NSS
Neutrosophic Sets and Systems, Vol. 81, 2025
University of New Mexico

Enhancing Missing Data Imputation for Migrants Data: A Neutrosophic Set-Based Machine Learning Approach

Doaa A. Abdo [1*] , A. A. Salama[2] , Alaa A. Abdelmegaly[3] , Hanan Khadari Mahdi Mahmoud[4]

[1]Applied statistics and insurance department, faculty of commerce, Mansoura university, Mansoura, Egypt, doaaashour@mans.edu.eg

[2]Dept. of Math and Computer Sci., Faculty of Science, Port Said Univ., Egyptdrsalama44@gmail.com, ahmed_salama_2000@sci.psu.edu.eg.

[3]Higher Institute of Advanced Management Sciences and Computers, Al-Buhayrah, Egypt bintmasr880@yahoo.com

[4]Depatrment of applied statistics at the Nile Higher Institute of commercial sciences and Computer Technology in Mansoura, hananhkodary@gmail.com.

* Correspondence: doaaashour@mans.edu.eg

Abstract: This study tackles the problem of missing data in migrant datasets by introducing a new framework that combines machine learning techniques with neutrosophic sets. These sets, which can represent uncertainty and ambiguity, are well-suited for managing the complex nature of missing information in sensitive fields like migration research. We test the effectiveness of KNN, SVM, decision tree, random forest, and Ada Boost algorithms on a migrant dataset, comparing their results using different imputation methods (mean/mode, model-based imputer (simple tree), and random values). Our research showed that our proposed approach, which used neutrosophic sets, improved imputation accuracy and strengthened model reliability. Our results underscored the potential of neutrosophic set-based machine learning for addressing missing data issues across various fields.

Keywords: Missing data imputation, neutrosophic sets, machine learning, migrant data, KNN, SVM, decision tree, random forest, Ada Boost, classification, accuracy, precision, recall, F1-score.

1. Introduction

In statistics, missing data, also known as missing values, occurs when no data value is saved for a variable in an observation. Incomplete data is a widespread issue in primary care studies, including clinical trials, observational research, and quality improvement projects. It refers to data points that are not available for analysis, such as migrants who drop out. Missing data is prevalent and can have a substantial impact on the conclusions formed from the data. Missing values in research can lead to bias, reduced validity, inaccurate conclusions and the loss of crucial information from study samples. When missing data is not correctly and handled improperly, it might result in statistical bias and hide the underlying relationship between variables. Missing data might lead to loss of knowledge, affecting study efficiency and interpretation. Also, missing data might reduce the validity and trustworthiness of research findings by increasing bias, causing information loss, and reducing statistical power (Suthar and Goswami, 2012; Hourarip et al., 2014; Ayilara et al., 2019).

The effect of missing values depends on how much data is missing, what type of data is missing, and why it's missing (Zhang, 2015). The most common methods which dealing with missing data is delectation missing values and using the mean to fill in missing values. This method resulted in biased estimation of parameters and uncertainty, and decreasing statistical power.

Machine learning methods can overcome the shortcomings of traditional approaches like mean, median, and regression imputation (Langkamp et al., 2010; Donders et al., 2006). Assessing these methods' performance requires careful planning and analysis. Factors like algorithm choice and sampling techniques are key. Therefore, selecting the right strategy for handling missing data is crucial, as improper treatment can lead to inaccurate results. Recently, advanced analysis models like neutrosophic models have gained popularity in studying the link between migration and missing persons. Neutrosophic statistics offers a robust framework for dealing with various uncertainties in statistical analysis, making it a powerful tool for predicting intervals in machine learning (Smarandache, 2022).

Neutrosophic machine learning combines neutrosophic theory, introduced by Florentin Smarandache, with machine learning algorithms to handle uncertain and incomplete data. While traditional machine learning assumes accurate and complete datasets, real-world data often contains ambiguities and gaps that can affect model performance. Neutrosophic methods offer a novel approach by directly incorporating uncertainty into the learning process, potentially improving model accuracy, flexibility, and interpretability. Neutrosophic sets provide a mathematical framework for representing information beyond simple true-false values, making them a powerful tool for managing uncertainty and indeterminacy (Salama et al., 2024). By integrating neutrosophic theory into machine learning algorithms, we can enhance model performance when dealing with missing data. Jdid et al. (2023, 2022) have explored the basics of neutrosophic simulation and its potential uses in handling randomness and uncertainty.

These models show promise in fields where data is often imperfect. However, implementing neutrosophic techniques can be complex, requiring significant customization of existing algorithms. Despite these challenges, neutrosophic machine learning has the potential to improve outcomes in various applications by better handling uncertainty and missing data. It represents a new direction in how models treat uncertainty and imprecision. While it currently faces obstacles and needs further validation, its ability to provide a richer and more flexible representation of data shows promise for increasing the accuracy and applicability of machine learning models across different fields (Maguina, et al., 2024).

Several studies have addressed missing data issues. For example, (Petersen et al., 2019) examined key implications of handling missing data using health indicator records in the UK. (Stiglic et al., 2019) highlighted challenges in electronic health records, demonstrating missed opportunities from incomplete data through simulations in predictive modeling. Their approach to missing data involved removing incomplete records. The rest of this study is designed as follows, in section 2, we proposed the related works, section 3 proposed the methodology of this study, in section 4, we presented missing data mechanisms, section 5, introduced machine learning algorithms, section 6, proposed neutrosophic machine learning and finally section 7, numerical analysis of this study.

2. Related Works

There are many studies related to this study like, Power and Freeman (2012) compared interpersonal psychotherapy (IPT) and cognitive-behavioral therapy (CBT) in primary care. Their study, which examined three treatments, encountered significant missing data. They suggested that complex imputation methods might offer the best estimates. Mirzaei et al. (2022) investigated handling missing survey data using various approaches, including deletion, calculation, and likelihood methods, providing guidance on their application. Nijman et al. (2022) applied machine learning to tackle missing data in model prediction, finding that deletion, complete-case analysis (CCA), and multiple imputation were most effective. Zhou et al. (2022) explored how missing data impacts comparative effectiveness research using electronic health records (EHRs), concluding that spline smoothing yielded results similar to having complete data. This study aims to employ neutrosophic set under machine learning to treat missing data issue.

3. Methodology

3.1. Data Collection and Processing

- **Dataset Acquisition:** A migrants dataset, either publicly available or proprietary, was obtained, containing relevant demographic, socioeconomic, and migration-related information.
- **Data Refinement:** The dataset underwent thorough cleaning to eliminate inconsistencies, errors, and outliers. Missing values were identified and classified based on their mechanisms like, missing completely at random (MCAR), missing at random (MAR), and missing not completely at random (MNAR).

3.2. Neutrosophic Set Application

Neutrosophic sets (NSs) introduced by (Smarandache,1998) which represented a generalization of fuzzy sets and intuitionistic fuzzy set, is a powerful tool to deal with incomplete, indeterminate and in consistent information, which exist in the real-world. Neutrosophic sets are determined by truth membership function (T), Indeterminacy membership function(I) and falsity membership function(F). This theory is very essential in various applications areas since indeterminacy is quantified explicitly and the determined memberships functions are independent. Wang, Smarandache, Zhang & Sunderraman (2010) proposed the concept of single – valued neutrosophic set. The single-valued neutrosophic set can convey truth-membership degree, indeterminacy-membership degree, and falsity-membership degree, addressing incomplete, indeterminate, and inconsistent data. The single-valued neutrosophic set accurately describes human thought due to the inadequacy of external knowledge.

- **Attribute Representation:** Each dataset attribute was represented as a neutrosophic set, assigning degrees of truth, falsity, and indeterminacy to possible values.
- **Distance Measures:** Appropriate measures (e.g., Hamming distance, Euclidean distance) were defined to compare neutrosophic sets, accounting for uncertainty and indeterminacy.

3.3. Machine Learning Algorithm Choice

- **Selected Algorithms:** The following algorithms were chosen based on their ability to handle missing data and incorporate neutrosophic sets:
  - K-Nearest Neighbors (KNN)
  - Support Vector Machines (SVM)
  - Decision Trees
  - Random Forest
  - AdaBoost
- **Algorithm Modification:** These algorithms were adapted to process neutrosophic data by incorporating neutrosophic distance measures or adjusting their decision-making processes.

3.4. Imputation Strategies

- **Conventional Imputation:** Standard imputation methods were employed as benchmarks, including:
  - Mean imputation
  - Median imputation
  - Multiple imputation
- **Neutrosophic Imputation:** An innovative imputation approach was developed, utilizing neutrosophic sets to fill in missing values based on the closest neighbors in the neutrosophic attribute space.

3.5. Model Training and Assessment

- **Model Training:** The chosen machine learning algorithms were trained on the imputed dataset using suitable training methods (e.g., cross-validation).
- **Model Assessment:** The models' performance was evaluated using common classification metrics, such as accuracy, precision, recall, and F1-score.
- **Comparison:** The effectiveness of the neutrosophic-based imputation method was compared to traditional imputation techniques to gauge its efficacy.

3.6. Sensitivity Analysis

- **Parameter Adjustment:** The sensitivity of the results to various algorithm parameters and imputation methods was examined.
- **Missing Data Patterns:** The effect of different missing data mechanisms (MCAR, MAR, MNAR) on model performance was explored.

3.7. Ethical Considerations

- **Data Protection:** Appropriate steps were taken to ensure data privacy and confidentiality throughout the study.
- **Fairness and Bias:** The potential for bias and unfairness in the dataset and models was addressed, with measures implemented to mitigate these issues.

4. Missing Data Mechanisms

The nature of missing data varies across studies based on the underlying mechanism of missingness. These mechanisms define the patterns of missing data. The three common types of missing data patterns are univariate, monotone, and non-monotone (Little and Rubin, 2019). Univariate patterns occur when missingness is related to a single variable (Demirtas, 2019), often seen in experimental studies (Lacerda et al., 2007). Monotone patterns are typically observed in ordered variables, such as those in longitudinal studies (Liu, 1995). Non-monotone patterns are the simplest to handle as missing values are easily identified through patterns (Dong and Peng, 2013). In this case, the missingness of one variable does not influence the missingness of others (Chen, 2020).

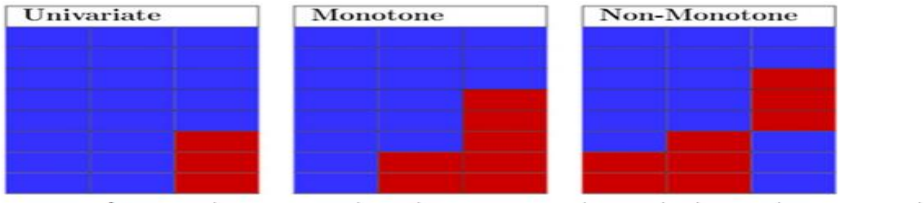


Figure 1: Missing data patterns.

The assumptions underlying missing data treatment methods are influenced by the mechanisms leading to data missingness. Therefore, it is crucial to understand these mechanisms. The theory of missing data, established by Rubin (1976), identifies three main mechanisms for missingness, which are determined based on the availability and absence of data.

To define the missing. Let Y be an entire data matrix, which separated into Y_0 observed data and Y_m missing data. Let R is a matrix of missing value known by,

$$R = \begin{cases} 0, & if\ Y\ is\ observed \\ 1, & if\ Y\ is\ missing \end{cases} \quad (1)$$

Let q a values vector, which determine the relation between the missing in R and the data set Y. The mechanisms of missing values defined by the probability of whether a value observed or missing as it displays below.

a) Missing at Random (MAR)

In MAR, the probability of missing data is solely linked to the observed data. This probability in MAR can be expressed as:

$$p(R|Y_0, q) \quad (2)$$

Missing at random (MAR) mostly appeared in medical studies data sets. The missing in this mechanism can be treated by observed predicator variables (Gomez et al., 2014).

b) Missing Completely at random (MCAR)

Occurs when the occurrence of missing data is unrelated to both observed and unobserved measurements. The probability of missing completely at random (MCAR) is represented as:

$$p(R|q) \quad (3)$$

c) Missing not at random (MNAR)

This situation occurs when the missing data is independent of the previous two mechanisms, and the missing values depend equally on both the missing and observed values. Dealing with missing data using this approach is challenging due to its connection with unobserved data.

The probability of MNAR is defined as: -

$$p(R|Y_0, Y_m, q) \quad (4)$$

The probability R depends on both Y_0 and Y_m and this mechanism is applied in various sectors like biomedicine in data sets of psychology and education (Grittner et al., 2011).

5. Machine Learning Algorithms.

The approaches of missing data vary from traditional to improved techniques of machine learning. As the imputation of traditional statistical methods has a bias at the final analysis of missing data. Therefore, this study introduces some machine learning methods for handling missing data problems. Such learning methods allow us to find challenging associations in data sets where exploratory analysis has failed to accurately determine the form of the underlying model. We do not want to provide an explicit formula for the distribution of our data; instead, we want the algorithm to figure out the pattern on its own, based on the facts. Learning methods allow us to challenge the assumptions underlying statistical methodology. Algorithms of machine learning imputation are used to develop a predictive technique to treat missing data. In addition, the estimation process of these methods relies on the availability of information from observed data using labeled or unlabeled data. The machine learning techniques are discussed below.

(5-1) K nearest Neighbor Classification

The algorithm of KNN conducts by classifying the nearest neighbors of missing values and using those neighbors for imputation using a distance measure between instances (Maillo et al., 2017). Different distance like, Euclidean distance, distance of cosine, Jaccard distance and distance of Minkowski can be used for imputation of KNN. However, the distance of Euclidean is the most distance measure used widely.

The imputation of KNN using distance of Euclidean distance can be written as follows:

$$Dist_{xy} = \sqrt{\sum_{k=1}^{m} \left(X_{ik} - X_{jk}\right)^2} \quad (5)$$

Where:-

- $\boldsymbol{Dist_{xy}}$: is the Euclidean distance.
- $\boldsymbol{k}$: data dimensions
- $\boldsymbol{X_{ik}}$: value for j .attribute involving missing data.
- $\boldsymbol{X_{jk}}$: is the value of j. Attribute including complete data.

The $\boldsymbol{k}$ points that have a minimum distance value are chosen, and then the estimation of weight mean is imputed as:

$$X_k = \frac{\sum_{j=1}^{J} W_j V_j}{\sum_{j=1}^{J} W_j} \quad (6)$$

Where:-

- $\boldsymbol{X_k}$: is the estimation.
- $\boldsymbol{J}$: is the parameters number with $j = 1,2,3, \ldots \ldots, k$.
- $\boldsymbol{V_j}$: are complete values on attributes including missing data.
- $\boldsymbol{W_j}$: is the observed nearest neighbors.

Hence, the weighted value given by the following equation:

$$W_j = \frac{1}{dis_{(x,y)^2}} \quad (7)$$

In both discrete and continuous data, The KNN calculation technique is used also, conducted as a handler of multiple missing data (Suthar et al., 2012 ; Mailo et al., 2017).

(5-2) Support Vector Machine (SVM)

This method represents the most used method for handling missing data (Honghai et al., 2005 ; Pelckmans et al., 2005). SVM discovers an optimal separating hyper- plane for a sample of a labeled training like the distance between the hyper- plane and the nearest data points is maximized (Stewart, 2018). The hyper- planes are known by:

$$w.\ x_1 + b \geq +1 \quad when \quad y_i = +1 \quad (8)$$

$$w.\ x_1 + b \leq -1 \quad when \quad y_i = -1 \quad (9)$$

Where:-

- $\boldsymbol{w}$: is a vector of weight.
- $\boldsymbol{x}$: is a vector of input.
- $\boldsymbol{b}$: is a bias.

Different researchers applied SVM in various studies. For instance, (Hong et al.,2005) handling missing data by using the SVM regression method. They concluded that this method gave precise results on the data set of SARS. In addition, (Smola et al., 2005) the authors treating missing data using SVM and Gaussian processes by using exponential families in feature peace. In another research (Ghazanfar and Prugel, 2013). The authors exchanged the missing values by using the findings assesses from applying the SVM classifier through the set of training and used SVM regression to treat the values.

(5-3) Decision tree

Describing a machine-learning algorithm that outlines all potential outcomes and the pathways to those outcomes in a tree structure format. This approach relies on constructing decision trees to analyze missing values for each variable, filling in these gaps by referencing the respective tree (Twala, 2009). Predicted missing values are displayed in the leaf nodes. Notably, this algorithm handles numerical and categorical variables with

equal effectiveness. Decision trees offer a lower bias in estimation compared to other methods, despite requiring a substantial amount of time for computation (Rockach, 2016).

(5-4) Random forest

This method enhances decision tree bagging by reducing correlation between trees. Similar to bagging, it involves training M decision trees on bootstrap samples. However, during tree construction, each split considers a random subset of predictors, allowing changes based only on this subset. In a random forest, each split restricts the use of most available predictors. The number of predictors used at each split, m, is typically the square root of the total number of predictors, p, rounded up:

$$m \approx \sqrt{p} \qquad (10)$$

It is necessary to utilize a small value of m when establishing a random forest in the presence of a large number of correlated predictors. The strategy of constructing a random forest, which containing$N$ trees, is as follows: -

1) Generate a bootstrap X_n for each tree.
2) Build each tree T_n on the sample X_n
3) At each tree split, the best predictor is selected from a random subset based on criteria like entropy or the Gini index. The tree continues to grow until its leaves contain no fewer than a specified minimum number of objects or the maximum tree depth is reached. This process results in a model (Boyko and Dypko, 2023):-

$$\widehat{f_{rand}}(x) = \frac{1}{M}\sum_{m=1}^{B} \hat{f}^m(x) \qquad (11)$$

(5-5) Ada Boost :-

This algorithm proposed by Frennd and Schapire in (1997) revolutionized ensemble modeling. This technique is extensively employed to tackle binary classification tasks. By converting numerous weak learners into formidable, strong learners, this potent algorithm significantly boosts prediction accuracy. The algorithm operates by initially constructing a model on the training dataset and subsequently creating a second model to rectify errors from the initial model. This iterative process continues until errors are minimized, ensuring accurate predictions on the dataset.

6. Neutrosophic Machine Learning:-

Neutrosophic statistics represents an extension of interval statistic, provides a robust framework for handling various indeterminacies in statistical analysis (Smarandache, 2022). The main concern of neutrosophic statistics is used for the analysis of the uncertainty observation data. Neutrosophic machine learning is an emerging branch that merges neutrosophic theory, introduced by Flor-entin Smarandache (Smarandache, 2022 & Smarandache, 1999) with algorithms of machine learning to handle uncertainty, imprecision and Missing data. Traditional machine learning supposes accurate and complete data, but data often includes ambiguities and gaps that can impact on model performance. Neutrosophic provides a novel approach by incorporating uncertainty .Directly into the process of learning, enhancing model accuracy, flexibility, and interpretability.

These models are useful in different fields where data is often imperfect. Although the complexity of implementing neutrosophic techniques because it requiring significant customizing of existing algorithms. Despite these challenges, neutrosophic machine learning has the potential to improve outcomes in various applications by better handling uncertainty and missing data. So, neytrosophic machine learning presents a new progression in the way models treat uncertainty and imprecision. However it currently faces hurdles and needs more validation, it is ability to give a richer and more flexible representation of data displays increasing of the accuracy and applying of machine learning models in different fields (Maguina, etal.,2024)**.**

To develop models that can learn from data to make prediction, machine learning uses mathematical formulations without being explicitly programmed to accomplish such tasks (Shinde and Shah, 2018).Interval prediction in machine learning refers to the approach of predicting a range of probable outcomes for a given input rather than a single point estimate. By offering intervals, these approaches give not just forecasts but also insight into the dependability and uncertainty of the predictions, which is vital for decision making in uncertain contexts.

For a data set with independent variables $X = [x_1, x_2, x_3, \dots., x_n]$ and a dependent variable, the main objective of regression analysis is to model the relationship between X and Y accurately.

This relationship is expressed mathematically as:

$$y \approx f(x; \theta) \qquad (12)$$

Where:-

- $\boldsymbol{y}$: is the dependent variable or the objective to be predicted.
- X: represents the independent variables that are used to predict
- f: Is the regression function, which can vary in shape depending on the type of regression model used (Linear, Polynomial, Logistic, etc).
- θ: are the parameters or coefficients of the model, adjusted during the training process to minimize a loss function, typically the mean square error(MSE) in regression (Anwar etal., 2024). For instance, the prediction interval for a new observation in a simple linear regression is given by:

$$\hat{y}_0 \pm t_{\frac{\alpha}{2}, n-2}.SE \qquad (13)$$

Where:-

- $\boldsymbol{\hat{y}_0}$: is the predicted value of y from the t distribution, for a specific confidence level $\propto$ and $n-2$ degrees of freedom.

Converting the traditional intervals into a neutrosophic number is enhanced to contain a component of indeterminacy. This addition captures the uncertainty and imprecision that is typically present in real- world data, offering a more nuanced understanding of data variability.

The interval neutrosophic treatment is as follows:-

$$\hat{y}_0 \pm t_{\frac{\alpha}{2}, n-2}.SE + \left(\hat{y}_0 \pm t_{\frac{\alpha}{2}, n-2}.SE\right) I_N \qquad (14)$$

I_N refers to the indeterminacy factor associated with the prediction. Where $I_N \in [I_l, I_u]$.

Where:-

- I_l: lower indeterminacy.
- I_u:upper indeterminacy.

Which define the range of possible deviations due to uncertain elements that affect the forecast (Delgado, etal., 2024). Neutrosophic techniques increasing the robustness and reliability of prediction models by presenting a more complete framework that takes into account many types of uncertainty. The neutrosopjic mean denoted as X_n, is calculated considering the neutrosophic inclusion I_N that connects to the interval $[I_l, I_u]$. This mean consists of two main elements X_l refers to the mean of the bottom part of the neutrosophic samples, and X_u is the mean of the top part.

The respective definitions are (Castro, etal., 2021):-

$$X_l = \frac{\sum_{i=1}^{n_l} X_{il}^1}{n_l} \quad (15)$$

$$X_u = \frac{\sum_{i=1}^{n_u} X_{iu}}{n_u} \quad (16)$$

Where n_l and n_u refers to the number of elements in the lower and upper parts of the neutrosophic samples respectively. Therefore, the neutrosophic mean X_n is defined as the sum of

X_l and X_u, handled by the interval of indeterminacy I_n(Sanchez etal., 2021).

$$X_N = X_l + X_u I_N; \quad I_N \in [I_l, I_u] \quad (17)$$

Where:-

- $I_l = 0$ and ,
- $I_u = \frac{X_u - X_l}{X_u}$

7. Numerical Analysis: -

This research utilized data obtained from the Missing Migrants Project, sourced from the International Organization for Migration. Specifically, the data originates from the Missing Migrants Project, a distinct initiative that monitors the deaths of migrants, including refugees, who have disappeared along complex migration paths globally, spanning from 2014 to June 2017. The dataset comprises 2420 individuals and was obtained from an open online repository available on the Missing Migrants and the dataset is publicly accessible at Global Missing Migrants Dataset (kaggle.com). The primary objective of this study revolves around handling missing data and determine the most appropriate machine learning algorithm for this dataset type, focusing on some accuracy metrics.

(7-1) Data set description: -

The Missing Migrants Project documents individuals who have perished or disappeared while attempting migration across international borders or en route to foreign destinations. This tally does not encompass fatalities within immigration detention centers, during deportation, or following the forced return of migrants to irregular statuses, such as those stemming from labor exploitation. Moreover, it excludes migrants who pass away or vanish after settling in a new residence, thus excluding deaths occurring in refugee camps or housing facilities.

The data available offers insights into evolving conditions and trends concerning migration paths and the individuals traversing them, providing valuable information for policy formulation and protective strategies. This dataset can aid in assessing the comparative risks associated with irregular migration routes. It comprises 10 predictor variables, including ID, cause of death, region of origin, nationality, missing persons, deceased individuals, incident region, date, latitude, and longitude. These variables play crucial roles in- depth analysis and discussion of numbers of missing and dead migrants around the world and the challenges involved in identification and tracking. But this analysis relied on eight variables that influence missing or death as displayed in table 1. Detailed description for the features of the missing and dead migrants are provided in table1.

Table1: Description of variables

Variable	Description
Id	Unique key documenting incident
Region of origin	Horn of Africa, null and other.
Missing persons	Counts (number of missing persons)
Dead	Counts (number of deaths)
Incident region (Source)	Region where incident was recorded.
Reliability	Verified, partially verified.
Latitude	Spatial coordinates
Longitude	Spatial coordinates

Table 1 introduced features description of eight variables for 2420 individuals. Firstly, the **id** variable represented numerical variable, and the key recorded for each incident. **The region** of individuals decomposed in three areas horn of Africa with percentage 20%, the percentage of persons with non-region was 18% and persons who belonged to other areas represented 62% of the total number of persons. The **incident – region** identified two places of missing or death north Africa contained 26% of instances, Mediterranean included 24% and the rest of percent (50%) represented other areas. **The source of** recording the states contained regional mixed, this source recorded 27% of migrants, 8% of people recorded by Pima county office of the medical examiner and the other sources recording 65% of people. Finally, **reliability,** the percent of verified individuals was 33% while partially verified involved 47% and the other percent 20%.

Table2: Feature statistics under missing data.

Variable	Mean	Mode	Median	Dispersion	Minimum	Maximum	Missing
Region- origin		Home of Africa(p)		2.19			443(18%)
Missing	39.66		10	2.12	0	750	2149(89%)
Dead	4.73	1	1	4.34	0	750	102(4%)
Incident – region		North Africa		1.91			10(0%)
Lat.	26.9014	12.5331	29.3489	0.428083	-26.2245	66.9672	4(0%)
Lon.	-13.9959	1.85869	14.4711	-4.40709	-117.071	116.225	4(0%)
Reliability		Partially verified		**0.908**			**324(13%)**
Id	95926.26	1	121178	0.65	1	184750	0(0%)

Table 2 presented the statistics of predicators using Orange software, under this study like mean, mode, median, dispersion, minimum, maximum value and missing number and percent of each variable. The highest variable of dispersion was missing. This proved that there were differences and variation between these variable values with a mean of 39.7 affected by missing observations. The less dispersion variable was dead variable this emphasized on symmetrical between its observation. This study depended on partially verified persons, which represented 13% from all instances with high dispersion equals 0.908. The number of missing migrants and the percentage of them through the covering period. The results showed that the highest percentage was for missing variable with 89%, then region origin with 18% and finally reliability with 13%. This variation in missing percentage relied on the number of missing persons to the total number of instances.

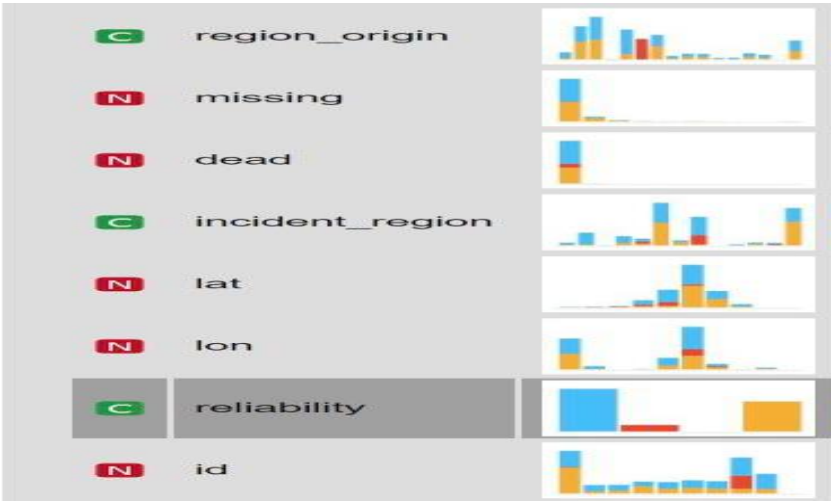


Figure 2: Distribution of study variables through missing data

Figure 2 proposed study predictors distributions. The figure displayed that the reliability of partially verified have a normal distribution, while missing is a right skewed variable, this emphasized on that the most extreme

values are on the right side and the mean or average is greater than the median because of the presence of missing values of some migrants.

The main objective of this study is to treat the missing values of people who are missing or partially verified in migration. This study depended on three methods to replace missing values they are , average/ most frequent method , a method of random values and model based imputer (simple tree). Also, using four precision measures to evaluate these methods under machine learning algorithms involving classification accuracy (CA), F_1 – score, precision and Recall. Classification accuracy represents the ratio of the number of correct predictions to the total number of input samples.

While F_1 is the weighted average of the accuracy and recall rates close to 1 indicating higher accuracy levels, F_1 – score equals: -

$$= \frac{2 \times precision \times Recall}{precision \times Recall} \qquad (18)$$

On the other hand, Recall or true positive rate (TPR) measures the percentage of relevant data points that were correctly identified by the model, represents a metric that measures how machine learning model correctly identifies positive instances. TPR can be written as follows: -

$$\frac{TP}{TP + FN} \qquad (19)$$

Finally, precision is one indicator of a machine learning model's performance- the quality of a positive prediction made by the model. Precision refers to the number of true positives divided by the total number of positive predictions (i.e. the number of true positives plus the number of false positives as follows: -

$$precision = \frac{Tp}{Tp + Fp} \qquad (20)$$

Table 3: Confusion matrix

		Predicted: Yes	**Predicted: No**
		1	0
Actual: Yes	1	TP	FN
Actual: No	0	FP	TN

Table 4: The effectiveness of replacing missing data under random values method

Model	**CA**	**F_1**	**Precision**	**Recall**
Tree	**0.924**	0.518	0.541	0.497
Logistic Regression	0.901	0.269	0.344	0.221
Naïve Bayes	0.869	0.539	0.379	**0.930**
SVM	0.898	0.283	0.333	0.246
KNN	0.905	0.346	0.396	0.307
Random forest	**0.919**	0.522	0.507	0.538

Ada boost	**0.919**	0.512	0.507	0.518

Table 4 proposed the imputation of random values, which is used to substitute missing values, there by constructing a data set with non-missing. Which this approach is straight forward, it doesn't utilize any information from the data set and may present randomness that affects subsequent analyses (Hui, et al., 2023). The precision measurements of machine learning under missing data issue introduced. The findings showed that the Recall measure for naïve Bayes has the highest precision with ratio 93%, then classified accuracy for tree algorithm, and after that random forest and ad boost have the same value under the same measure.

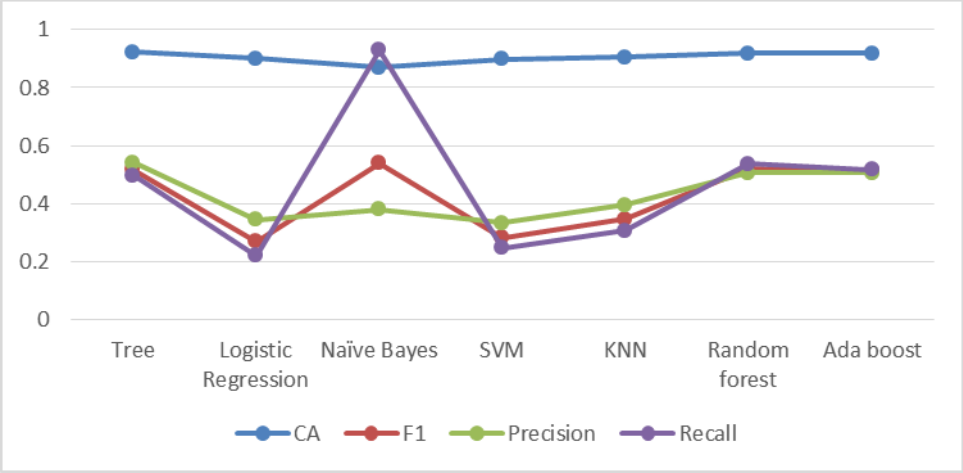


Figure 3 : Comparison of Machine Learning Algorithms for Missing Data Imputation and Classification (Random Values Method)

Figure 3 introduced (CA, F_1, precision and Recall) measures for proposed machine learning algorithms under random values method. The previous figure showed how these measures varies through different algorithms introduced. This figure displayed that the Naïve Bayes algorithm is the best for Recall measure with percent 93%, then the three algorithms of (tree, Ada boost and random forest) are the best for classified accuracy (CA) measure with percent (0.924, 0.919, 0.919) respectively.

Table 5: Handling missing data using average/ most frequent method

Model	CA	F_1	Precision	Recall
Tree	0.924	0.431	0.452	0.412
Logistic regression	**0.929**	0.023	0.333	0.012
Naïve Bayes	0.844	0.414	0.281	0.782
SVM	0.854	0.314	0.234	0.476
KNN	0.918	0.261	0.357	0.206
Random forest	**0.929**	0.387	0.495	0.318
Ada boost	0.924	0.454	0.461	0.447

Table 5 exhibited average/ most frequent method for fixing missing data. This method counted on replacing missing entries with the average (mean) or most frequent value(mode). This is quick and easy approach, but it

can introduce bias if the missing data is not randomly distributed (Tamboli, 2024). This study built on missing completely at random, hence the observations followed normal distribution, so, it reckoned on mean to substitute missing data. The findings clarified that logistic regression and random forest were the best methods under this technique according to CA measure with proportion 92.9.

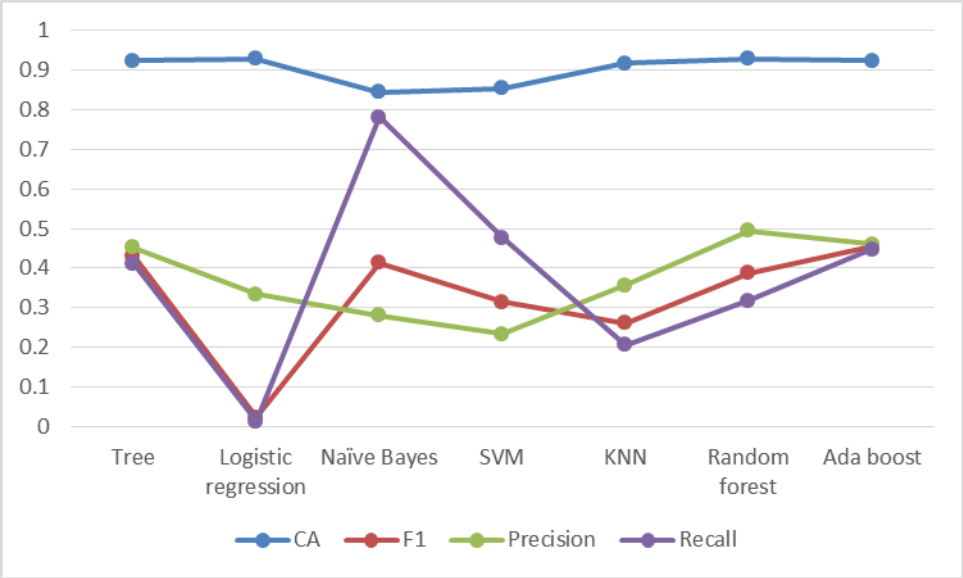


Figure 4: Comparison of Machine Learning Algorithms for Missing Data Imputation and Classification (Average/Most Frequent Method)

Figure 4 illustrated four different measures of accuracy to compare between (tree, logistic regression, Naïve Bayes, SVM, KNN, random forest and Ada boost) algorithms based on average/ most frequent enhancing missing data method. The figure concluded that according to classified accuracy measure, the logistic regression and random forest machine learning algorithms have the highest percent of accuracy with ratio(92.9).

Table 6: Treating missing data by model-based imputer (simple tree) through machine learning methods

Model	CA	F_1	Precision	Recall
Tree	0.922	0.420	0.442	0.400
Logistic regression	**0.928**	0.022	0.250	0.012
Naïve Bayes	0.845	0.420	0.285	0.800
SVM	0.874	0.278	0.232	0.341
KNN	0.918	0.250	0.351	0.194
Random Forest	**0.926**	0.382	0.466	0.324
Ada boost	**0.926**	0.464	0.470	0.459

Table 6 illustrated using model based imputer (simple tree) for replacing missing values. This methods strategr depended on constructs a model for predicting the missing values based on values of other attributes, a separate

model is constructed for each attribute. By another way involves creating astatistical model to predict the missing values based on other features in the data. This can be a powerful technique but it requires complex computations(Tamboli, 2024).This method displayed good results for logistics regression with percent 92.8.

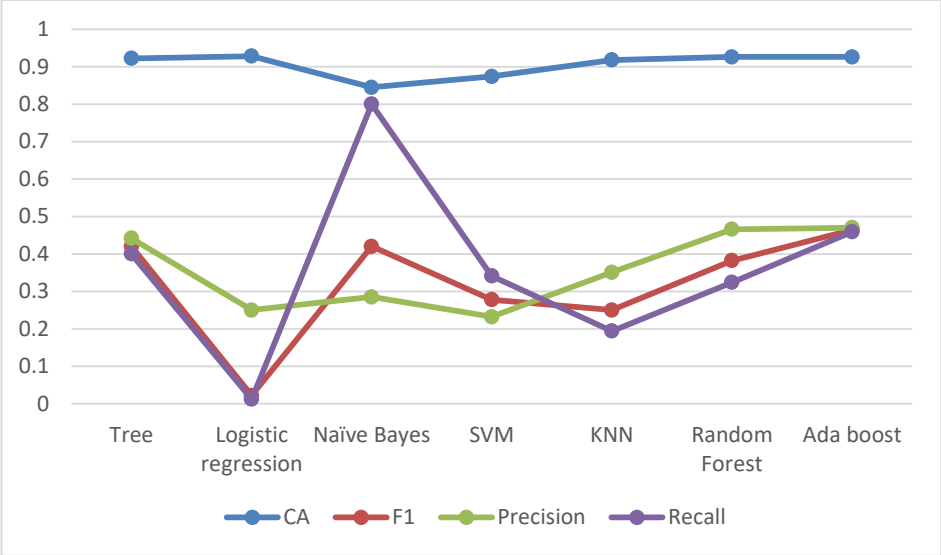
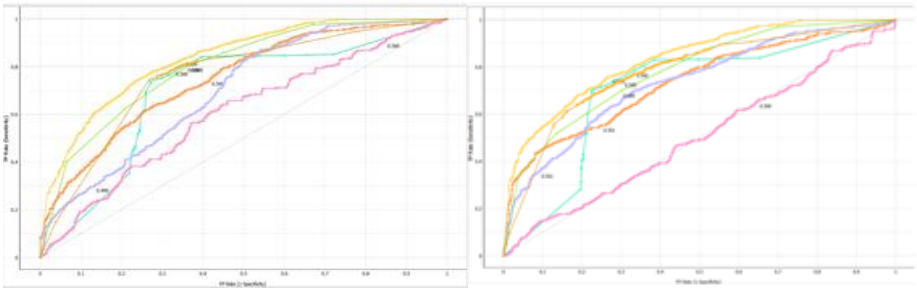
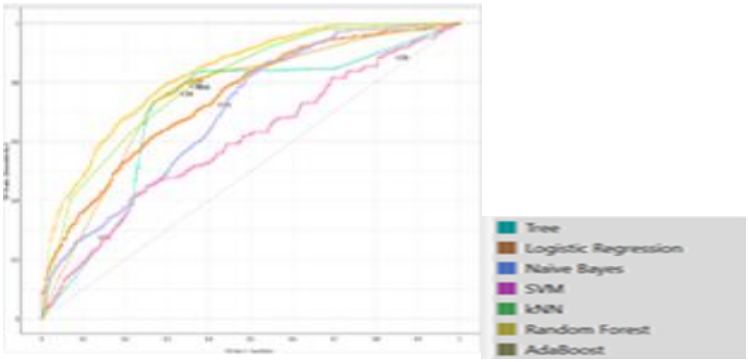


Figure 5: Comparison of Machine Learning Algorithms for Missing Data Imputation and Classification(simple tree)

Figure 5 conducted the seventh machine learning presented algorithms using simple tree treating missing data method and comparing between these algorithms relied on various measures of accuracy. The logistic regression , random forest and Ada boost algorithms displayed superiotity versus other methods poposed with (0.928, 0.926, 0.926) percents respectively



g.1 ROC curve of random values methods g.2 ROC curve of average method



g.3 Roc curve of simple tree method

Figure 6: ROC curve under missing data handling methods.

Figure 6 introduced a ROC curve of proposed methods, which illustrated the performance of a binary classifier model at varying threshold values, it also can be thought of as a plot of the statistical power as a function of type 1 error. The area under Roc curve (AUC) is widely recognized as the measure of a diagnostic test's discriminatory power (Fan et al., 2006). The maximum value for the AUC is 1.0, thereby identifying a (theoretically) perfect test (i.e., 100% sensitive and 100% specific). An AUC value of 0.5 indicates no discriminative value (i.e., 50% sensitive and 50% specific) and is represented by a straight, diagonal line extending from the lower left corner to the upper right. There are various scales for AUC value interpretation but, in general, ROC curves ≤0.75 are not useful and an AUC of 0.97 has a very high value, correlating with likelihood ratios of approximately 10 and 0.1 (Faraggi and Reiser, 2002). For all presented curves, the most dominant curve is the logistic regression curve and Ada Boost curve, introduced accurate tests under proposed methods in this study with 0.92% AUC.

Table 7: Merging between some machine learning algorithms.

Model	CA	F_1	Precision	Recall
Tree	0.922	0.420	0.442	0.400
Logistic Regression	**0.928**	0.022	0.250	0.012
Naïve Bayes	0.845	0.420	0.285	0.800
SVM	0.874	0.276	0.232	0.341
KNN	0.918	0.250	0.351	0.194
Random forest	**0.926**	0.382	0.466	0.324
Ada Boost	**0.926**	0.464	0.470	0.459
Merging	**0.926**	**0.151**	**0.381**	**0.094**

Table 7 presented merging between machine learning algorithms to get the best fit and the most effective methods of handling missing values of migration. Merging contained all proposed machine learning algorithms in this study. Also, merging logistic regression, random forest and Ada boost with values of classified accuracy 0.928, 0.926 and0.926 respectively. Merging results showed that the best merger was merging Ada boost and

logistic regression with accuracy measure equals 0.926. This emphasized that merging and using machine learning handling missing values with 92.6 percent.

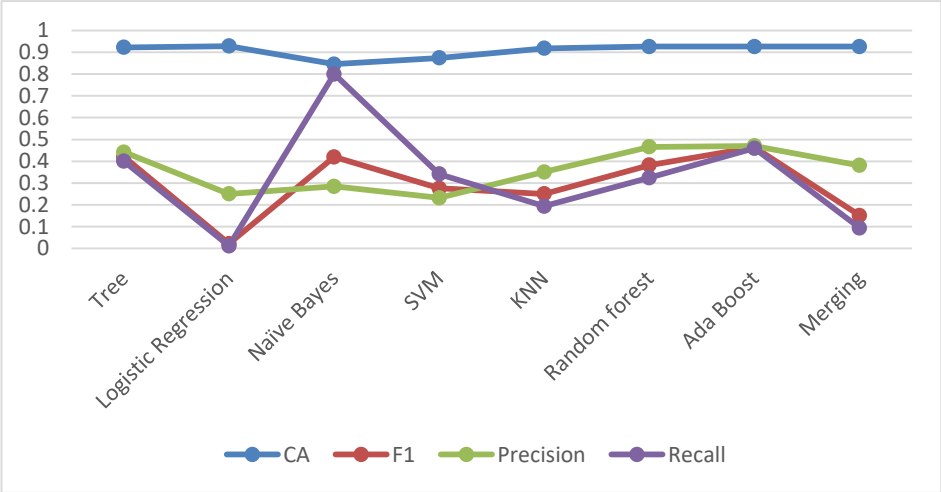


Figure7: Comparison of Machine Learning Algorithms for Missing Data Imputation and Classification (Merged Models)

Figure 7 revealed the accuracy of merging between proposed algorithms at this study. The merged models conducted that a best performance with accuracy was merging Ada boost and logistic regression algorithms with percent 0.926.

Table8: Feature statistics under methods of treating missing data.

Method	Variable	Mean	Mode	Median	Dispersion	Mini.	Max.	Missing
Simple tree	**Region-origin**		Home of Africa(p)		2.02			0(0%)
	Missing	40.66	1	10.21	2.22	0	750	0 (0%)
	Dead	5.02	1	1	4.07	0	750	0(0%)
	Incident – region		North Africa		1.9			0(0%)
	Lat.	26.9011	12.533	29.34	0.428083	-26.2245	66.9672	0(0%)
	Lon.	-13.97	1.85869	14.51	-4.41274	-117.071	116.225	0(0%)
	Reliability		Partially verified		0.862			0(0%)
	Id	95926.3	1	121178	0.65	1	184750	0(0%)
	Region- origin		**Home of Africa(p)**		**2**			**0(0%)**
	Missing	40.66	1	10.21	2.22	0	750	0 (0%)
	Dead	5.02	1	1	4.25	0	750	0(0%)
	Incident –		North Africa		1.9			0(0%)

Average/ mode value	**Region**							
	Lat.	26.9011	12.5331	29.34	0.428083	-26.2	66.9672	0(0%)
	Lon.	-13.9687	1.85869	14.4711	-4.41274	-117.1	116.225	0(0%)
	Reliability		Partially verified		0.862			0(0%)
	Id	95926.26	1	121178	0.65	1	184750	0(0%)
Random values	**Region-origin**		**Home of Africa(p)**		**2.19**			**0(0%)**
	Missing	40.66	1	10.21	2.22	0	750	0 (0%)
	Dead	4.69	1	1	4.29	0	750	0(0%)
	Incident Region				1.91			0(0%)
	Lat.	26.9011	12.5331	29.3489	0.42791	-26.2245	66.9672	0(0%)
	Lon.	-14.0435	1.85869	14.4711	-4.44	-117.071	116.225	0(0%)
	Reliability		Partially verified		0.912			0(0%)
	Id	95926.3	1	121178	0.65	1	184750	0(0%)

Table8 illustrated the statistics of variables after handling missing data using three methods of treatment, simple tree, average/ adequate value and random values. The percentage of missing after applying these techniques equals 0% with lower dispersion for both methods of simple tree and average/ adequate value. The less dispersion method for all variables is average/adequate method.

Figure8: Feature distributions after handling missing values

Figure 8 displayed the distributions of features after treatment of missing data. The figure showed that the distribution of some variables closed to normal distribution like, incident region, lat., and Lon. Also, reducing

the outliers value for both missing and reliability variables. This demonstrates on treating missing data improving the distribution of the data and hence getting more accurate results.

Table9: Accuracy Measures of Average Method for Handing Missing Data Under neutrosophic numbers

Model	CA_N	$F1_N$	$Prec_N$	$Recall_N$
Tree	0.996+0.996IN; IN ϵ[0,0.000]	0.97+0.97 IN ; IN ϵ[0,0.000]	0.976+0.976 IN ; IN ϵ[0,0.000]	0.965+0.965 IN ; IN ϵ[0,0.000]
Logistic Regression	0.917+0.918 IN ; IN ϵ[0,0.001]	0.082+0.092 IN ; IN ϵ[0,0.109]	0.184+0.208 IN ; IN ϵ[0,0.115]	0.053+0.059 IN ; IN ϵ[0,0.113]
Naive Bayes	0.843+0.844 IN ; IN ϵ[0,0.001]	0.449+0.449 IN ; IN ϵ[0,0.000]	0.298+0.298 IN ; IN ϵ[0,0.000]	0.906+0.912 IN ; IN ϵ[0,0.007]
SVM	0.856+0.858 IN ; IN ϵ[0,0.009]	0.335+0.336 IN ; IN ϵ[0,0.003]	0.243+0.248 IN ; IN ϵ[0,0.020]	0.518+0.547 IN ; IN ϵ[0,0.053]
kNN	0.995+0.995 IN ; IN ϵ[0,0.000]	0.965+0.965 IN ; IN ϵ[0,0.000]	0.959+0.959 IN ; IN ϵ[0,0.000]	0.971+0.971 IN ; IN ϵ[0,0.000]
Random Forest	0.996+0.997 IN ; IN ϵ[0,0.001]	0.97+0.976 IN ; IN ϵ[0,0.006]	0.988+1 IN ; IN ϵ[0,0.012]	0.953+0.953 IN ; IN ϵ[0,0.000]
AdaBoost	0.996+0.996 IN ; IN ϵ[0,0.000]	0.971+0.973 IN ; IN ϵ[0,0.002]	0.971+0.976 IN ; IN ϵ[0,0.005]	0.971+0.971 IN ; IN ϵ[0,0.000]
Merging	**0.997+0.998 IN ; IN ϵ[0,0.001]**	0.979+0.985 IN ; IN ϵ[0,0.006]	**0.994+1 IN ; IN ϵ[0,0.006]**	0.965+0.971 IN ; IN ϵ[0,0.006]

Table10: Accuracy Measures of Simple Tree Method for Handing Missing Data Under neutrosophic numbers

Model	CA_N	$F1_N$	$Prec_N$	$Recall_N$
Tree	0.996+0.996 IN ; IN ϵ[0,0.000]	0.97+0.97 IN ; IN ϵ[0,0.000]	0.976+0.976 IN ; IN ϵ[0,0.000]	0.965+0.965 IN ; IN ϵ[0,0.000]
Logistic Regression	0.915+0.920 IN ; IN ϵ[0,0.005]	0.072+0.085 IN ; IN ϵ[0,0.153]	0.154+0.22 IN ; IN ϵ[0,0.300]	0.047+0.053 IN ; IN ϵ[0,0.113]
Naive Bayes	0.841+0.843 IN ; IN ϵ[0,0.002]	0.443+0.445 IN ; IN ϵ[0,0.004]	0.29+0.296 IN ; IN ϵ[0,0.007]	0.9+0.9 IN ; IN ϵ[0,0.000]
SVM	0.858+0.871 IN ; IN ϵ[0,0.015]	0.292+0.317 IN ; IN ϵ[0,0.079]	0.238+0.24 IN ; IN ϵ[0,0.008]	0.376+0.471 IN ; IN ϵ[0,0.202]
kNN	0.995+0.995 IN ; IN ϵ[0,0.000]	0.965+0.965 IN ; IN ϵ[0,0.000]	0.959+0.959 IN ; IN ϵ[0,0.000]	0.971+0.971 IN ; IN ϵ[0,0.000]
Random Forest	0.996+0.996 IN ; IN ϵ[0,0.000]	0.97+0.973 IN ; IN ϵ[0,0.003]	0.994+1 IN ; IN ϵ[0,0.006]	0.947+0.947 IN ; IN ϵ99.[0,0.000]
AdaBoost	0.995+0.995 IN ; IN ϵ[0,0.000]	0.968+0.968 IN ; IN ϵ[0,0.000]	0.965+0.965 IN ; IN ϵ[0,0.000]	0.971+0.971 IN ; IN ϵ[0,0.000]

Merging	**0.998+0.998 IN ; IN ϵ[0,0.000]**	0.985+0.982 IN ; IN ϵ[0,0.003]	0.994+1 IN ; IN ϵ[0,0.006]	0.971+0.971 IN ; IN ϵ[0,0.000]

Table11: Accuracy Measures of Random Values Method for Handing Missing Data Under neutrosophic numbers

Model	CA_N	$F1_N$	$Prec_N$	$Recall_N$
Tree	0.969+0.978 IN ; IN ϵ[0,0.009]	0.817+0.863 IN ; IN ϵ[0,0.053]	0.797+0.861 IN ; IN ϵ[0,0.074]	0.838+0.865 IN ; IN ϵ[0,0.031]
Logistic Regression	0.902+0.903 IN ; IN ϵ[0,0.001]	0.1+0.125 IN ; IN ϵ[0,0.200]	0.194+0.227 IN ; IN ϵ[0,0.145]	0.067+0.086 IN ; IN ϵ[0,0.221]
Naive Bayes	0.866+0.868 IN ; IN ϵ[0,0.002]	0.53+0.54 IN ; IN ϵ[0,0.019]	0.367+0.377 IN ; IN ϵ[0,0.027]	0.948+0.949 IN ; IN ϵ[0,0.001]
SVM	0.882+0.9 IN ; IN ϵ[0,0.020]	0.219+0.297 IN ; IN ϵ[0,0.263]	0.238+0.338 IN ; IN ϵ[0,0.296]	0.203+0.264 IN ; IN ϵ[0,0.231]
KNN	0.981+0.984 IN ; IN ϵ[0,0.003]	0.88+0.896 IN ; IN ϵ[0,0.018]	0.923+0.927 IN ; IN ϵ[0,0.004]	0.838+0.87 IN ; IN ϵ[0,0.037]
Random Forest	0.98+0.986 IN ; IN ϵ[0,0.006]	0.869+0.906 IN ; IN ϵ[0,0.041]	0.916+0.97 IN ; IN ϵ[0,0.056]	0.827+0.85 IN ; IN ϵ[0,0.027]
Ada Boost	0.971+0.975 IN ; IN ϵ[0,0.004]	0.829+0.846 IN ; IN ϵ[0,0.020]	0.811+0.827 IN ; IN ϵ[0,0.019]	0.848+0.865 IN ; IN ϵ[0,0.020]
Merging	**0.982+0.986 IN ; IN ϵ[0,0.004]**	0.882+0.907 IN ; IN ϵ[0,0.028]	0.932+0.965 IN ; IN ϵ[0,0.034]	0.838+0.855 IN ; IN ϵ[0,0.020]

Tables 9, 10 &11 provided algorithms of machine learning and merging between these models under neutrosophic sets. These tables contained random values, average/adequate method and simple tree method to enhance missing data. For table 9, under average method and after entering neutrosophic sets on the data, we found that the classified accuracy increasing to 99.8 percent. While at table10 under simple tree method, neutrosophic sets increased accuracy to 99.8 percent. Finally at table 11, neutrosophic sets under random values method raised the accuracy to 98.2 percent. The findings from these tables proved on that using neutrosophic sets provided more accuracy rather than using algorithms only.

Conclusion: -

In this paper, we introduced seven machine learning algorithms namely, decision tree, logistic regression, naïve Bayes, support vector machine, K nearest neighbor, random forest and Ada Boost to handle missing data issue. These algorithms were applied on data of missing migrations to treat missing data. Handling missing data was conducted using three replacing missing data methods involving, random values, average value and simple tree method. Additionally accuracy criteria measurements were conducted for these algorithms and utilized in the analysis.

The analysis findings revealed that Naïve Bayes was superior performance in fitting missing data for random values method, logistic regression, while logistic regression was the best algorithm for both average method

and simple tree. Finally, merging between proposed algorithms emerged logistic regression as the best algorithm for CA criteria. Additionally applying neutrosophic sets gave more performance of accuracy, as it raised the ratio of accuracy for all proposed algorithms as shown in tables 9,10 and 10. The classified accuracy of merging increased from 92.6 percent to 99.8 percent, this emphasized on the good performance of proposed neutrosophic technique.

For future scope, we seek to apply type-2 Neutrosophic- logic to handle missing data. Also, we plan to develop a hybrid framework between a neutrosophic set and type-2 neutrosophic which incorporates machine learning. Furthermore, more applications will be conducted to test the efficiency of the proposed model with other membership functions or a higher number of membership functions.

References: -

1. Anwar, M. B., Hanif, M., Shahzad, U., Emam, W., Anas, M. M., Ali, N., & Shahzadi, S. (2024). Incorporating the neutrosophic framework into kernel regression for predictive mean estimation. Heliyon, 10(3).
2. Ayilara, O. F., Zhang, L., Sajobi, T. T., Sawatzky, R., Bohm, E., & Lix, L. M. (2019). Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. Health and quality of life outcomes, 17, 1-9. https://doi.org/10.1186/s12955-019-1181-2.
3. Boyko, N., & Dypko, O. (2023). Machine Learning Methods for the Detection of Misinformation in News Content. International Journal of Multidisciplinary and Current Educational Research (IJMCER), 5(30), 31-42.
4. Castro Sánchez, F., Almeida Blacio, J. H., Flores Bracho, M. G., Andrade Santamaria, D. R., & Sánchez Casanova, R. (2021). Neutrosophic and Plithogenic Statistical Analysis in Educational Development. Neutrosophic Sets and Systems, 44(1), 26.
5. Cerna Maguiña, H. F., Chumpitaz Ramos, D. G., Gallegos Ruiz Conejo, A. L., Pool Painted, N. I., & Baltazar Ángeles, J. I. (2024). Uncertainty Analysis in Prediction Intervals Using Neutrosophic Numbers. Neutrosophic Sets and Systems, 71(1), 13.
6. Chen, Y. C. (2022). Pattern graphs: a graphical approach to non-monotone missing data. The Annals of Statistics, 50(1), 129-146. https://doi.org/10.1214/21-AOS2094.
7. Delgado Estrada, S. M., Rocio Ruiz Molina, K. D., Ponce Orella-na, F. E., & Chabusa Vargas, J. L. (2024). Neutrosophic Statistics for Enhanced Time Series Analysis of Unemployment Trends in Ecuador. Neutrosophic Sets and Systems, 67(1), 14.
8. Dong, Y., & Peng, C. Y. J. (2013). Principled missing data methods for researchers. SpringerPlus, 2, 1-17. https://doi.org/10.1186/2193-1801-2-222.
9. Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic (ROC) curves. Canadian Journal of Emergency Medicine, 8(1), 19-20. https://doi.org/10.1017/s1481803500013336.
10. Faraggi, D., & Reiser, B. (2002). Estimation of the area under the ROC curve. Statistics in medicine, 21(20), 3093-3106. https://doi.org/10.1002/sim.1228.
11. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.https://doi.org/10.1006/jcss.1997.1504.

12. Ghazanfar, M. A., & Prugel, A. (2013). The advantage of careful imputation sources in sparse data-environment of recommender systems: Generating improved svd-based recommendations. Informatica, 37(1).
13. Gómez-Carracedo, M. P., Andrade, J. M., López-Mahía, P., Muniategui, S., & Prada, D. (2014). A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemometrics and Intelligent Laboratory Systems, 134, 23-33. https://doi.org/10.1016/j.chemolab.2014.02.007
14. Honghai, F., Guoshun, C., Cheng, Y., Bingru, Y., & Yumei, C. (2005, September). A SVM regression based approach to filling in missing values. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (pp. 581-587). Berlin, Heidelberg: Springer Berlin Heidelberg.
15. Houari, R., Bounceur, A., Tari, A. K., & Kecha, M. T. (2014, June). Handling missing data problems with sampling methods. In 2014 International conference on advanced networking distributed systems and applications (pp. 99-104). IEEE. https://doi.org/10.1109/INDS.2014.25.
16. Hui, G., Chen, Z., Wang, Y., Zhang, D., & Gu, F. (2023). An integrated machine learning-based approach to identifying controlling factors of unconventional shale productivity. Energy, 266, 126512. https://doi.org/10.1016/j.energy.2022.126512.
17. Jdid, M., Alhabib, R., & Salama, A. A. (2022). Fundamentals of neutrosophical simulation for generating random numbers associated with uniform probability distribution. Neutrosophic Sets and Systems, 49(1), 6.
18. Jdid, Maissam; Rafif Alhabib; and A. A. Salama. "The Basics of Neutrosophic Simulation for Converting Random Numbers Associated with a Uniform Probability Distribution into Random Variables Follow an Exponential Distribution." Neutrosophic Sets and Systems 53, 1 (2023).
19. KERVANCI, I. S. (2023). A Review Hybrid Structure of Neutrosophy and Machine Learning Algorithms for Different Types of Problems. *2023 Neutrosophic SuperHyperAlgebra And New Types of Topologies*, 133.
20. Liu, C. (1995). Missing data imputation using the multivariate t distribution. Journal of multivariate analysis, 53(1), 139-158. https://doi.org/10.1006/jmva.1995.1029.
21. Liu, J. (2021). Exploring teacher attrition in urban China through interplay of wages and well-being. *Education and Urban Society*, *53*(7), 807-830.
22. Maillo, J., Ramírez, S., Triguero, I., & Herrera, F. (2017). kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data. Knowledge-Based Systems, 117, 3-15. https://doi.org/10.1016/j.knosys.2016.06.012.
23. Pelckmans, K., De Brabanter, J., Suykens, J. A., & De Moor, B. (2005). Handling missing values in support vector machine classifiers. Neural Networks, 18(5-6), 684-692. https://doi.org/10.1016/j.neunet.2005.06.025.
24. Rokach, L. (2016). Decision forest: Twenty years of research. Information Fusion, 27, 111-125. https://doi.org/10.1016/j.inffus.2015.06.005
25. Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592. https://doi.org/10.1093/biomet/63.3.581
26. Salama, A. A.; Mahmoud Y. Shams; Sherif Elseuofi; and Huda E. Khalid. "Exploring Neutrosophic Numeral System Algorithms for Handling Uncertainty and Ambiguity in Numerical Data: An Overview

and Future Directions." Neutrosophic Sets and Systems 65, 1 (2024). https://digitalrepository.unm.edu/nss_journal/vol65/iss1/15

27. Sánchez, F. C., Blacio, J. H. A., Bracho, M. G. F., Santamaria, D. R. A., & Casanova, R. S. (2021). *Neutrosophic and Plithogenic Statistical Analysis in Educational Development*. Infinite Study.
28. Sawant, S., Savakhande, R., Sankhe, O., & Tamboli, S. (2024, February). Phishing Detection by integrating Machine Learning and Deep Learning. In 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 1078-1083). IEEE. https://doi.org/10.23919/INDIACom61295.2024.10499100.
29. Shinde, P. P., & Shah, S. (2018, August). A review of machine learning and deep learning applications. In 2018 Fourth international conference on computing communication control and automation (ICCUBEA) (pp. 1-6). IEEE.
30. Smarandache, F. (1999). A unifying field in Logics: Neutrosophic Logic. In *Philosophy* (pp. 1-141). American Research Press.
31. Smarandache, F. (2022). Neutrosophic Statistics is an extension of Interval Statistics, while Plithogenic Statistics is the most general form of statistics (second version) (Vol. 2). Infinite Study.
32. Smola, A. J., Vishwanathan, S. V. N., & Hofmann, T. (2005, January). Kernel methods for missing variables. In International Workshop on Artificial Intelligence and Statistics (pp. 325-332) PMLR.
33. Stewart, T. G., Zeng, D., & Wu, M. C. (2018). Constructing support vector machines with missing data. Wiley Interdisciplinary Reviews: Computational Statistics, 10(4), e1430. https://doi.org/10.1002/wics.1430.
34. Suthar, B., Patel, H., & Goswami, A. (2012). A survey: classification of imputation methods in data mining. International Journal of Emerging Technology and Advanced Engineering, 2(1), 309-12.
35. Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. Applied Artificial Intelligence, 23(5), 373-405. https://doi.org/10.1080/08839510902872223.
36. Zhang, Z. (2015). Missing values in big data research: some basic skills. Annals of translational medicine, 3(21). https://doi.org/10.3978%2Fj.issn.2305-5839.2015.12.11

Received: Sep 5, 2024. Accepted: Feb 10, 2025