Feature Imputation using Neutrosophic Set Theory in Machine Learning Regression Context

Yamen El Touati
Department of Computer Science, Faculty of Computing and Information Technology, Northern Border University, Saudi Arabia
yamen.touati@nbu.edu.sa (corresponding author)

Walid Abdelfattah
Department of Mathematics, College of Arts and Science, Northern Border University, Saudi Arabia
walid.abdelfattah@nbu.edu.sa

Received: 11 February 2024 | Revised: 28 February 2024 | Accepted: 11 March 2024

ABSTRACT

The prediction context of machine learning aims to discern the underlying patterns that dictate the characteristics needed to forecast the output. This prediction, however, lacks precision when the input data are not accurate or precise. The current study focuses on feature imputation through the application of the neutrosophic set theory. The primary concept involves substituting the feature data, which may exhibit accuracy and correctness issues, with neutrosophic variables considering the degrees of truth, indeterminacy, and falsity to produce more precise and resilient predictions. The proposed method was implemented in a specific case study, and the results are analyzed.

Keywords-neutrosophic set theory; machine learning; prediction; features imputation

I. INTRODUCTION

In the field of machine learning [1], specifically in the context of regression models [2], the correctness of data characteristics is of utmost importance. The efficacy of a regression model depends on its ability to identify patterns and correlations within the input data, therefore establishing the accuracy of these attributes crucial in the overall prediction process [3-6]. Flawed or untrustworthy data characteristics might add interference and hinder the model's ability to make generalizations, resulting in unreliable predictions and impaired effectiveness. Furthermore, the collaboration among specialists in a certain field and professionals specialized in data analysis is crucial in enhancing data feature precision. Experts in the field can provide useful insights into the intricacies of the data, enhancing the comprehension of the latter’s traits and their possible influence on the model's performance. Consistent surveillance and verification of the data during the model's lifespan are equally crucial, as the dataset attriubutes may change over time.

Experts face significant obstacles when evaluating and formalizing the correctness and precision of data, especially when it comes to aspects that directly affect the accuracy of predictions in machine learning and regression [3, 7]. A significant obstacle arises from the ever-changing nature of real-world data, where inaccuracies and imprecisions might originate from many different sources, such as human fallibility, measurement devices, or environmental influences. Experts must confront the complexities of detecting and reducing these mistake sources, which often necessitate a sophisticated understanding of both the particular field and the subtleties of the machine learning algorithms used. Furthermore, there is also difficulty in rreassuring the presence of ongoing monitoring and adjustment, as the data environment changes over time. Achieving a suitable equilibrium between the aspiration for the highest accuracy and the pragmatic limitations of collecting and analyzing data is still another complex facet. Collaboration between experts in a specific field and specialized data analysis professionals becomes crucial to address these challenges. This is because the former utilizes the combined expertise to improve the accuracy and precision of characteristics, ultimately strengthening the dependability of machine learning and regression models to make well-informed predictions. The task of developing a formal system to handle imprecise data in machine learning becomes increasingly challenging when dealing with a wide range of sources [8]. Inclusion of diverse datasets brings about complexity, necessitating the use of advanced statistical models to reconcile the differences between them. The need to measure inaccuracy and resolve uncertainty in feature values requires a well-balanced and flexible formal framework.

The neutrosophic logic concept [9-18] is an important notion that addresses, among other ones, the issue of data inaccuracies. Neutrosophic logic is a theoretical framework that expands on classical and fuzzy logic to incorporate indeterminacy, uncertainty, and ambiguity in greater depth. It is especially significant when it comes to addressing data inaccuracies, and it allows for the representation of not only binary values (true and false), but also values that are uncertain or undefined, recognizing the inherent inaccuracies in real-world data. Therefore, it improves the reliability of decision-making processes, particularly in areas, like machine learning and data analysis, by offering a systematic method to deal with uncertainty. This is particularly important since a precise representation of uncertainties is vital in these domains. The shift in paradigm towards a more refined depiction of truth values provides a valuable means of addressing the difficulties emerged by imperfections in datasets, thus enhancing the accuracy and dependability of outcomes in data-driven applications. Within this specific context, this study aimed to present a strategy that utilizes neutrosophic theory to address data imprecision and ameliorate prediction models in real-world scenarios.

II. MACHINE LEARNING AND REGRESSION TECHNIQUES

Machine learning is at the forefront of a data analysis and predictive modeling revolution, providing deep insights in diverse disciplines, namely finance, healthcare, marketing, and environmental sciences. Regression approaches are highly effective tools in machine learning to understand and forecast continuous results. These tools allow us to see patterns in the data, build connections between variables, and predict future events with exceptional precision. Regression models serve as the foundation for analyzing various phenomena, such as forecasting stock market patterns, estimating patient recovery durations, and predicting customer buying schemes. This section focuses on five distinct regressors, each of them having a distinct ability to model intricate relationships within the data. Consequently, these regressors improve decision-making processes and strategic planning by providing more comprehensive insights.

A. Linear Regression

Linear Regression [19] is a fundamental and traditional regression method that seeks to represent the correlation between a dependent variable and one or more autonomous ones by adjusting a linear equation to the collected data. The equation can be expressed as $Y = aX + b$, where Y is the dependent variable, X is the independent variable, a represents the slope, and b is the intercept. Linear Regression is most effective when there is a linear relationship between the independent and the dependent variables, where the data should be relatively free from noise. It is also efficient when the amount of features is not excessively large, as it can be susceptible to overfitting situations in areas with a high number of dimensions.

B. Ridge Regression

Ridge regression is a form of linear regression that incorporates a regularization term, often known as the L2 penalty, and is obtained by taking the squared magnitude of the coefficients [20]. Multiple regression is used when the data exhibit multicollinearity, which refers to a situation where the independent variables are strongly correlated. Ridge Regression is effective in situations where it is necessary to prevent overfitting and when the number of predictor variables surpasses the number of observations.

C. Support Vector Regression (SVR)

SVR is a machine-learning approach that is derived from the Support Vector Machine (SVM) algorithm [21]. It identifies the optimal line, or hyperplane in higher dimensions, which includes the highest number of points within a specific threshold margin. SVR demonstrates efficacy in handling nonlinear interactions by utilizing kernel tricks. Furthermore, SVR exhibits proficiency in both small and large datasets. In addition, it performs exceptionally well when there is a distinct separation between the data points and any outliers present.

D. Random Forest Regression

Random Forest Regression is an ensemble learning technique that utilizes decision tree regressors [22]. The system functions by generating numerous decision trees throughout the training process and producing the average prediction of each individual tree. This technique demonstrates great accuracy when the dataset comprises a combination of category and numerical variables and when the data exhibit intricate structures that are beyond the capability of basic models, such as linear regressors, to capture. Additionally, it displays resilience against overfitting, particularly when dealing with data that contain noise.

E. Gradient Boosting Regression

The Gradient Boosting Regressor is an ensemble technique that uses decision trees as its foundation [20]. Unlike Random Forest, which constructs trees simultaneously, Gradient Boosting constructs one tree at a time in a sequential manner, with each subsequent tree aiming to rectify the mistakes produced by the previously trained ones. Optimizing an arbitrary differentiable loss function enhances the performance of the model. Gradient Boosting is particularly effective in situations where the data are uneven and the main focus is on achieving high predicted accuracy. It is highly efficient for a broad spectrum of regression problems, particularly when the connection between the input and target variables is intricate and nonlinear. Nevertheless, meticulous adjustment of its parameters is necessary, as it has the potential to overfit in datasets with noise if it is not adequately regularized.

Each of these regressors possesses distinct advantages and optimal scenarios for usage. Linear Regression is highly competent in generating rapid and straightforward models and interpretations, particularly when the data exhibit linearity. Ridge Regression is a preferred method for addressing multicollinearity. SVR is proficient at accurately representing nonlinear relationships. Random Forest and Gradient Boosting are adept at identifying complex correlations within the data, but they demand substantial computational resources and may exhibit overfitting tendencies if not appropriately fine-tuned. The accuracy of a regression model is determined by the degree to which its predictions align with the observed data. It can be affected by the selection of model parameters, the accuracy of the input data, and the degree to which the assumptions of the regression model align with the actual relationships in the data. It is advisable to evaluate the model's performance on data that have not been seen before and adjust the model's hyperparameters to optimize its prediction accuracy.

III. BACKGROUND ON NEUTROSOPHIC SET THEORY

In the realm of neutrosophic logic [6-15], the concepts of truth, indeterminacy, and falsity degrees form the foundational pillars for understanding and analyzing information that is imprecise, incomplete, or inconsistent. These degrees provide a more nuanced approach to reasoning under uncertainty compared to traditional binary logic. The truth degree represents the extent to which a statement is considered genuine. Unlike binary logic, where a statement can only be true or false, the truth degree allows for varying levels of truthfulness, acknowledging that knowledge about the world is often partial or probabilistic. The indeterminacy degree measures the level of uncertainty or lack of information regarding the truth value of a statement. This is particularly important in scenarios where information is missing, ambiguous, or the existing knowledge about a situation is incomplete, making it difficult to fully ascertain the veracity of a statement. The falsity degree quantifies the extent to which a statement is considered false. Just as with the truth degree, the falsity degree embraces the complexity of real-world information, where statements may not be entirely false but contain elements of falsehood. These three degrees are not mutually exclusive and are designed to capture the multifaceted nature of truth in the real world, where statements can be partially true, partially false, and partially indeterminate at the same time. Thus, the neutrosophic set theory is seen as a generalized extension of fuzzy and intuitionistic fuzzy theories. Furthermore, it exhibits a higher degree of resemblance to human cognition, since it more accurately replicates human decision-making processes by considering evidence that involves uncertainty. Following the first introduction of the concept of neutrosophic sets [9], other extensions have been suggested, the most renowned being the single-value neutrosophic set [12]. Single-valued neutrosophic numbers are a subset of single-valued neutrosophic sets and play a crucial role in neutrosophic multiattribute decision-making situations, as they accurately represent an uncertain quantity [14]. Neutrosophic sets and their extensions have been applied in many sectors [11].

Triangular neutrosophic numbers are a significant extension of neutrosophic theory [17]. This triangle structure enables a more comprehensive representation of ambiguity and incomplete truths, which frequently arise in real-life situations where facts and human perceptions are rarely absolute. Triangular neutrosophic numbers offer a strong instrument for intricate decision-making procedures that require a subtle approach to ambiguity and subjectivity [16]. A triangular neutrosophic number is defined as $\tilde{a} = \langle (a^1, a, a^2), t_{\tilde{a}}, i_{\tilde{a}}, f_{\tilde{a}} \rangle$, incorporating lower bound a^1, median value a, and upper bound a^2 besides measures of truth, indeterminacy, and falsity [15]. For de-neutrosiphication, the variation degree of $\tilde{a}$ is:

$$\theta_{\tilde{a}(\alpha,\beta,\gamma)} = \frac{1}{4}\left[\frac{\alpha}{t_{\tilde{a}}} + 2\frac{(1-\beta)}{1-i_{\tilde{a}}} + \frac{(1-\gamma)}{1-f_{\tilde{a}}}\right], \theta_{\tilde{a}(\alpha,\beta,\gamma)} \in [0,1] \quad (1)$$

where α denotes the minimal degree of acceptance, $\alpha \in [0, t_{\tilde{a}}]$, β denotes the maximal degree of indeterminacy, $\beta \in [d_{\tilde{a}}, 1]$, and γ denotes the maximal degree of rejection, $\gamma \in [f_{\tilde{a}}, 1]$. The proposed approach posits that a triangular neutrosophic number, represented as $\tilde{a} = \langle (a^1, a, a^2), t_{\tilde{a}}, i_{\tilde{a}}, f_{\tilde{a}} \rangle$ can be simplified to its equivalent interval value in the following manner:

- When the acceptance degree α is at its minimal level, and the indeterminacy and rejection degrees β and γ are at their greatest levels, the value of $\tilde{a}$ will fall within the range defined by its lower and upper bounds, thus $\tilde{a} = [a^1, a^2]$.
- When the acceptance degree α is at its highest level and the indeterminacy and rejection degrees β and γ are at their lowest levels, the value of $\tilde{a}$ will tend to be a specific crisp value, which is logically considered to be equal to its median value. Therefore, $\tilde{a} = [a, a] = a$.

The neutrosophic theory is a notable expansion of classical and fuzzy logic that addresses the inherent ambiguity and uncertainty of real-world data. The mathematical framework offers a means of representing uncertain, inconsistent, and incomplete information by utilizing its key elements: truth, indeterminacy, and falsity. Triangular neutrosophic numbers are a specific form of representation in this theory, distinguished by a three-part structure that measures the extent of truth, indeterminacy, and falsehood for each element. This structure is highly skilled at representing the intricate range of real-life situations where data frequently deviate from binary or fuzzy logic.

IV. USING NEUTROSOPHIC VARIABLES TO DEAL WITH DATA IMPRECISION

The adoption of a neutrosophic paradigm allows for a more accurate depiction of fine details in empirical data. Utilizing neutrosophic variables is a deliberate and calculated reaction to the widespread existence of uncertain and contradictory features within datasets. It expands the range of analysis, allowing predictive systems to understand and combine the inherent uncertainties of the data. This enhances the decision-making process by providing a variety of insights that cannot be shown by traditional binary logic.

Consider $A = (a_1, \ldots, a_n)$ to be a vector of reals that represents a feature. Consider a_i as a component of A that represents one experiment value of the feature A. a_i^1 and a_i^2 are proposed to be computed as follows.

$$a_i^1 = a_i(1 - |CVar(A)|)$$

$$a_i^2 = a_i(1 + |CVar(A)|)$$

where $CVar(A) = \left|\frac{\bar{X}(A)}{\delta_A}\right|$, $\bar{X}(A)$ represents the mean of A, and δ_A represents the standard deviation. The coefficient of variation $CVar(A)$ is a normalized metric that quantifies the extent of variability in probability or frequency distributions, and it is commonly used to assess the level of variation among distinct datasets, even if their means differ significantly.

The proposed method involves using $CVar(A)$ to modify each component a_i of the feature vector A, resulting in the creation of two additional components, a_i^1 and a_i^2. This adjustment accounts for the fluctuation in the characteristic compared to its average value. A higher coefficient of variation corresponds to larger modifications made to a_i, indicating a greater dispersion of data. On the other hand, a low coefficient of variation results in smaller adjustments, indicating that the data points are closer to the mean. Each component a_i is then represented by the triangular neutrosophic variable as follows:

$$\tilde{a}_i = \langle (a_i^1, a_i, a_i^2), t_{a_i}, i_{\tilde{a}_i}, f_{\tilde{a}_i} \rangle$$

where t_{a_i} is the truth degree related to a_i, $i_{\tilde{a}_i}$ is the indeterminacy degree related to a_i, and $f_{\tilde{a}_i}$: is the falsity degree related to a_i. By integrating $CVar(A)$ into the neutrosophic variable representation, the relative dispersion in the data, which may improve the model's ability to manage and interpret feature fluctuations, can be considered. The model has the potential to enhance accuracy in regression analysis by acknowledging and incorporating the inherent variability in the data into the forecasting process. It is crucial to include this aspect in datasets where the variability greatly affects the behavior of the dependent variable being modeled.

To comprehend each crisp component a_i or its substitute (a_i^1, a_i, a_i^2) in a dataset, the former is linked to a triplet $(t_{a_i}, i_{a_i}, f_{a_i})$ that denotes its levels of truth, indeterminacy, and falsity, respectively. To make this procedure more efficient, one might select a predetermined set of configurations that accurately represent the core characteristics of each variable. This collection was carefully chosen by an expert who has a thorough understanding of the subject matter. It is therefore ensured that the given values for truth, uncertainty, and falsehood are perfectly linked with the variable's real-world meanings and behaviors. This is particularly important when data are collected from many sources that differ in terms of quality and precision. Neutrosophic logic extends the basic notions of truth, indeterminacy, and falsehood by introducing the ideas of minimal degree of acceptance α, maximal degree of indeterminacy β, and maximal degree of rejection γ. These concepts help to enhance the examination of assertions in situations of uncertainty.

The notion of a minimal degree of acceptance establishes the minimum level at which a proposition is deemed sufficiently truthful within a specific context. The purpose of this criterion is to exclude irrelevant or unreliable information and prioritize content that passes a minimum standard of reliability or relevance. The maximal degree of indeterminacy refers to the highest level of uncertainty that can be accepted in a decision-making process. It clearly defines the border between acceptable and excessive vagueness, guaranteeing that judgments are made with the highest level of knowledge feasible. The maximal degree of rejection sets the upper limit for a statement to be considered false. It acts as a standard for filtering out mostly erroneous information, allowing for a more precise differentiation between helpful and misleading information.

The expression θ (1) reflects the overall assessment of the variable within the neutrosophic context. Integrating truth, indeterminacy, and falsity degrees with the minimal degree of acceptance, maximal degree of indeterminacy, and maximal degree of rejection provides a robust framework for navigating the complexities of real-world information. By quantifying the nuances of truthfulness, uncertainty, and falseness, neutrosophic logic allows for a more sophisticated analysis of statements, facilitating informed decision-making even in the presence of incomplete or contradictory information. This holistic approach underscores the adaptability of neutrosophic logic in addressing the challenges of uncertainty, making it a valuable tool in fields ranging from artificial intelligence and data science to philosophy and theoretical mathematics. The final step involves adjusting the original variable a_i using the neutrosophic assessment to obtain two de-neutrosophied variables a_i^l and a_i^u as follows:

$$a_i^l = a_i^1 + \theta_{\tilde{a}_i}(a_i - a_i^1)$$

$$a_i^u = a_i^2 + \theta_{\tilde{a}_i}(a_i^2 - a_i)$$

The de-neutrosophied variables a_i^l and a_i^u serve as more nuanced representations that consider the variable's corresponding levels of truth, indeterminacy, and falsity. In other words, these two variables represent the lower and upper bounds, respectively, that could be considered projections of $\tilde{a}$ via the proposed de-neutrosophication procedure. By incorporating neutrosophic factors into predictive models, the accuracy is expected to be enhanced as these variables enable a more comprehensive depiction of data. Neutrosophic logic captures additional intricacies of the real world that could be overlooked by classical precise and crisp values. By carefully evaluating several factors, this approach allows the model to recognize and adapt to the underlying complexities and uncertainties in the data, resulting in predictions that better reflect the intricate nature of reality.

V. CASE STUDY: HAPPINESS INDICATOR

The dataset used for this case study was derived from the annual happiness report and the country statistics report published by the UN [21]. The former consists of a variety of measures that evaluate the welfare and contentment of inhabitants in different countries. This dataset encompasses various measures, including Gross Domestic Product (GDP) per capita, Economy Services, Agricultural Production Index, Infant Mortality, Health Total Expenditure, CO2 Emission, Energy Production, and several more indicators (a total of 38 indicators). These metrics jointly contribute to the calculation of a happiness score for each country. The specific dataset also classifies countries according to their regions and provides demographic context to augment the examination of happiness levels.

This study focused on the GDP metric as one of the most important key features of happiness. The proposed method was applied using three configurations, as shown in Table I, related to the truth, indeterminacy, and falsity values. These configurations were assigned to each measurement of GDP, mainly according to its belonging region.

TABLE I. USED CONFIGURATIONS OF TRUTH, INDETERMINACY, AND FALSITY

	Truth	Indeterminacy	Falsity
Configuration 1	0.9	0.2	0.2
Configuration 2	0.6	0.4	0.3
Configuration 3	0.2	0.7	0.4

By incorporating neutrosophic logic, traditional GDP values are transformed into a pair of neutrosophic variables that capture the inherent uncertainty and variability within the economic data. New variables, represented as GDP^l and GDP^u, were derived from the original GDP values while adjusting for the degree of acceptance, indeterminacy, and rejection ascribed to each data point by an expert. In the first version of the analysis, traditional regression methods were employed to predict happiness scores using the original dataset, which included GDP as a conventional, crisp variable. Linear Regression, Ridge Regression, SVR, Random Forest regression, and Gradient Boosting regression were among the techniques used. These regressors drew from the raw data, aiming to capture the direct relationship between GDP and the reported levels of happiness in different countries. This approach is grounded in classical statistical analysis, where the impact of each predictor is considered in its original form, with the assumption that the relationships between variables are linear and directly observable.

For the second version, the GDP variable was substituted with two new de-neutrosophied values derived from the neutrosophic logic-based equations. Not only did these new values aim to encapsulate the economic output, but also the relative uncertainty and perception variations in economic well-being that GDP figures might imply. The same regressors were then applied to this transformed dataset. The prediction models could potentially account for the nuanced influences of economic factors on happiness by considering these broader implications. Figure 1 illustrates the R^2 scores obtained with both experiments.

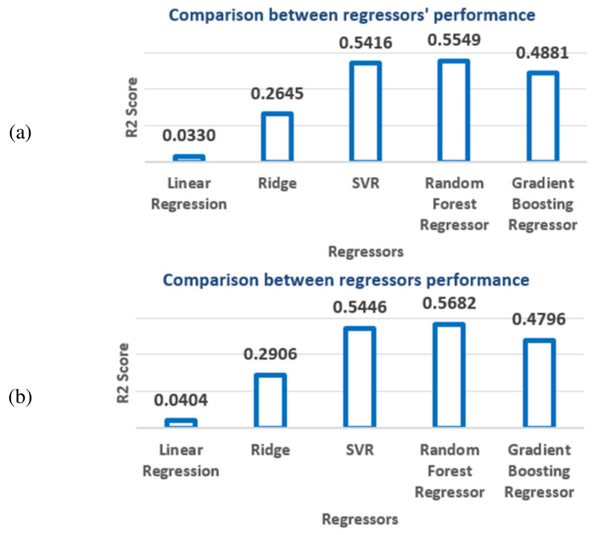

Fig. 1. R^2 score with (a) original dataset, and (b) updated dataset with substituted GDP with GDP^l and GDP^u.

This technique demonstrates a more dynamic and possibly meaningful understanding of the data, suggesting that the overall welfare of a country's inhabitants may not be adequately represented solely by conventional economic measures. The purpose was to determine whether these augmented data could improve the accuracy of the predictions and provide a more profound understanding of the factors that influence happiness. The varying prediction performance using the original dataset and the updated dataset is portrayed in Tables II and III, respectively.

TABLE II. PERFORMANCE WITH ORIGINAL DATASET

Applied Regressor	MSE	RMSE	R^2
Linear Regression	1.31506912	1.14676463	0.03297209
Ridge	1.00025007	1.00012503	0.26447232
SVR	0.62337761	0.78954266	0.54160314
Random Forest	0.60534793	0.77804108	0.55486116
Gradient Boosting	0.69617557	0.83437136	0.48807162

TABLE III. PERFORMANCE WITH UPDATED DATASET

Applied Regressor	MSE	RMSE	R^2
Linear Regression	1.30499936	1.14236569	0.04037682
Ridge Regression	0.96465932	0.98217072	0.29064376
SVR	0.61924131	0.78691887	0.54464475
Random Forest	0.58726705	0.76633351	0.56815682
Gradient Boosting	0.70768057	0.84123752	0.47961149

When comparing the results of the two experiments, traditional prediction models using original data versus models employing neutrosophic logic-enhanced variables, a clear distinction in performance emerges. The first experiment, which deployed original GDP data, demonstrated a range of effectiveness across different types of regression. Linear Regression, constrained by its simplicity and assumption of linearity, lagged in performance. Ridge Regression offered moderate improvements, possibly due to its penalization of coefficient size, which helps to manage multicollinearity. SVR, Random Forest, and Gradient Boosting, known for handling nonlinear patterns and complex relationships, performed better. In particular, the Random Forest and SVR models stood out, indicating their robustness in handling the dataset's complexity. The second experiment introduced the neutrosophic logic-enhanced GDP variables, aiming to capture the uncertainty and vagueness inherent in economic indicators. This transformation was intended to provide richer, more nuanced input for the regressors. The results disclosed that the Random Forest regressor using the new variables had an improvement in the R^2 score, suggesting a better model fit. The robustness of the Random Forest to noise and its ability to handle nonlinear relationships possibly contributed to this improved performance. Conversely, Linear Regression's minimal change in performance reaffirmed its limitations when dealing with complex, transformed datasets. Gradient Boosting and SVR also maintained high performance, indicating their ability to capitalize on nuanced data, although they did not show significant improvement over the traditional method. This could suggest that, while the new variables added information, they did not substantially alter the predictive relationships these models were able to capture.

All in all, the introduction of neutrosophic logic-based variables appeared to bolster the performance of models capable of handling complexity and nonlinearity. The enhanced performance of the Random Forest regressor points to its suitability in scenarios where the independent variables are enriched with additional interpretative layers. Linear Regression's underperformance underscores the importance of model selection in predictive analytics, as simpler models may not suffice when variables capture more profound economic and psychological constructs, such as those implied by neutrosophic logic.

VI. CONCLUSION

In the realm of predictive analytics, accurately modeling and forecasting subjective measures is a complex problem. Traditional linear models that utilize crisp data often fail to capture the multifaceted nature of such phenomena. This study aimed to address the limitations discussed above by introducing a novel approach that incorporates neutrosophic logic into the predictive modeling process, transforming a crisp variable into a pair of de-neutrosophied values. This novel approach was applied on a happiness assessment dataset based on multiple parameters, including GDP. The GPD crisp values were modified adopting the proposed de-neutrosophication technique. Not only was the purpose of this process to encompass economic productivity, but also to incorporate the relative unpredictability and variability in the perception of economic welfare. The results of this method were promising. Compared to traditional regression models, the neutrosophic method showed an improvement in model performance, particularly with the Random Forest regressor, which demonstrated a significant increase in the R^2 score. This indicates a stronger correlation between the predictive model and the actual happiness scores, suggesting that neutrosophic variables provide a more accurate and nuanced reflection of the underlying economic factors that contribute to happiness.

The conclusions drawn from these experiments are twofold. First, the application of neutrosophic logic in predictive modeling can offer a meaningful improvement in the accuracy of predictions, especially in complex and subjective contexts, such as happiness indices. Second, the choice of regressor is crucial. Models that can leverage the complexity and nuances introduced by neutrosophic variables, like the Random Forest regressor, tend to perform better. Although these findings demonstrate the potential of neutrosophic logic, caution should be exercised in generalizing these conclusions to the broader field of machine learning prediction models. There is a need for further investigation and exploration in diverse domains to validate the applicability of neutrosophic set theory across a wider range of predictive models.

ACKNOWLEDGMENT

The authors gratefully acknowledge the approval and support of this research from the Deanship of Scientific Research, Northern Border University, Arar, Kingdom of Saudi Arabia by the CSCR-2022-11-1243 grant.

REFERENCES

[1] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, https://doi.org/10.1126/science.aaa8415.

[2] R. Trinchero and F. Canavero, "Machine Learning Regression Techniques for the Modeling of Complex Systems: An Overview," *IEEE Electromagnetic Compatibility Magazine*, vol. 10, no. 4, pp. 71–79, 2021, https://doi.org/10.1109/MEMC.2021.9705310.

[3] C. J. M. Maas and J. J. Hox, "Robustness issues in multilevel regression analysis," *Statistica Neerlandica*, vol. 58, no. 2, pp. 127–137, 2004, https://doi.org/10.1046/j.0039-0402.2003.00252.x.

[4] Y. E. Touati, "Deadline Verification for Web Services Using Timed Automata," *Engineering, Technology & Applied Science Research*, vol. 12, no. 1, pp. 8013–8016, Feb. 2022, https://doi.org/10.48084/etasr.4611.

[5] U. Khan, K. Khan, F. Hassan, A. Siddiqui, and M. Afaq, "Towards Achieving Machine Comprehension Using Deep Learning on Non-GPU Machines," *Engineering, Technology & Applied Science Research*, vol. 9, no. 4, pp. 4423–4427, Aug. 2019, https://doi.org/10.48084/etasr.2734.

[6] B. Trstenjak, D. Donko, and Z. Avdagic, "Adaptable Web Prediction Framework for Disease Prediction Based on the Hybrid Case Based Reasoning Model," *Engineering, Technology & Applied Science Research*, vol. 6, no. 6, pp. 1212–1216, Dec. 2016, https://doi.org/10.48084/etasr.753.

[7] R. Chen and I. C. Paschalidis, "A Robust Learning Approach for Regression Models Based on Distributionally Robust Optimization," *Journal of Machine Learning Research*, vol. 19, no. 13, pp. 1–48, 2018.

[8] E. Hüllermeier, "Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization," *International Journal of Approximate Reasoning*, vol. 55, no. 7, pp. 1519–1534, Oct. 2014, https://doi.org/10.1016/j.ijar.2013.09.003.

[9] F. Smarandache, *A Unifying Field in Logics: Neutrosophic Logic*. American Research Press, 1999.

[10] F. Smarandache, A Unifying Field in Logics. Neutrosophy: Neutrosophic Probability, Set and Logic. American Research Press, 1999.

[11] F. Smarandache and S. Pramanik, *New Trends in Neutrosophic Theory and Applications*. Pons Editions, 2016.

[12] H. Wang, Florentin Smarandache, Y. Zhang, and R. Sunderraman, "Single valued neutrosophic sets," in *Collected Papers. Volume XIV: Neutrosophics and other topics*, Global Knowledge, 2022.

[13] C. H. Wang, C. C. Chuang, and C. C. Tsai, "A fuzzy DEA–Neural approach to measuring design service performance in PCM projects," *Automation in Construction*, vol. 18, no. 5, pp. 702–713, Aug. 2009, https://doi.org/10.1016/j.autcon.2009.02.005.

[14] I. Deli and Y. Şubaş, "A ranking method of single valued neutrosophic numbers and its applications to multi-attribute decision making problems," *International Journal of Machine Learning and Cybernetics*, vol. 8, no. 4, pp. 1309–1322, Aug. 2017, https://doi.org/10.1007/s13042-016-0505-3.

[15] W. Abdelfattah, "Data envelopment analysis with neutrosophic inputs and outputs," *Expert Systems*, vol. 36, no. 6, 2019, Art. no. e12453, https://doi.org/10.1111/exsy.12453.

[16] M. Abdel-Basset, M. Mohamed, Y. Zhou, and I. Hezam, "Multi-criteria group decision making based on neutrosophic analytic hierarchy process," *Journal of Intelligent & Fuzzy Systems*, vol. 33, no. 6, pp. 4055–4066, Jan. 2017, https://doi.org/10.3233/JIFS-17981.

[17] M. Abdel-Basset, M. Mohamed, A. N. Hussien, and A. K. Sangaiah, "A novel group decision-making model based on triangular neutrosophic numbers," *Soft Computing*, vol. 22, no. 20, pp. 6629–6643, Oct. 2018, https://doi.org/10.1007/s00500-017-2758-5.

[18] M. Abdel-Basset, M. Mohamed, and F. Smarandache, "An Extension of Neutrosophic AHP–SWOT Analysis for Strategic Planning and Decision-Making," *Symmetry*, vol. 10, no. 4, Apr. 2018, Art. no. 116, https://doi.org/10.3390/sym10040116.

[19] G. A. F. Seber and A. J. Lee, *Linear Regression Analysis*. John Wiley & Sons, 2012.

[20] G. C. McDonald, "Ridge regression," *WIREs Computational Statistics*, vol. 1, no. 1, pp. 93–100, 2009, https://doi.org/10.1002/wics.14.

[21] M. Awad and R. Khanna, "Support Vector Regression," in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, M. Awad and R. Khanna, Eds. Berkeley, CA, USA: Apress, 2015, pp. 67–80.

[22] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geology Reviews*, vol. 71, pp. 804–818, Dec. 2015, https://doi.org/10.1016/j.oregeorev.2015.01.001.

[23] S. Peter, F. Diego, F. A. Hamprecht, and B. Nadler, "Cost efficient gradient boosting," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[24] Elai, "Elaishalev/Countries_Happiness." Mar. 17, 2020, [Online]. Available: https://github.com/Elaishalev/Countries_Happiness.