Neutrosophic Hybrid Machine Learning Algorithm for Diabetes Disease Prediction

A. Bermúdez del Sol[1,*], Edison Sotalin Nivela[1], Edwin Miranda Solis[1], Yasser H. Elawady[2]

[1]Universidad Regional Autónoma de los Andes, Ecuador

[2]Computer Engineering Department, Misr Higher Institute for Engineering and Technology, Mansoura, Egypt

Emails: ua.abdelbermudez@uniandes.edu.ec; us.medicina@uniandes.edu.ec; ua.edwinmiranda@uniandes.edu.ec; y.alawady@engmet.edu.eg

Abstract

Because of its far-reaching effects, diabetes remains a major health problem on a worldwide scale. It's a metabolic illness that causes hyperglycemia and a host of other health issues, including cardiovascular disease, renal failure, and neuropathy. Many scientists have spent time and energy over the years trying to develop a reliable diabetes prediction model. Researchers are forced to adopt big data analytics and machine learning (ML)-based methodologies since there are still major open research concerns in this area owing to a lack of acceptable data sets and prediction techniques. This study seeks solutions by way of an examination of healthcare predictive analytics. The major purpose of this research was to explore the potential applications of big data analytics and machine learning-based approaches in the field of diabetes. In this study, we used the neutrosophic AHP as a feature selection method. The neutrosophic AHP is used to compute the importance of features, then apply the machine learning methods to these features. This study applied logistic regression, support vector machine (SVM), and random forest (RF) to predict the disease of diabetes.

Keywords: Machine Learning; Diabetes Disease; Neutrosophic, Random Forest; Support Vector Machine; Feature Selection.

1. Introduction

These days, diabetes is a major killer in third-world nations. Both authorities and private citizens are funding medical research to discover a cure for the deadly illness. In those with diabetes, the inability to produce enough insulin causes the amount of sugar in their blood to steadily increase. Diabetic individuals have trouble converting the glucose sugar they ingest into the energy they need to carry out their daily tasks. This causes blood sugar levels to rise gradually over time. Since not all cells in the body can absorb glucose, it stays in the circulation[1]–[3].

The process of diagnosis is the backbone of modern medicine. They approach it from various angles. This approach involves sorting a set of data into categories according to predetermined rules[4], [5]. Several variables may determine whether a person has diabetes. Diabetes is a disease caused by the body's failure to release enough insulin. Symptoms of an insulation problem include increased thirst and hunger, elevated blood sugar, and a need to urinate often. Untreated diabetes often leads to other issues that worsen health. To lessen the impact of diabetes, accurate early detection is essential[6]–[8].

DOI: https://doi.org/10.54216/IJNS.210207
Received: February 19, 2023 Revised: April 27, 2023 Accepted: May 28, 2023

Predictive analysis may range from simple observations to complex machine learning techniques. This method involves categorizing historical data in order to learn anything about the future. Predictive analysis may benefit greatly from the use of machine learning and regression approach. Machine learning is often regarded as the most important component of AI. It helps the computer system learn from the history without being explicitly programmed. Therefore, machine learning is the best option for minimizing manual labor since it allows for automation with almost no room for mistake[9]–[11].

Machine learning yields preferred findings for diabetes diagnosis. Logistic Regression, Random Forest Classifier, and Support Vector Machine are all examples of different machine learning algorithms[12], [13]. The purpose of this study is to offer a machine learning model and neutrosophic AHP method for estimating the likelihood that a person has diabetes[14], [15].

The neutrosophic AHP method used as a feature selection method to select feature as a input to machine learning method. One of the most well-known MCDM methods, AHP was introduced by Thomas L. Saaty in 1980. In this approach, the interdependence of the criteria is taken into account. Real-world criteria rely on one another; therefore, their interdependence must be taken into consideration during decision-making. AHP is useful for complex decision-making situations that include several qualitative and quantitative factors. The foundation of the AHP method is the generation of a comparison matrix based on the weights assigned to each pair of factors. The technique's utilization of the joy analogy simplifies decision making and computation. Figure 1 shows the proposed model[16], [17].

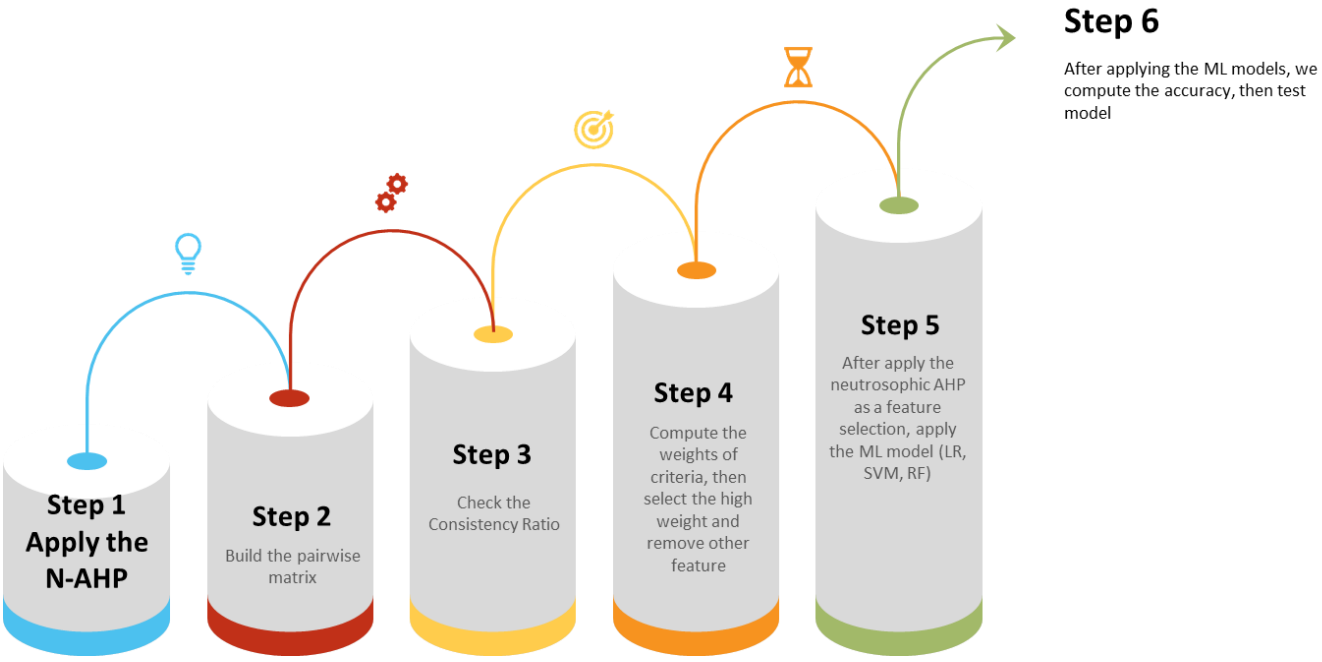

Figure 1: The steps of the proposed model.

2. Machine Learning Models

2.1 Logistic Regression (LR)

The statistics department's LR algorithms have been got. This approach has been modified to work with statements involving issues with binary classification. The primary goal of LR is to find the coefficient values. The LR normalizes the range to 0-1 for use. The LR model decides if the provided data instance belongs to the class with a 0 or 1 probability. while we have several factors to consider while making a prediction, this method might be useful[18], [19].

DOI: https://doi.org/10.54216/IJNS.210207
Received: February 19, 2023 Revised: April 27, 2023 Accepted: May 28, 2023

$$LR = \frac{1}{1+e^{-(\gamma_0+\gamma_0 x)}}$$

Where x refers to the dataset, and γ_0 refers to the constant parameter of LR

2.2 Support Vector Machine (SVM)

The technique of SVM is an example of supervised ML. If your data collection is relatively small and has some outliers, this approach may be a good fit. Finding the hyperplane that best separates the data points is the key. The determined hyperplane will partition a pair of spaces into discrete regions. Similar information kinds will cluster together in such a domain[20].

$$|| x \ || = \sqrt{x_1^2 + x_2^2 + x_3^2 + x_4^2 + \cdots \ldots . x_n^2}$$

2.3 Random Forest

The supervised learning framework includes the ML technique known as "random forest." Multiple decision trees representing the many classes in the input data set make up the RF classifier. The mean of the subsets from each tree is used to increase the accuracy of predictions. Instead than relying on just one decision tree, RF uses the predictions of all of them to arrive at a final forecast. A data-related inquiry is carried out at each node in the decision tree[21].

3. Neutrosophic Feature Selection

It's possible that the initial collection of features had both input features that are pertinent to the future of the system and those that are not, as well as redundant features[22], [23]. In order to ensure that only useful features are kept for creation of models, the choice of features is used to weed out the superfluous ones. Filter methods, wrapper methods, and embedding methods are the most common types of feature selection approaches[5], [24]–[26].

Better judgements may be made in the actual world with the aid of Neutrosophic Sets (NS), a generalization of crisp, fuzzy, and intuitionistic sets. Decisions are better characterized when people shift their preferences among the truth, falsity, and indeterminacy functions, even if human thought is not usually articulated in crisp numbers and may be unclear, insufficient, incompatible, vague, etc. Multi-Criteria Decision Making (MCDM) assignment approaches for dependability have hitherto ignored such data. The resulting relative weight for every component is then utilized in an optimization problem to distribute reliabilities to components in a way that maximizes system-wide software dependability[27].

To combine developer and user feedback, AHP uses a hierarchy developed by Zahedi and Ashrafi. In a Neutrosophic setting, consumers, technicians, and developers compare functions, programmers, and modules at various hierarchical levels to account for and make use of the accessible knowledge's inherent fuzziness, imprecision, and inconsistency[28], [29].

4. Results

The training dataset is partitioned when all processing is complete. There were four different ML classification algorithms used. To get the best outcomes for the available data, we used hyper-parameter tweaking and cross-validation. The LR, SVM, and RF machine learning techniques were used, as was previously mentioned. Below, we detail the models we ran, the hyper-parameters we tuned, and the outcomes we found. Metrics like the B1 score, recall, precision, and accuracy are used to assess ML algorithmic effectiveness. Here is the equation in question:

$$PPV = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

DOI: https://doi.org/10.54216/IJNS.210207
Received: February 19, 2023 Revised: April 27, 2023 Accepted: May 28, 2023

$$Accuracy = \frac{True\ Positive + Trure\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} * 100$$

$$Specificity = \frac{True\ Negative}{True\ Positive + False\ Positive} * 100$$

First apply the N-AHP method as a feature selection method. Start with building the pairwise comparison matrix by experts. Then normalize the pairwise comparison as shown in Table 1. Then compute the weights of criteria as shown in Figure 2. From Figure 2, the lowest criterion will remove.

Table 1: Normalization matrix by N-AHP.

	FSC_1	FSC_2	FSC_3	FSC_4	FSC_5	FSC_6	FSC_7	FSC_8	FSC_9
FSC_1	0.047941	0.017704	0.018047	0.028469	0.03607	0.080502	0.080202	0.040626	0.140034
FSC_2	0.159803	0.059013	0.027317	0.015629	0.05133	0.067961	0.064037	0.060483	0.149837
FSC_3	0.217913	0.177216	0.082033	0.073081	0.035714	0.091839	0.031011	0.16141	0.050568
FSC_4	0.123559	0.277056	0.082362	0.073374	0.018038	0.091839	0.029942	0.061212	0.140034
FSC_5	0.068497	0.05925	0.118373	0.209641	0.051536	0.058063	0.010961	0.067224	0.050724
FSC_6	0.060769	0.088608	0.091147	0.081527	0.090573	0.102043	0.032883	0.157767	0.140034
FSC_7	0.053268	0.082122	0.235726	0.218376	0.418991	0.27654	0.089113	0.066677	0.033141
FSC_8	0.214982	0.17775	0.092588	0.218376	0.139664	0.117833	0.243479	0.182179	0.140034
FSC_9	0.053268	0.06128	0.252408	0.081527	0.158085	0.113381	0.418372	0.202421	0.155594

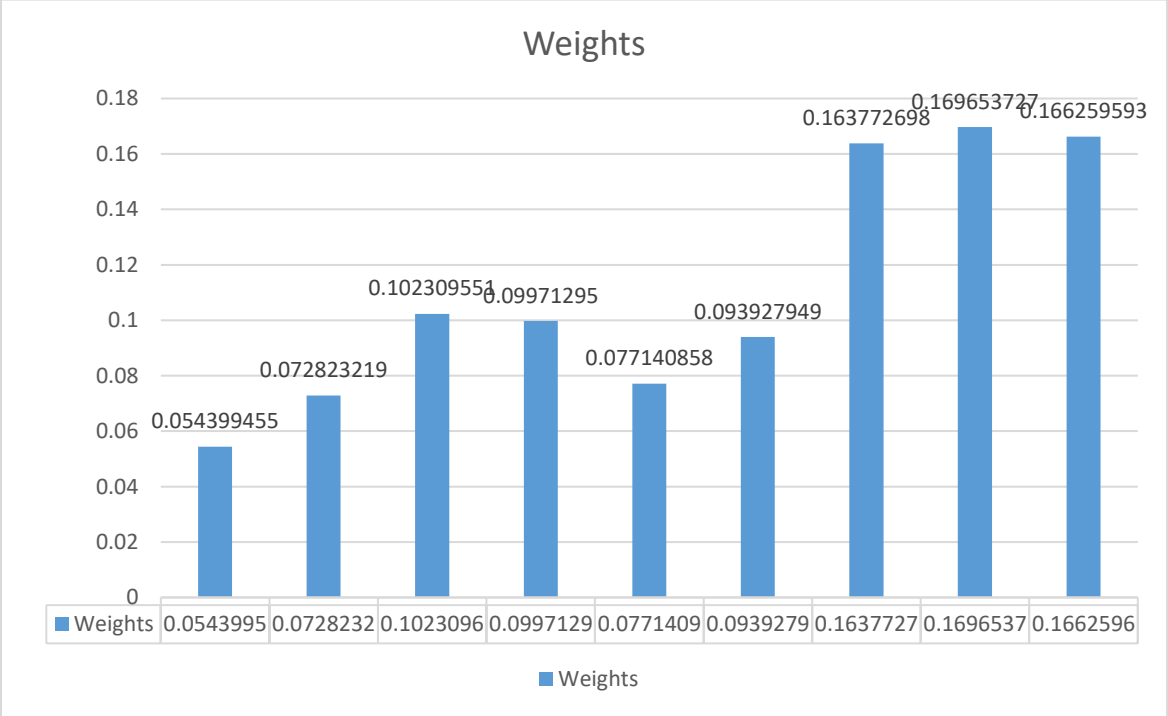

Figure 2: The feature importance by the N-AHP.

DOI: https://doi.org/10.54216/IJNS.210207
Received: February 19, 2023 Revised: April 27, 2023 Accepted: May 28, 2023

Then apply the machine learning model on diabetes disease dataset from Kaggle. The dataset has nine features. Figure 3 shows the data in the outcome for two classes.

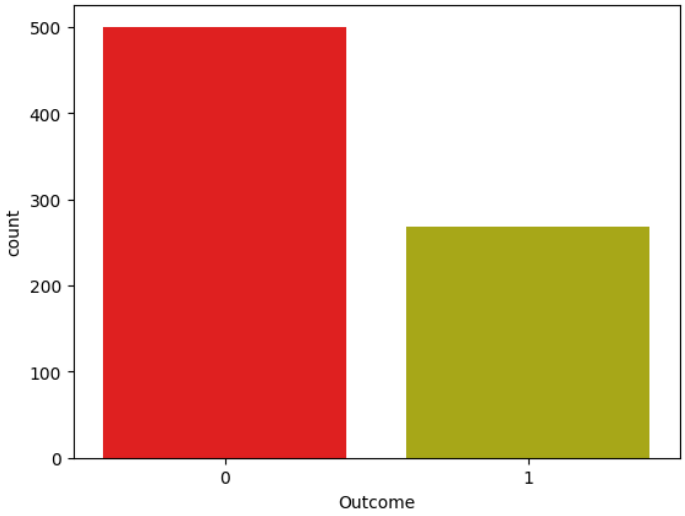

Figure 3: The outcome class.

Table 2 shows the accuracy of the proposed model. The best accuracy in the RF model and the lowest accuracy in the LR model.

Table 2: The results of the ML models.

	Accuracy	Precision	F1 Score
LR	76%	71%	61%
SVM	77%	72%	62%
RF	79%	75%	67%

Figure 4 shows the heatmap between the dataset. Finally, Figure 5 shows the pair plot of the dataset.

DOI: https://doi.org/10.54216/IJNS.210207
Received: February 19, 2023 Revised: April 27, 2023 Accepted: May 28, 2023

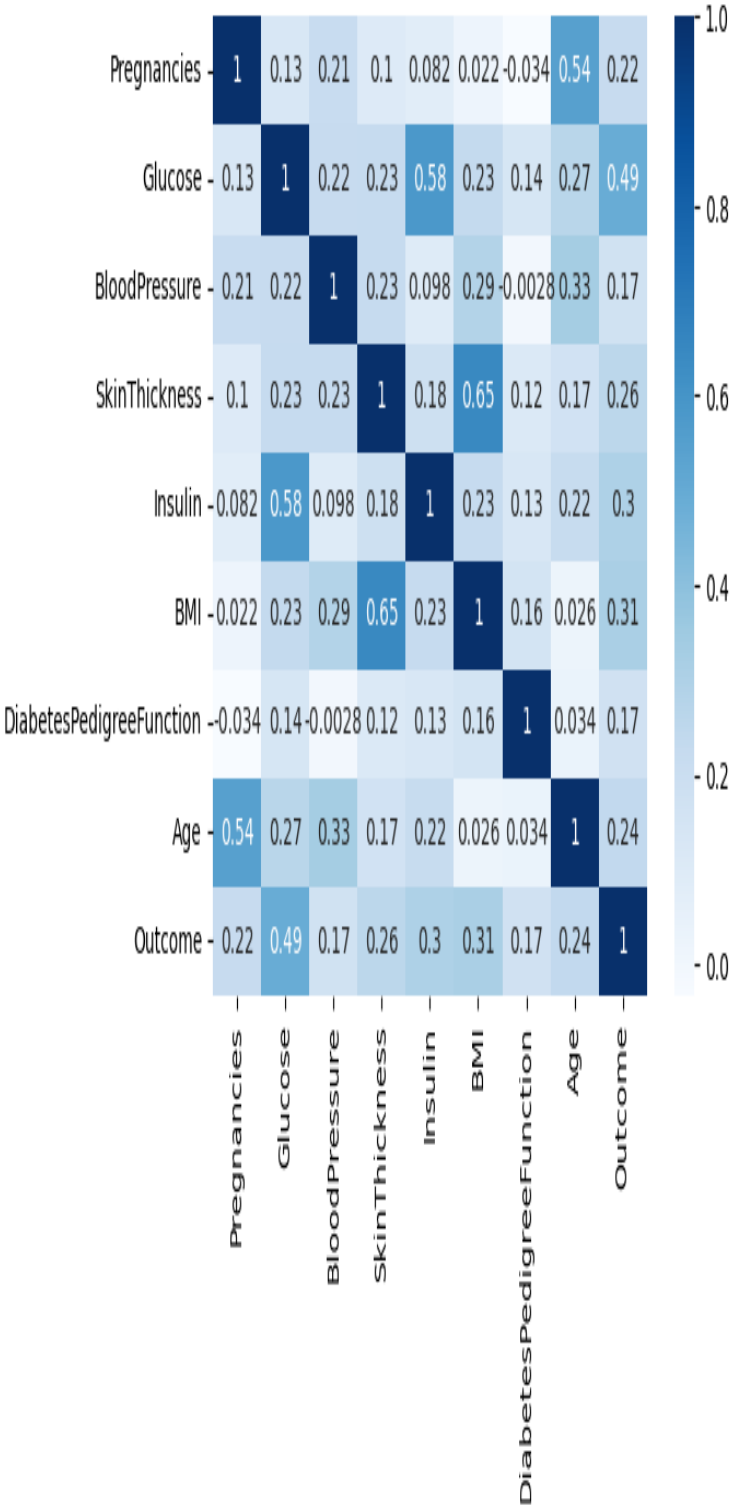

Figure 4: The heat map.

DOI: https://doi.org/10.54216/IJNS.210207
Received: February 19, 2023 Revised: April 27, 2023 Accepted: May 28, 2023

Figure 5: The pair plot of the dataset.

5. Conclusion

The dataset provides the patient data used in this study. In this research, we constructed an algorithm for diabetes diagnosis by employing a number of machine-learning techniques for categorization. The methods used include Logistic Regression, Random Forest Classifier, and Support Vector Machine. Additionally, the accuracy of the results is measured as a means of gauging the effectiveness of the machine learning method. The neutrosophic sets used in this paper with the AHP method to select bets feature as a feature selection. The lowest feature is removed and the highest rank is employed in the machine learning model. The dataset has nine features. The random forest has the highest rank in accuracy but the logistic regression has the lowest rank.

DOI: https://doi.org/10.54216/IJNS.210207
Received: February 19, 2023 Revised: April 27, 2023 Accepted: May 28, 2023

References

[1] M. K. Hasan, M. A. Alam, D. Das, E. Hossain, and M. Hasan, "Diabetes prediction using ensembling of different machine learning classifiers," *IEEE Access*, vol. 8, pp. 76516–76531, 2020.

[2] S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, and T. Saba, "Current techniques for diabetes prediction: review and case study," *Appl. Sci.*, vol. 9, no. 21, p. 4604, 2019.

[3] N. S. El_Jerjawi and S. S. Abu-Naser, "Diabetes prediction using artificial neural network," 2018.

[4] M. Komi, J. Li, Y. Zhai, and X. Zhang, "Application of data mining methods in diabetes prediction," in *2017 2nd international conference on image, vision and computing (ICIVC)*, IEEE, 2017, pp. 1006–1010.

[5] J. Li *et al.*, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, 2017.

[6] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, 2018.

[7] S. I. Ayon and M. M. Islam, "Diabetes prediction: a deep learning approach," *Int. J. Inf. Eng. Electron. Bus.*, vol. 12, no. 2, p. 21, 2019.

[8] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Comput. Sci.*, vol. 165, pp. 292–299, 2019.

[9] N. Nai-Arun and R. Moungmai, "Comparison of classifiers for the risk of diabetes prediction," *Procedia Comput. Sci.*, vol. 69, pp. 132–142, 2015.

[10] T. N. Joshi and P. P. M. Chawan, "Diabetes prediction using machine learning techniques," *Ijera*, vol. 8, no. 1, pp. 9–13, 2018.

[11] N. Jayanthi, B. V. Babu, and N. S. Rao, "Survey on clinical prediction models for diabetes prediction," *J. Big Data*, vol. 4, pp. 1–15, 2017.

[12] B. Mahesh, "Machine learning algorithms-a review," *Int. J. Sci. Res. (IJSR).[Internet]*, vol. 9, pp. 381–386, 2020.

[13] G. Bonaccorso, *Machine learning algorithms*. Packt Publishing Ltd, 2017.

[14] M. Attya, M. K. El-Sayed, A. Sakr, and H. Ahmed, "An evaluation framework for selecting cloud service provider in neutrosophic environment and Modified Generative Adversarial Network," *IJCI. Int. J. Comput. Inf.*, vol. 10, no. 1, pp. 78–89, 2023.

[15] Shimaa Said , Mahmoud M. Ibrahim , Mahmoud M. Ismail, An Integrated Multi-Criteria Decision-Making Approach for Identification and Ranking Solar Drying Barriers under Single-Valued Triangular Neutrosophic Sets (SVTNSs), Neutrosophic and Information Fusion, Vol. 2 , No. 1 , (2023) : 35-49 (Doi : https://doi.org/10.54216/NIF.020103).

[16] F. Yiğit, "A Three-Stage Fuzzy Neutrosophic Sets-Based Methodology for Training Assignment," *Available SSRN 4341819*, 2023.

[17] A. Aliahmadi and H. Nozari, "Evaluation of security metrics in AIoT and blockchain-based supply chain by Neutrosophic decision-making method," in *Supply Chain Forum: An International Journal*, Taylor & Francis, 2023, pp. 31–42.

[18] S. Ray, "A quick review of machine learning algorithms," in *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, IEEE, 2019, pp. 35–39.

[19] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, Ieee, 2016, pp.

DOI: https://doi.org/10.54216/IJNS.210207
Received: February 19, 2023 Revised: April 27, 2023 Accepted: May 28, 2023

1310–1315.

[20] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *J. Intell. Learn. Syst. Appl.*, vol. 9, no. 01, p. 1, 2017.

[21] V. K. Ayyadevara, "Pro machine learning algorithms," *Apress Berkeley, CA, USA*, 2018.

[22] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[23] J. Miao and L. Niu, "A survey on feature selection," *Procedia Comput. Sci.*, vol. 91, pp. 919–926, 2016.

[24] V. Kumar and S. Minz, "Feature selection: a literature review," *SmartCR*, vol. 4, no. 3, pp. 211–229, 2014.

[25] B. Venkatesh and J. Anuradha, "A review of feature selection and its methods," *Cybern. Inf. Technol.*, vol. 19, no. 1, pp. 3–26, 2019.

[26] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine learning proceedings 1992*, Elsevier, 1992, pp. 249–256.

[27] Ahmed M. Ali, A Multi-Criteria Decision-Making Approach for Piston Material Selection under Single-Valued Trapezoidal Neutrosophic Sets, Neutrosophic and Information Fusion, Vol. 2 , No. 1 , (2023) : 23-43 (Doi : https://doi.org/10.54216/NIF.020102)

[28] F. A. Alzahrani, N. Ghorui, K. H. Gazi, B. C. Giri, A. Ghosh, and S. P. Mondal, "Optimal Site Selection for Women University Using Neutrosophic Multi-Criteria Decision Making Approach," *Buildings*, vol. 13, no. 1, p. 152, 2023.

[29] Mona Mohamed, Financial Risks Appraisal based on Dynamic Appraisal Framework, Neutrosophic and Information Fusion, Vol. 2 , No. 1 , (2023) : 50-58 (Doi : https://doi.org/10.54216/NIF.020104)

DOI: https://doi.org/10.54216/IJNS.210207
Received: February 19, 2023 Revised: April 27, 2023 Accepted: May 28, 2023