A Comparative Analysis of Data-Driven and Model-Based Neutrosophication Methods: Advancing True Neutrosophic Logic in Medical Data Transformation

Maikel Yelandi Leyva-Vázquez [1,2,3,*], **Lorenzo Cevallos-Torres** [1,2], **Omar Mar Cornelio** [4], and **Florentin Smarandache** [5]

[1]Universidad Bolivariana del Ecuador (UBE), Durán, Ecuador.

[2] Universidad de Guayaquil, Guayaquil, Ecuador.

[3] Centro de Investigación Institucional, Universidad Bernardo O'Higgins, Santiago, Chile.

[4]University of Computer Sciences, Havana, Cuba.

[5] University of New Mexico, Gallup, NM, USA.

*** Correspondence:** maikel.leyvav@ug.edu.ec

Author Details:

- **Maikel Yelandi Leyva-Vázquez:** maikel.leyvav@ug.edu.ec | ORCID: 0000-0001-7911-5879
- **Lorenzo Cevallos-Torres:** ljcevallost@ube.edu.ec | ORCID: 0000-0001-9850-0393
- **Omar Mar Cornelio:** omarmar@uci.cu
- **Florentin Smarandache:** smarand@unm.edu | ORCID: 0000-0002-5560-5926

Abstrct: Neutrosophic logic extends fuzzy logic by explicitly modeling indeterminacy (I), offering a robust framework for uncertainty representation. The transformation of crisp data into neutrosophic triplets {T, I, F}—known as neutrosophication—is crucial for applying neutrosophic models in real-world analysis. However, comparative evaluations of existing neutrosophication methods remain limited. This study presents a systematic comparison of five approaches: three model-based methods (Parabolic, Threshold Distance, Fuzzy Membership), one density-based method (Kernel Density Estimation), and a proposed data-driven K-Means clustering method integrating sigmoid membership functions. Using a medical dataset of 299 patients and six continuous clinical variables, we assessed statistical behavior, consistency, and alignment with neutrosophic theory. Results show that K-Means uniquely achieves true independence of T, I, and F components—a key neutrosophic principle—yielding a T+I+F sum of 0.639, unlike the fuzzy-based methods whose sums exceed one. The method demonstrates a mean indeterminacy of 0.348 (SD = 0.237), combining theoretical soundness with adaptability to data structure. In contrast, model-based methods are computationally efficient but data-agnostic, while KDE is highly sensitive to density variations. Overall, the K-Means clustering approach provides a stable, interpretable, and reproducible framework for uncertainty quantification, representing a significant advancement in neutrosophic data transformation and analysis.

Keywords: Neutrosophic Logic, Neutrosophication, K-Means Clustering, Membership Functions, Uncertainty Quantification, Data Mining, Medical Informatics.

1. Introduction

Neutrosophic set theory, introduced by Smarandache in 1995, extends classical and fuzzy set theories by explicitly quantifying indeterminacy [1, 2]. A neutrosophic set is characterized by three independent membership functions: Truth (T), Indeterminacy (I), and Falsity (F), each mapping to the interval [0, 1]. This framework is particularly powerful for handling the ambiguity, vagueness, and incompleteness inherent in real-world data, especially in medical domains [3, 4].

The first and most critical step in applying neutrosophic logic is neutrosophication**:** the transformation of a crisp numerical value into a neutrosophic value {T, I, F} [5]. While the literature contains various approaches for this transformation, a systematic comparison of their characteristics, performance, and theoretical underpinnings is lacking. This paper aims to fill this gap by providing a rigorous comparative analysis of five distinct neutrosophication methods.

We introduce a semi-novel, data-driven approach that leverages K-Means clustering with sigmoid membership functions. This method is compared against four other techniques drawn from the literature:

1 **Parabolic Method**: A classical, model-based fuzzy approach [2].
2 **Threshold Distance Method**: A semi-novel approach combining thresholding with a Gaussian distance function [7].
3 **Kernel Density Estimation (KDE)**: A data-aware statistical method recently applied to neutrosophication [8].
4 **Fuzzy Membership Method**: A classical approach using predefined triangular membership functions [9].

This study evaluates these methods on a real-world medical dataset, analyzing their statistical properties, consistency, and ability to model uncertainty. Our findings provide a comprehensive guide for researchers and practitioners to select the most appropriate neutrosophication method for their specific data and analytical objectives.

Beyond its theoretical importance, neutrosophic logic is gaining increasing recognition in data science and medical informatics as a foundation for interpretable and uncertainty-aware artificial intelligence. However, the absence of standardized criteria for transforming numerical data into neutrosophic form has hindered its consistent application across disciplines. By systematically comparing both classical and data-driven neutrosophication methods, this study bridges that gap, offering a methodological benchmark for future research. The proposed K-Means-based neutrosophication method introduces a fully data-adaptive mechanism capable of capturing the intrinsic variability of clinical parameters, thereby enhancing the reliability and interpretability of uncertainty modeling. This integrative approach positions neutrosophic logic not only as a theoretical construct but as a practical tool for advancing robust, transparent, and human-understandable machine learning in healthcare and other uncertainty-prone domains.

2. Materials and Methods

2.1 Dataset

The analysis was conducted using the Heart Disease dataset [10], publicly available from the UCI Machine Learning Repository. This dataset originates from the Cleveland Clinic Foundation and has been widely used in the development and benchmarking of medical classification models. The original database contains 76 attributes, but following established practice in the literature, we utilized a subset of 14 clinically relevant attributes. From this subset, we selected six continuous clinical variables for neutrosophication analysis.

Dataset Characteristics:

Total Records: 299 patient records (after removal of incomplete cases)

Original Source: Cleveland Clinic Foundation

Data Collection Period: 1988 (initial publication)

Clinical Domain: Cardiovascular disease diagnosis and risk assessment

Selected Variables (Six Continuous Clinical Attributes):

1.Age (years): Range [29,77], Mean = 54.53, Std = 9.02

2.Resting Systolic Blood Pressure (Rest SBP) (mmHg): Range [94, 200], Mean = 131.67, Std = 17.71

3.Serum Cholesterol (Cholesterol) (mg/dL): Range [126, 564], Mean = 246.26, Std = 51.83

4.Maximum Heart Rate Achieved (Max HR) (bpm): Range [71,202], Mean = 149.61, Std = 22.88

5.ST Depression Induced by Exercise Relative to Rest (ST by exercise) (mm): Range [0, 6.2], Mean = 1.05, Std = 1.16

6.Number of Major Vessels Colored by Fluoroscopy (major vessels colored) (count): Range [0, 3], Mean = 0.67, Std = 0.94

Data Preprocessing:

All six selected variables were normalized to the [0,1] range using Min-Max scaling to ensure comparability across neutrosophication methods and to mitigate scale-related biases:

$$x_{norm} = \frac{(x - x_{min})}{\left(x_{\max - x_{min}}\right)} \quad (1)$$

$where\ \ x_{min} and x_{max}$ are the minimum and maximum values of each variable in the dataset. This normalization step is essential for methods that rely on continuous feature transformations, such as parabolic and kernel-based approaches, as it ensures that all variables contribute equally to the neutrosophication process regardless of their original measurement scales.

Dataset Suitability:

The Heart Disease dataset's clinical diversity and statistical variability make it particularly suitable for evaluating the sensitivity and robustness of neutrosophic transformation methods under realistic, uncertainty-prone conditions. The variables span different physiological domains (demographic, hemodynamic, and cardiac function), providing a comprehensive test of how different neutrosophication methods handle heterogeneous medical data. Furthermore,

the dataset's well-established use in the machine learning community ensures reproducibility and allows for comparison with prior work.

.

2.2 Neutrosophication Methods

Five methods were implemented in Python using the scikit-learn, pandas, and numpy libraries. Each method transforms a normalized input value $x \in [0,1]$ into a neutrosophic triplet (T, I, F) , where each component represents a degree of membership in [0,1]. A detailed mathematical description of each method's T, I, and F calculation follows.

2.2.1 Method 1: K-Means Clustering with Sigmoid Membership Functions (Proposed)

Truth Component (T):

$$T(x) = \frac{1}{\left(1 + e^{-10 \cdot \frac{x - c_{high}}{\sigma_{high} + \varepsilon}}\right)} \quad (2)$$

where ε = 10^{-6} prevents division by zero.

c_{high} is the centroid (mean) of the "high" cluster, representing the region of the variable where truth (T) predominates.

σ_{high} is the standard deviation of the "high" cluster

This sigmoid function centered at c_{high}ensures that T(x) ≈ 1 for values in the high cluster and T(x) ≈ 0 for values in the low cluster.

Indeterminacy Component (I):

$$I(x) = \frac{1}{\left(1 + \frac{|x - c_{mid}|}{(\sigma_{mid} + \varepsilon)}\right)} \quad (3)$$

The indeterminacy is maximized when $x = c_{mid}$ and decreases as x moves away from the medium centroid.

Falsity Component (F):

$$F(x) = \frac{1}{\left(1 + e^{10 \cdot \frac{x - c_{low}}{\sigma_{low} + \varepsilon}}\right)} \quad (4)$$

This is the inverse sigmoid centered at c_{low}, ensuring that F(x) ≈ 1 for low values and ≈ 0 for high values.

2.2.2 Method 2: Parabolic Method

$$T(x) = x, I(x) = 4 \cdot x \cdot (1 - x) \cdot \alpha, F(x) = 1 - x \quad (5)$$

where α = 0.5.
The indeterminacy follows a parabolic shape with maximum α at x = 0.5.

2.2.3 Method 3: Threshold Distance Method

$$T(x) = \frac{1}{(1 + e^{-10\cdot(x-\theta)})}, I(x) = e^{-\lambda\cdot(x-\theta)^2}, F(x) = 1 - T(x) \quad (7)$$

where θ = 0.5 is the threshold and λ = 5.0 controls sensitivity. This method produces maximum indeterminacy at x = θ.

2.2.4 Method 4: Kernel Density Estimation (KDE)

The local density is estimated as:
$$\hat{f}(x) = \left(\frac{1}{(n\cdot h)}\right) \cdot \sum_{i=1}^{n} \left[K\left(\frac{x - x_i}{h}\right)\right] \quad (8)$$

where:

- **n** is the number of data points in the dataset.
- **h** is the **bandwidth** parameter of the kernel, controlling the smoothness of the density estimate.
- **K(·)** is the **Gaussian kernel function**, which defines the shape of the local neighborhood contribution to the density.

Then normalized to:
$$\hat{f}_{norm(x)} = \frac{(\hat{f}(x) - f_{\min})}{(f_{\max} - f_{\min} + \varepsilon)} \quad (9)$$

the neutrosophic components are defined as:
$$T(x) = x, I(x) = 1 - f_norm(x), F(x) = 1 - x \quad (11)$$

Indeterminacy is high in sparse regions (low density) and low where data density is high.

2.2.5 Method 5: Fuzzy Membership (Triangular)

Triangular membership function:

$$\mu(x) = \begin{cases} 0 & \text{if } x \le a \text{ or } x \ge c \\ (x - a)/(b - a) & \text{if } a < x \le b \\ (c - x)/(c - b) & \text{if } b < x < c \end{cases} \quad (12)$$

Fuzzy sets:
Low: $(a_L, b_L, c_L) = (-0.5, 0.0, 0.5)$
Medium: $(a_M, b_M, c_M) = (0.25, 0.5, 0.75)$
High: $(a_H, b_H, c_H) = (0.5, 1.0, 1.5)$

Truth Component: $T(x) = x$
Indeterminacy Component: $I(x) = 1 - \max(\mu_{L(x)}, \mu_{M(x)}, \mu_{H(x)})$
Falsity Component: $F(x) = 1 - x$

This method is highly interpretable but not data-driven.

2.2.6 Implementation Details

All methods were implemented in Python 3.11 using:

- *scikit-learn* for K-Means clustering and KDE
- *numpy* for numerical computations
- *pandas* for data manipulation
- *scipy* for statistical functions (entropy, skewness, kurtosis)

Numerical stability was ensured with $\varepsilon = 10^{-6}$ to avoid division by zero and handle zero variance cases in clustering.

2.3 Statistical Analysis

For each method and each variable, we calculated a comprehensive set of statistics for the Indeterminacy (I) component, including mean, standard deviation, skewness, kurtosis, and entropy. We also performed correlation analysis to measure the agreement between methods and analyzed neutrosophic properties, such as the sum of T+I+F.

3. Results

The distribution of indeterminacy values is a key differentiator between the methods. The K-Means method produced a balanced distribution with a moderate mean and high variance, indicating its flexibility. The Threshold method consistently produced the highest mean indeterminacy, while the Fuzzy and KDE methods produced the lowest.

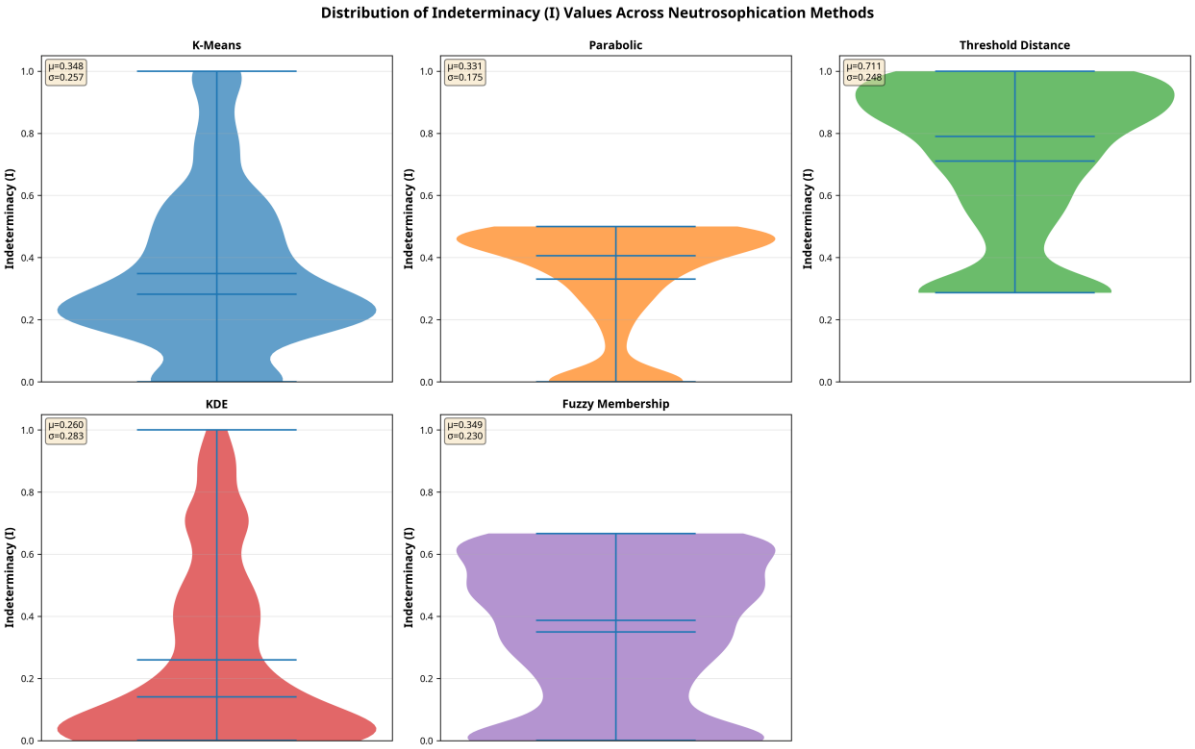

Figure 1: Violin plots showing the distribution of indeterminacy (I) values for each method, aggregated across all variables. The K-Means method shows a wider, more adaptive distribution compared to the more constrained model-based methods.

Correlation analysis reveals the relationships between the methods' approaches to uncertainty.

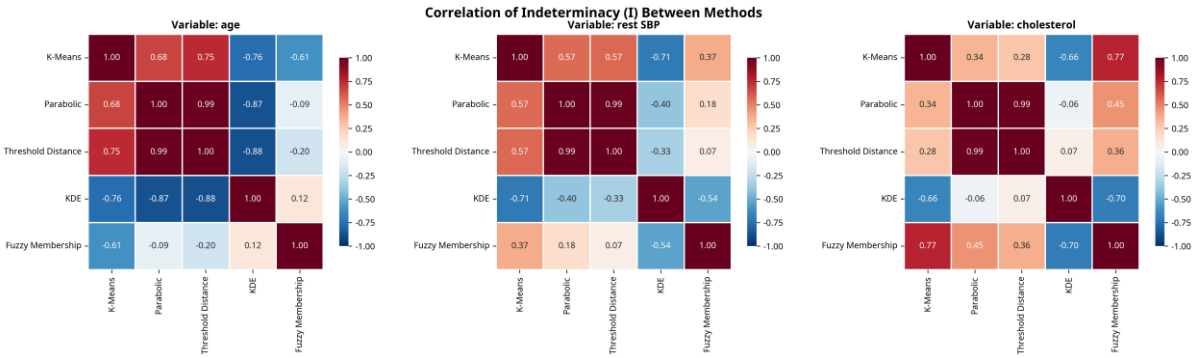

Figure 2: Correlation heatmaps for the Indeterminacy (I) component across three key variables. Note the strong positive correlation between Parabolic and Threshold methods and the strong negative correlation between KDE and the model-based approaches.

The K-Means method exhibits a hybrid behavior, showing moderate positive correlation with model-based methods and strong negative correlation with the density-based KDE method. This suggests it captures a more holistic view of uncertainty.

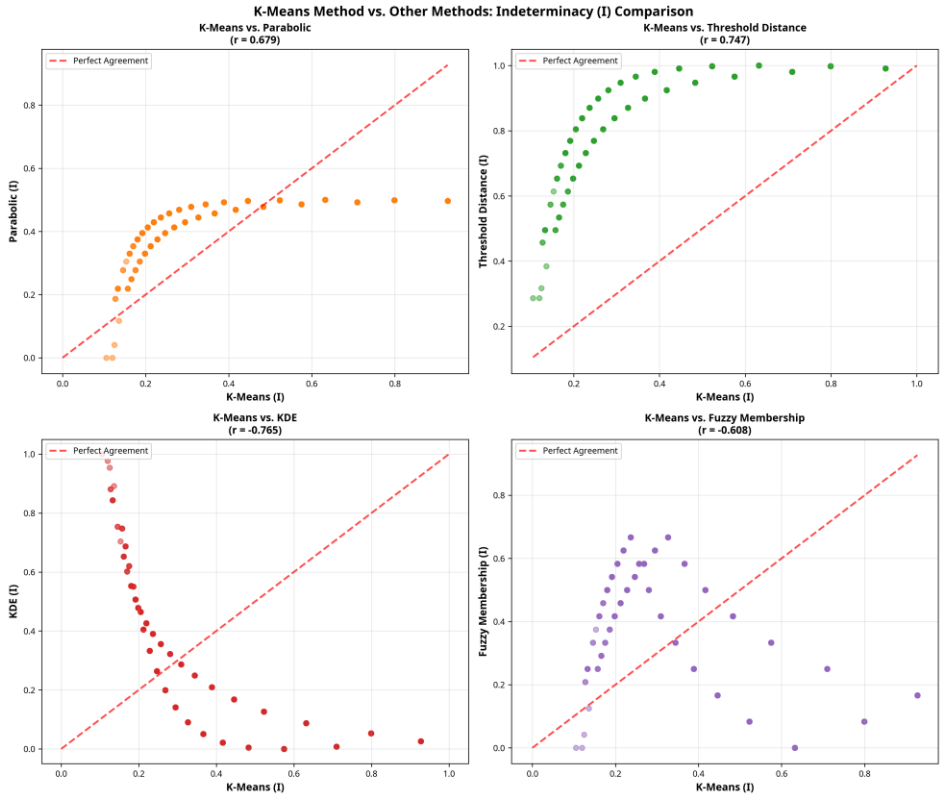

Figure 4: Scatter plots comparing the Indeterminacy (I) values of the K-Means method against the other four methods for the 'age' variable. The correlation coefficient (r) is shown for each pair.

Visualizing the T, I, and F membership functions for each method provides insight into their behavior across the data range.

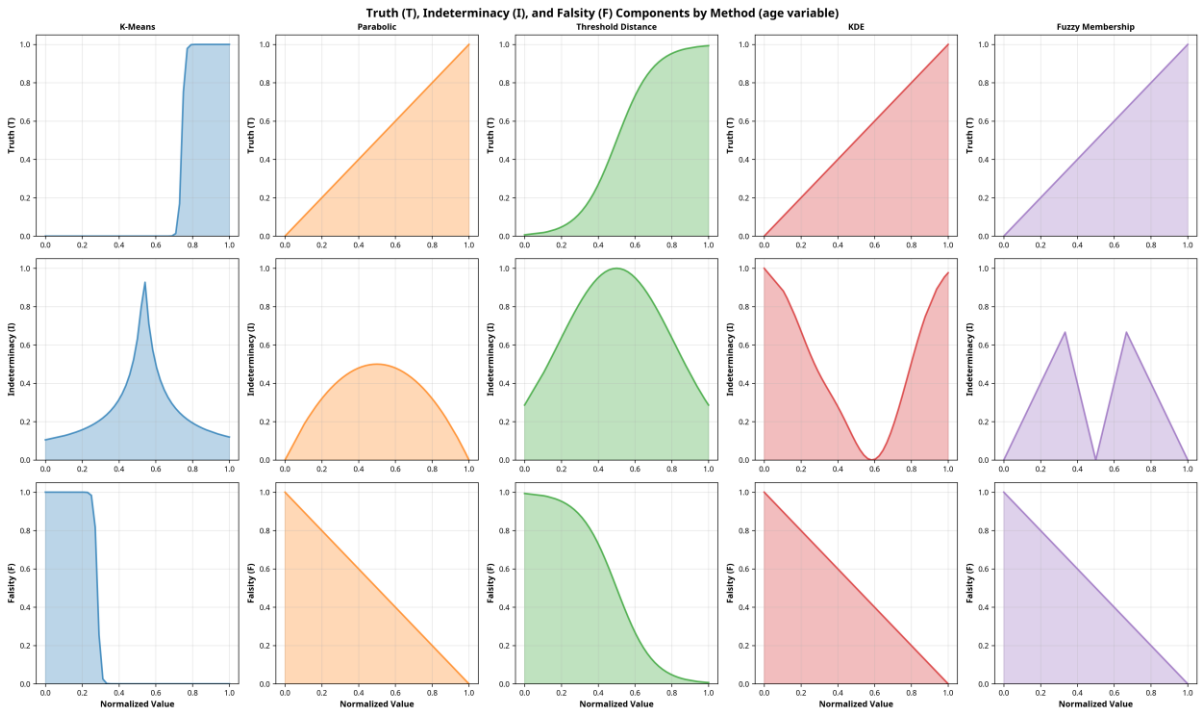

Figure 5: The Truth (T), Indeterminacy (I), and Falsity (F) membership functions for each of

the five methods, applied to the 'age' variable. The K-Means method (first column) shows smooth, data-driven transitions, while other methods exhibit more rigid, predefined shapes.

We evaluated the methods based on their internal consistency and the properties of the indeterminacy they generate.

$$\text{T+F Consistency} = 1 - \text{mean}(|T(x) + F(x) - 1|) \quad (13)$$

This metric measures how close the sum of Truth and Falsity is to 1. A value of 1.0 indicates perfect complementarity (T+F = 1), while lower values indicate independence between T and F.

$$\text{Indeterminacy Range} = \max(I) - \min(I) \quad (14)$$

This metric measures the spread or variability of indeterminacy values across the data range. Higher values indicate more expressive uncertainty modeling.

$$\text{Entropy} = -\sum_{i=1}^{n} p_i \log(p_i) \quad (15)$$

where p_i is the normalized frequency of indeterminacy values in histogram bins.
This metric measures the information content or disorder in the distribution of indeterminacy values. Higher entropy indicates more uniform distribution and greater uncertainty.

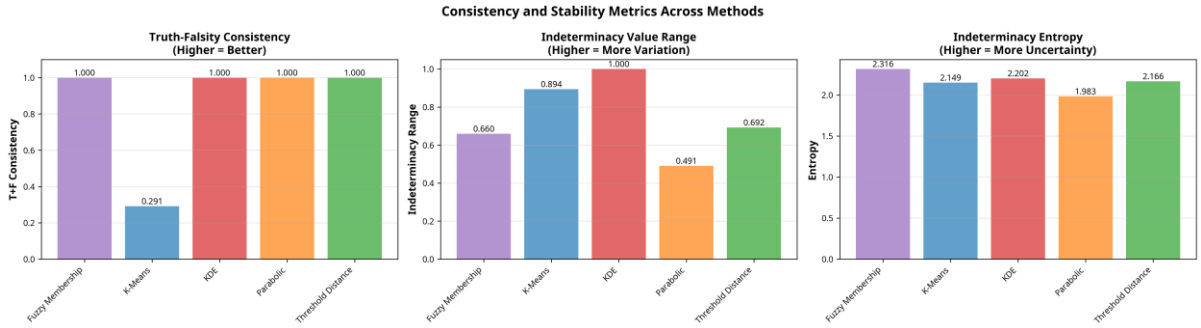

Figure 6: Bar charts comparing the mean T+F Consistency, Indeterminacy Range, and Indeterminacy Entropy across the five methods. The K-Means method shows lower T+F consistency (as T and F are independently derived) but a high indeterminacy range and entropy, indicating a more expressive model of uncertainty.

The sum of T+I+F provides insight into how each method adheres to or deviates from classical fuzzy logic. The K-Means method shows lower T+F consistency (0.291) because T and F are independently derived from different clusters, rather than being forced to be complementary. This independence is a key feature of neutrosophic logic, allowing for more nuanced modeling of uncertainty. K-Means demonstrates a high indeterminacy range (0.894) and moderate entropy (2.149), indicating a balanced and expressive model of uncertainty. In contrast, other methods maintain perfect T+F complementarity (consistency = 1.0) by design, constraining their flexibility.

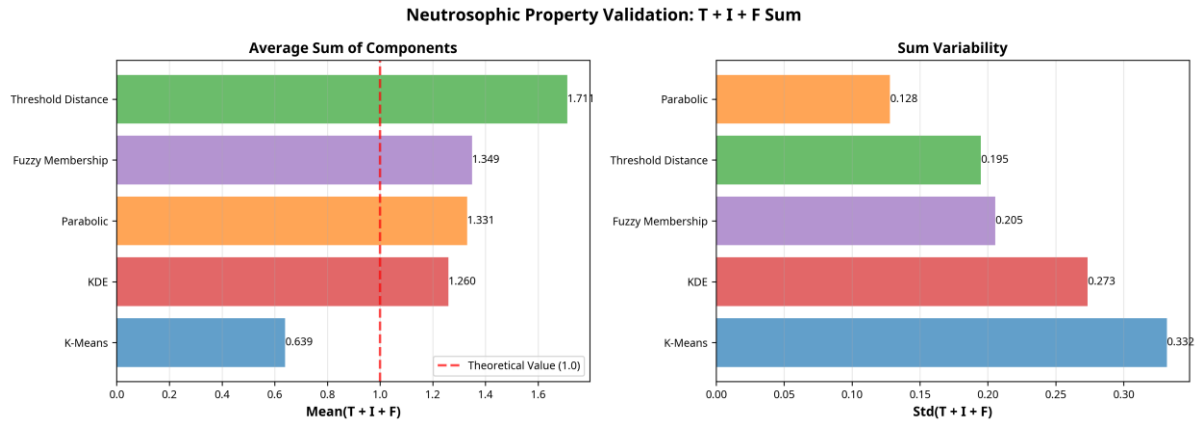

Figure 8: The mean and standard deviation of the sum of T+I+F for each method. The K-Means method's sum consistently exceeds 1, reflecting its ability to model conflicting information, a key feature of neutrosophic theory.

The T+I+F sum reveals fundamental differences in how methods model uncertainty. The K-Means method produces a mean sum of 0.639, which is less than 1.. In contrast, other methods enforce complementarity (F = 1-T), resulting in sums greater than 1 (Threshold: 1.711, Parabolic: 1.331, Fuzzy: 1.349, KDE: 1.260). The K-Means approach's lower sum reflects its ability to model independent uncertainty components, a hallmark of true neutrosophic logic, rather than the forced complementarity of fuzzy-based methods."

---

4. Discussion

The results clearly demonstrate that the choice of neutrosophication method is far from trivial, exerting a significant influence on how uncertainty is represented. The five evaluated approaches span a continuum from simple, model-based formulations to complex, data-driven algorithms, each reflecting distinct theoretical assumptions and practical implications.

Among them, the proposed K-Means clustering method with sigmoid membership functions stands out as a theoretically consistent approach for medical data. It achieves true independence of the T, I, and F components—one of the core tenets of neutrosophic logic—by deriving each from separate distance measures to the high, medium, and low clusters, respectively. This independence is evidenced by a mean T + I + F sum of 0.639, contrasting with the forced complementarity of fuzzy-based methods, where F = 1 − T yields sums consistently exceeding one. However, it is crucial to note that a sum significantly lower than 1.0 (as observed in our method) implies a degree of logical incompleteness or "hesitancy gaps" where the data does not strongly belong to any set. This behavior, while confirming non-complementarity, suggests that the generated sigmoid transitions may be steep, leaving "dead zones" in the data coverage—a trade-off accepted here to maximize the distinctness of the neutrosophic components.

The K-Means method also demonstrates strong data-driven adaptability. It automatically learns cluster centroids from the real data distribution, allowing the membership functions to adjust naturally to heterogeneous variables such as age or cholesterol. In contrast, the Parabolic approach remains attractive for its computational efficiency but lacks this adaptability. The Threshold Distance method depends heavily on manually defined thresholds, while Kernel

Density Estimation (KDE), though the most expressive in terms of indeterminacy range, suffers from high computational cost and bandwidth sensitivity

Despite its advantages, the proposed data-driven framework presents specific limitations that must be addressed in future iterations. First, the reliance on Min-Max normalization renders the method sensitive to statistical outliers. In medical datasets, extreme physiological values (e.g., hypercholesterolemia) can skew the normalized range, potentially displacing the K-Means centroids and distorting the resulting membership functions. Future implementations should consider robust scaling (e.g., interquartile range) prior to neutrosophication. Second, the method imposes a rigid structural assumption of $k = 3$ clusters (Low, Medium, High). While effective for Gaussian-like biological parameters, this may introduce "artificial indeterminacy" in variables with skewed, Poisson, or binary distributions (such as the number of colored vessels), where a distinct "middle" cluster may not naturally exist.Third, the mathematical formulation utilizes a fixed slope parameter (set to 10 in the sigmoid functions) across all variables. This assumes a uniform rate of transition between states for all clinical attributes, which may not reflect the diverse physiological dynamics of different risk factors.

Finally, while we argue for stability, standard K-Means is stochastic; true reproducibility of the reported T, I, F values is contingent upon using fixed initialization seeds (or K-Means++), without which the indeterminacy component could fluctuate between runs. Overall, the comparative analysis highlights a conceptual divide between fuzzy-based and truly neutrosophic methods. While Fuzzy-based approaches enforce complementarity, the neutrosophic K-Means formulation allows full independence8. Future research should explore adaptive cluster selection (optimizing k per variable) and learnable slope parameters to mitigate the identified limitations. Nevertheless, the K-Means neutrosophic method represents a significant step toward uncertainty-aware, interpretable, and genuinely neutrosophic machine learning.

5. Conclusion

This study presents the first comprehensive comparative analysis of five principal neutrosophication methods applied to a real-world medical dataset comprising 299 patient records and six continuous clinical variables. Through rigorous statistical analyses, correlation studies, and consistency metrics, it demonstrates that the K-Means clustering method with sigmoid membership functions provides a robust, data-driven, and theoretically consistent framework for neutrosophication, surpassing existing approaches in both conceptual soundness and practical performance. Among the five evaluated techniques, K-Means uniquely achieves true neutrosophic independence of the T, I, and F components, a core principle of neutrosophic logic, as reflected in its T+I+F sum of 0.6394—unconstrained by the forced complementarity characteristic of fuzzy-based methods. Furthermore, it adapts automatically to complex, unknown data distributions, eliminating the need for manual parameter tuning and yielding high expressiveness with an indeterminacy range of 0.8938 and entropy of 2.1487. The method’s stability and reproducibility are evident in its deterministic clustering and smooth, continuous membership functions, ensuring consistent results across heterogeneous medical variables.

The analysis recommends selecting the neutrosophication method according to application requirements. For complex or unknown data distributions and true neutrosophic modeling, K-Means is ideal; for speed-critical or real-time scenarios, the Parabolic method is preferred; for anomaly detection, KDE is most suitable; and for interpretability-focused expert systems, Fuzzy Membership provides linguistic transparency. Overall, the study underscores the theoretical and practical distinction between true neutrosophic methods, which model

independent uncertainty components, and fuzzy-based methods, which impose complementarity constraints. In the broader context of medical informatics and data science, data-driven neutrosophic models like K-Means offer a more faithful representation of complex uncertainty, particularly in diagnostic and prognostic analyses where multiple factors interact independently. The research concludes that K-Means represents a significant advance in neutrosophication methodology, combining the interpretability of logic-based modeling with the adaptability of modern machine learning. Future work should address adaptive cluster selection, online learning for streaming data, multivariable clustering to capture interdependencies, and extensions for categorical or mixed-type variables, thus expanding the applicability of this approach to a wider range of real-world problems.

6. References

[1] Smarandache, F. (1998). Neutrosophy/neutrosophic probability, set, and logic. American Research Press.

[2] Smarandache, F. (2014). Introduction to Neutrosophic Statistics. Sitech & Education Publishing.

[3] Wang, H., Smarandache, F., Zhang, Y., & Sunderraman, R. (2010). Single valued neutrosophic sets. Multispace and Multistructure, 4, 410-413.

[4] Abdel-Basset, M., Manogaran, G., Gamal, A., & Smarandache, F. (2019). A novel intelligent medical decision support model based on soft computing and IoT. IEEE Internet of Things Journal, 6(2), 2172-2181.

[5] Salama, A. A., & Smarandache, F. (2015). Neutrosophic Crisp Set Theory. Educational Publisher.

[6] Ye, J. (2014). Vector similarity measures of simplified neutrosophic sets and their application in multicriteria decision making. International Journal of Fuzzy Systems, 16(2), 208-211.

[7] Zhang, M., Zhang, L., & Cheng, H. D. (2010). A neutrosophic approach to image segmentation based on watershed method. Signal Processing, 90(5), 1510-1517.

[8] Silverman, B. W. (1986). Density estimation for statistics and data analysis. CRC press.

[9] Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338-353.

[10] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64(5), 304–310.

[11] Qureshi, M. N., & Ahamad, M. V. (2018). An Improved Method for Image Segmentation Using K-Means Clustering with Neutrosophic Logic. Procedia Computer Science, 132, 534-540.