Proficiency Profile

The National Proficiency Profile which serves as a reference in determining the proficiency of each educator was shown in Figure 2. It was taken from the mean ratio of all the samples answered from the certain scale. The individual proficiency ratio was subtracted to the national proficiency ratio in each scale. Then, greatest positive difference will determine the proficiency of the educators in a specific domain or dimension. The mean of the greatest positive difference of all domain or dimensions will correspond to the overall proficiency of the educators.

B. Testing Different Imputation Algorithms for Incomplete Data

To test the accuracy of the proposed missing data imputation methods, a series of statistical experiments on the national proficiency data set was done. In these experiments, full data set was used in which all entries are known, and random patterns of missing data for various percentages ranging from 10% to 50% will be generated. The proponents took the full data set that has no missing entries to be the ground truth.

The most used methods were run as benchmark methods for data imputation on these data sets to predict the missing value. The imputation methods in this comparison are:

a. *K-Nearest Neighbors* – imputes missing values using the K-nearest neighbors of an observation based upon Euclidean distance. [6]

b. *Decision Tree* – used to go from observation about an item to conclusion about the item's target value. [7]

c. *Random Forest* – operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of mean prediction of the individual trees [8]

d. *Gaussian Naive Bayes* – is the method which imputes based on the mean and standard deviation of the training data. [9]

e. *Linear Discriminant Analysis* – a method to find a linear combination of features that characterizes or separates two or more classes of objects. [10]

C. Creating a Predictive Model for Incomplete Data

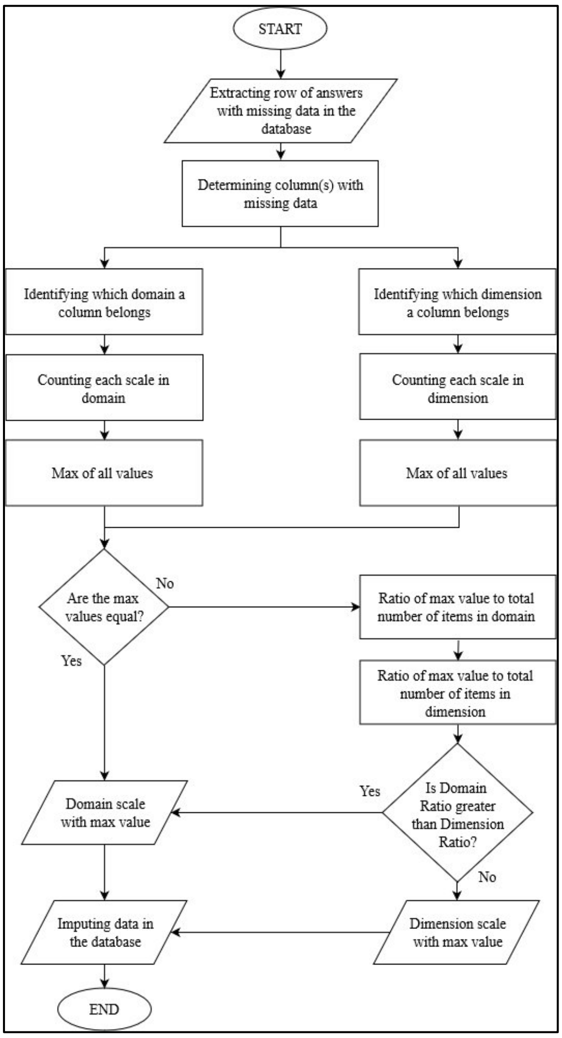
START
Extracting row of answers with missing data in the database
Determining column(s) with missing data
Identifying which domain a column belongs
Identifying which dimension a column belongs
Counting each scale in domain
Counting each scale in dimension
Max of all values
Max of all values
Are the max values equal?
No
Yes
Ratio of max value to total number of items in domain
Ratio of max value to total number of items in dimension
Is Domain Ratio greater than Dimension Ratio?
Yes
No
Domain scale with max value
Dimension scale with max value
Imputing data in the database
END

Fig. 3. Flowchart of the Development of Incomplete Data

Figure 3 shows the process of identifying, predicting, and replacing the missing values, in case the educators provide an incomplete data in the self-rating proficiency indicators. This will examine the items with no response from the present user and identify its corresponding domain and dimension. The method used for the data imputation is K-Nearest Neighbor, particularly the mode imputation technique wherein frequency of answers in each scale was obtained to get the most frequent answer. Then, the most frequent scale of the domain as well as the dimension was compared. In the case that they were equal, the most frequent scale of the domain was the chosen response for the missing item. If not, the individual proficiency of both domain and dimension in a certain scale was compared to obtain which is greater. In an instance that the domain is greater, it was the chosen response for the missing item. On the other hand, if the dimension is greater, its scale was the response for the missing value. Finally, the predicted response was imputed in the database.

III. Results and Discussion

Each method of data imputation was used on national proficiency data set with different percentage of missing data. This data set consists of 1,507 in size and 60 in dimension. In the following sections, the quality of imputation methods by getting each one of their accuracies was first shown. Next, classifying each imputation methods on downstream classification and regression tasks. Finally, the computational speed of each method was compared.

A. Imputation Accuracy for the Incomplete Data

Given the national proficiency data set with complete data, a fixed percentage of missing data was generated and was performed each of the imputation methods under the combined missing pattern. Each of the above methods generated a single set of imputed values. The imputation quality on the imputed set was evaluated by measuring how closely the imputed values resemble the ground truth values. In particular, the mean absolute error (MAE) between true and imputed values for each imputation method was calculated. Lower values indicate closer imputation, and perfect imputation corresponds to a MAE of zero. Another metric of imputation quality used was root mean squared error (RMSE). For each imputation method, the combination of parameters that achieves the lowest MAE in validation or RMSE was selected.

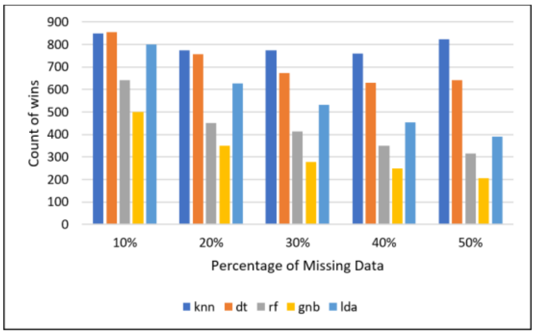

Fig. 4a. Counts based on lowest Mean Absolute Error (MAE)

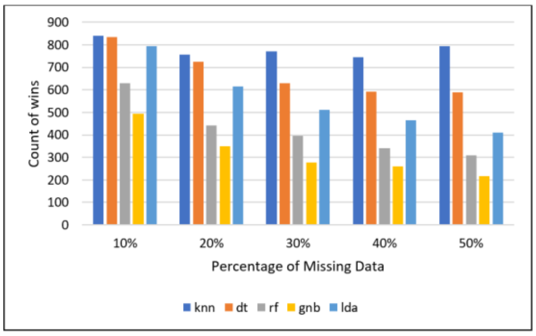

Fig. 4b. Counts based on lowest Root Mean Squared Error (RMSE)

In Figure 4a, the imputation method of different missing data percentages with winning counts was summarized. It showed the number of times that each method achieves the best overall imputation with lowest MAE and RMSE under five different missing data percentages. In all missing data scenarios, the K-Nearest Neighbors produced the best imputations of the data set according to both performance metrics.

In Figure 4b, the summary results of the MAE and RMSE values as geometric means across the national proficiency data set for each missing percentage were presented, with the confidence bands representing one geometric standard deviation multiplied above and divided below by the mean. Comparatively, K-nearest neighbor achieved the lowest average MAE and RMSE values for all missing percentages.

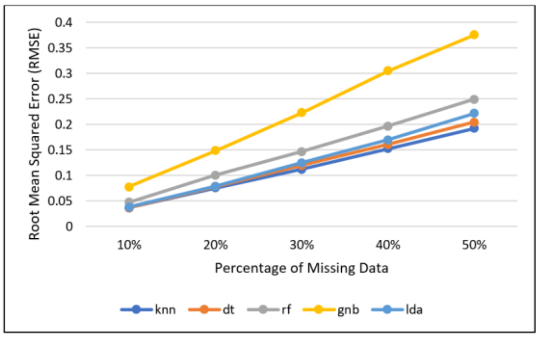

Fig. 5b. Average Root Mean Squared Error (RMSE)

Figure 5: (5a) Mean Absolute Error (MAE) and (5b) Root Mean Squared Error (RMSE) of national proficiency data set for each imputation method. The lines are geometric mean with one geometric standard deviation multiplied above and divided below by the mean. The x-axis corresponds to the percentage of missing entries.

B. Performance on Downstream Tasks

For each imputation, national proficiency data set was divided first using 50% training over testing split. Next was to randomly sample a fixed percentage of entries in data set to be missing completely at random ranging from 10% to 50%. For each missing percentage, the missing values was imputed in the training set and then was fitted standard machine learning algorithms to obtain a classification and regression model. The missing values will be imputed in the testing set by running the imputation methods on the full data set. For the regression tasks, cross-validated lasso and SVR models was fitted and was computed the out-of-sample accuracy on the imputed testing set. For the classification tasks, cross-validated SVM and optimal trees models was also fitted and was compute the out-of-sample R2 on the imputed testing set. [11]

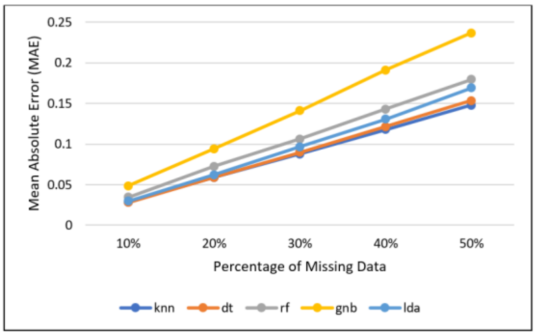

Fig. 5a. Average Mean Absolute Error (MAE)

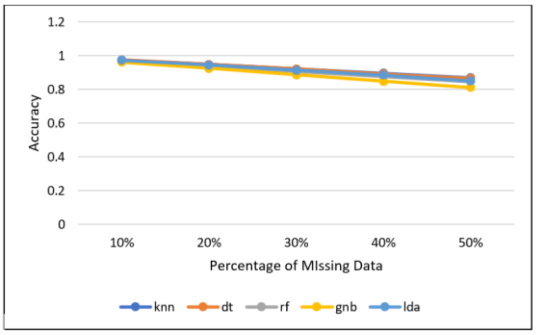

Fig. 6a. Average out-of-sample Accuracy values

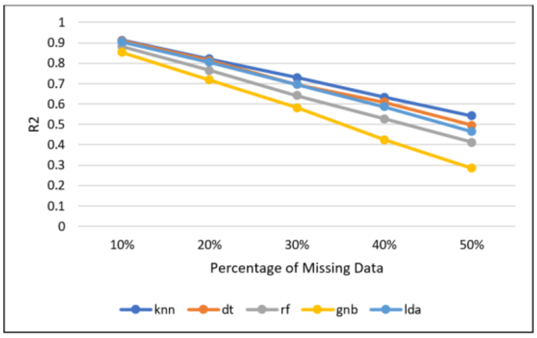

Fig. 6b. Average out-of-sample R squared values

Figure 6 shows the average out-of-sample performance, (6a) accuracy and (6b) R squared, of imputation method trained on data imputed via benchmark methods across a range of missing data percentages. These show how the imputation method chosen impacts the performance for classification and regression tasks, across the national proficiency data set and different missing data percentages. For each imputation methods, the improvement of k-nearest neighbor over all methods was statistically significant for all missing percentages in both classification and regression tasks. Moreover, this improvement in out-of-sample accuracy and R^2 was monotonically increasing with the missing percentage. At 50% missing data, the average improvement in out-of-sample accuracy is 2.63% for classification tasks, and the average improvement in out-of-sample R2 is 0.039 for regression tasks.

C. Computational Speed

Imputation Methods	Time per item (Seconds)				
	10 %	20 %	30 %	40 %	50 %
K-Nearest Neighbor	0.326507	0.478152	0.631176	0.742864	0.851657
Decision Tree	0.326885	0.474385	0.605735	0.711518	0.798641
Random Forest	0.537088	0.770505	1.071362	1.354010	1.636737
Gaussian Naive Bayes	0.322120	0.439977	0.557833	0.667972	0.769888
Linear Discriminant Analysis	0.509628	0.878600	1.105108	1.292778	1.480448

Table 1. Computational time comparison of different imputation methods

The computational time required for all imputation methods for different missing data patterns was compared. Each method was run with a limit of four (4) hours and the average time per run are recorded as shown in Table 1. For small percentage of missing data, all imputation methods finish quickly with a span of less than 1 second per item. As the size of missing data percentage increases, the time required to finish per item increases with a maximum span of 1.5 sec. Among the method used, as can be seen in Table 1, Gaussian Naïve Bayes method has the fastest time to finish per item followed by K-Nearest Neighbor.

IV. Conclusion

In this study, a Python-based instrument was first made to perform the same task as the Fortran-based instrument for the completed data. On the subject of incomplete data set, several algorithms were employed and compared to yield high accuracy and performance. The algorithms were K-Nearest Neighbor, Decision Tree, Random Forest, Gaussian Naive Bayes, and Linear Discriminant Analysis. K-Nearest Neighbor achieved the lowest average MAE and RSME values for all missing percentages. Subsequently, on all imputation methods used, the improvement of k-nearest neighbor over all methods was statistically significant for all missing percentages in both classification and regression tasks. Although Gaussian Naive Bayes has the fastest time to finish per item, K-Nearest Neighbor was utilized on the instrument to classify the proficiency of the educator for incomplete data set.

References

[1] J. C. Anito and M. P. E. Morales, “The pedagogical model of Philippine steam education: Drawing implications for the reengineering of Philippine steam learning ecosystem,” *Univers. J. Educ. Res.*, vol. 7, no. 12, pp. 2662–2669, 2019.

[2] M. P. E. Morales, J. C. Anito, R. A. Avilla, E. L. R. Abulon, and C. P. Palisoc, “proficiency_indicators_for_Philippine_STEAM_,” vol. 148, no. June, pp. 263–275, 2019.

[3] S. Z. Christopher, T. Siswantining, D. Sarwinda, and A. Bustaman, “Missing Value Analysis of Numerical Data using Fractional Hot Deck Imputation,” *ICICOS 2019 - 3rd Int. Conf. Informatics Comput. Sci. Accel. Informatics Comput. Res. Smarter Soc. Era Ind. 4.0, Proc.*, 2019.

[4] S. J. Hadeed, M. K. O’Rourke, J. L. Burgess, R. B. Harris, and R. A. Canales, “Imputation methods for addressing missing data in short-term monitoring of air pollutants,” *Sci. Total Environ.*, vol. 730, p. 139140, 2020.

[5] C. P. Palisoc, M. P. E. Morales, R. A. Avilla, T. O. Ayuste, B. Ramos-Butron and N. A. Casilla, *"Developing STEAM Educators' Proficiency Scoring Framework"*.

[6] T. Aljuaid and S. Sasi, “Intelligent imputation technique for missing values,” in *2016 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016*, Nov. 2016, pp. 2441–2445.

[7] D. V. Patil and R. S. Bichkar, “A hybrid evolutionary approach to construct optimal decision trees with large data sets,” *Proc. IEEE Int. Conf. Ind. Technol.*, pp. 429–433, 2006.

[8] H. Lan and Y. Pan, “A crowdsourcing quality prediction model based on random forests,” *Proc. - 18th IEEE/ACIS Int. Conf. Comput. Inf. Sci. ICIS 2019*, pp. 315–319, 2019.

[9] A. H. Jahromi and M. Taheri, “A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features,” *19th CSI Int. Symp. Artif. Intell. Signal Process. AISP 2017*, vol. 2018-January, no. 1, pp. 209–212, 2018.

[10] S. Xie and Y. Feng, "A Recommendation System Combining LDA and Collaborative Filtering Method for Scenic Spot" *2015 2nd International Conference on Information Science and Control Engineering*, 2015.

[11] D. ertsimas, C. Pawlowski, and Y. D. Zhuo, "From Predictive Mehods to Missing Data Imputation: An Optimization Approach", *The Journal of Machine Learning Research (JMLR)*, January 2017.

Bryan D. Brillo finished his studies in elementary and high school at Tinabunan Elementary School and Emiliano Tria Tirona Memorial National High School, respectively. Later on, he finished a degree in Bachelor of Science in Electronics Engineering major in ICT at Technological University of the Philippines – Manila on 2020.

Jesmar R. Enerio finished his studies in elementary and high school at St. Therese of Lisieux School and Divine Light Academy, respectively. Later on, he finished a degree in Bachelor of Science in Electronics Engineering major in microelectronics at Technological University of the Philippines – Manila on 2020.

Rodnick Jose G. Prudente finished his studies in high school at Elizabeth Seton School and Cavite School of Life – Bacoor. He later finished a degree in Bachelor of Science in Electronics Engineering major in microelectronics at Technological University of the Philippines – Manila on 2020.

Eloisa Mae B. Ramirez finished her studies in elementary and high school at Ligas 1 Elementary School and Cavite School of Life – Bacoor, respectively. Later on, she finished a degree in Bachelor of Science in Electronics Engineering major in microelectronics at Technological University of the Philippines – Manila on 2020.

Chester Keith A. Rivera finished his studies in elementary and high school at Las Piñas Elementary School Central and Las Piñas National High School, respectively. He later finished a degree in Bachelor of Science in Electronics Engineering major in ICT at Technological University of the Philippines – Manila on 2020.