Colorectal Cancer Prediction Using Machine Learning and Neutrosophic MCDM Methodology: A Case Study

Juan Viteri Rodríguez[1,*], Julio Rea Martínez[2], Freddy F. Jumbo Salazar[3]

[1]Docente de la carrera de Medicina de la Universidad Regional Autónoma de los Andes (UNIANDES Ambato), Ecuador

[2]Docente de la carrera de Medicina de la Universidad Regional Autónoma de los Andes (UNIANDES Santo Domingo), Ecuador

[3]Docente de la carrera de Medicina de la Universidad Regional Autónoma de los Andes (UNIANDES Ambato), Ecuador

Email: ua.juanviteri@uniandes.edu.ec; ua.juliorm92@uniandes.edu.ec; ua.freddyjumbo@uniandes.edu.ec;

Abstract

The third most common disease worldwide, colorectal cancer (CRC) is responsible for around 10% of annual cancer diagnoses. The success of personalized treatment hinges on the ability to recognize biomarkers linked with CRC longevity and forecast the prognosis of CRC patients. The goal of this research is to provide a novel approach to doing multi-attribute colorectal cancer analysis by using machine learning algorithms with multi-criteria decision-making (MCDM) methods and neutrosophic set (NS). The NS is used to overcome the uncertainty in the dataset. This paper used the neutrosophic AHP method to get the weights of features in the used dataset. Then the machine learning algorithms are used to give analysis and prediction of colorectal cancer. The decision tree (DT) and support vector machine (SVM) is used to analyze and predict colorectal cancer. The dataset has nine features like age, gender, dukes stage, location, and Disease-free survival. This paper shows the analysis of the dataset and the correlation among the features.

Keywords: Machine Learning; AHP; MCDM; Neutrosophic Set; Colorectal Cancer.

1. Introduction

CRC has an elevated prevalence and fatality rate, making it the third deadliest cancer in the globe. Annual deaths attributable to CRC are estimated at 900,000. With almost 521,000 cases reported and an estimated 248,000 fatalities in 2018, CRC has risen to rank as the third most common cancer overall in China, behind only lung cancer and stomach cancer. Patients diagnosed at the initial stage or second stage of CRC in the United States had a 5-year relative survival rate of 80.1% to 88.1%, whereas those diagnosed at stage fourth had just a 12.6% chance of survival[1]–[4]. There was a significant risk of recurrence in patients who had undergone surgery to remove tumors[5], [6].

DOI: https://doi.org/10.54216/IJNS.210211
Received: February 19, 2023 Revised: May 01, 2023 Accepted: June 06, 2023

Between 30 and 50 percent of colorectal cancer patients will have a recurrence within 2 years after surgery[7]. Nevertheless, only a tiny fraction of patients live beyond 5 years following surgery. However, it is not yet known what variables influence postoperative survival. Therefore, reliable biomarkers are essential for determining the prognosis of CRC patients. This would aid physicians in determining patient survival rates and pinpointing those in danger of further treatment[8]–[10].

The use of machine learning in medical diagnostics is growing in significance as the field of machine learning matures. This paper used the decision tree and support vector machine in the nalysis of CRC[11], [12].

Since they provide means of dealing with complicated issues, "MCDM" strategies have recently gained the focus of officials. These methods aid in creating judgments that are more consistent by being more clear, rational, and efficient. For all "MCDM" techniques, the AHP is the most well-known and widely used method. AHP has several useful features that recommend it as a testing approach for this investigation[13]–[15].

The goal of this work, introduce the framework that combines the neutrosophic AHP method with the machine learning algorithm for the analysis and prediction of CRC. The Neutrosophic AHP is used to compute the weights of features in the dataset. The purpose of using the NS is to overcome the vague information in the dataset. The machine learning method was used after the neutrosophic AHP to analyze and predict the CRC. The decision tree and support vector machine are used in this paper. The idea of the decision tree is to divide the dataset as a tree and predict cancer. The idea of a support vector machine uses the kernel to predict the class.

2. Machine Learning and Neutrosophic Sets

This section presented the neutrosophic AHP and machine learning methods like decision tree (DT) and support vector machine (SVM). The framework of this study is shown in Figure 1.

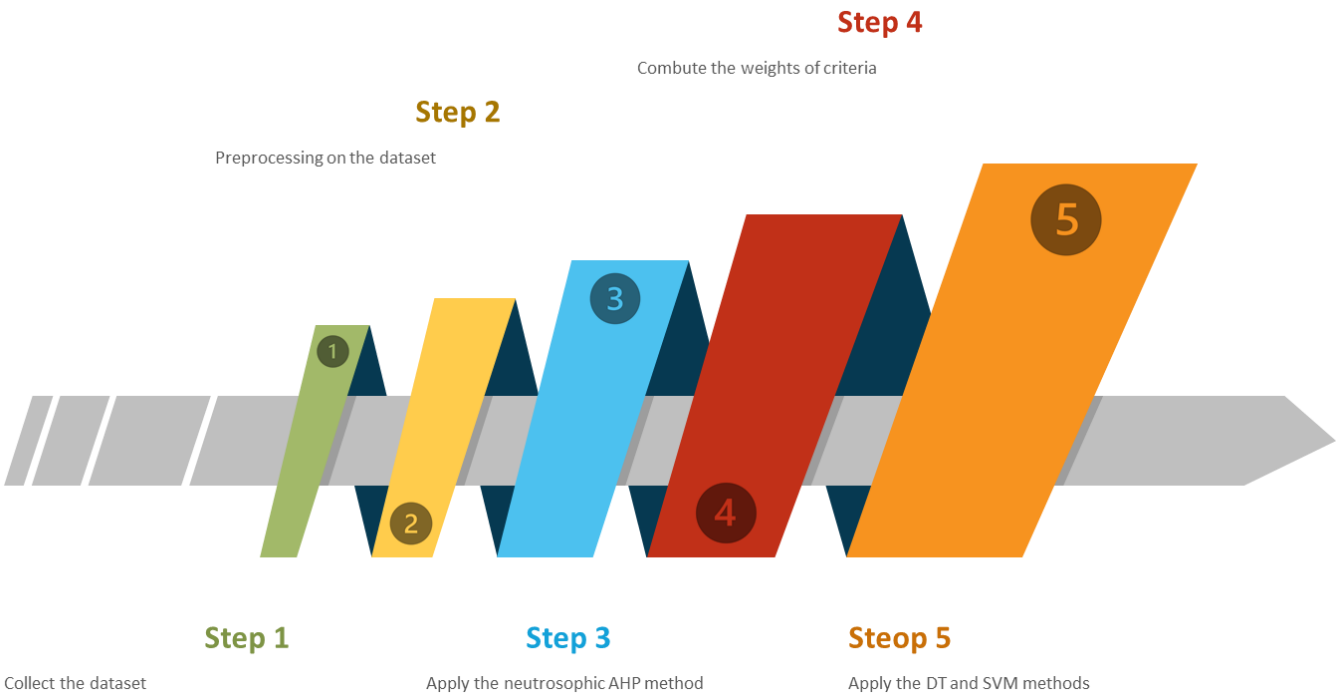

Figure 1: The steps of the suggested method.

DOI: https://doi.org/10.54216/IJNS.210211
Received: February 19, 2023 Revised: May 01, 2023 Accepted: June 06, 2023

2.1 Neutrosophic AHP

The NS may be used to the description of data that is imprecise, indeterminate, or inconsistent. Neutrosophy is the study of agnostic philosophy. When contrasted with fuzzy sets and intuitionistic fuzzy sets, it exemplifies a crucial distinction. One's level of connection or lack thereof determines the truthfulness or untruth of a statement. The indeterminacy criterion has nothing to do with the validity or falsehood of a statement[16]–[18]. The boundaries between truth, uncertainty, and falsehood are completely open. Each of these degrees has a range within which it might change [0,1].

$$B = \{< x, T_B(x), I_B(x), F_B(x) >: x \in X\} \quad (1)$$

The operations on the two single-valued neutrosophic sets (SVNSs) can be presented as:

$$x_1 = \left(T_{x_1}, I_{x_1}, F_{x_1}\right), and\ x_2 = \left(T_{x_2}, I_{x_2}, F_{x_2}\right)$$

The sum of two single-valued neutrosophic numbers (SVNNS) can be computed as:

$$x_1 + x_2 = \left(T_{x_1} + T_2 - T_{x_2}T_{x_1}, I_{x_1} + I_{x_2} - I_{x_1}I_{x_2}, F_{x_1} + F_{x_2} - F_{x_1}F_{x_2}\right) \quad (2)$$

The multiplication of two SVNNS can be computed as:

$$x_1x_2 = \left(T_{x_1}T_{x_2}, I_{x_1}I_{x_2}, F_{x_1}F_{x_2}\right) \quad (3)$$

The multiplication constant by the SVNNs can be computed as:

$$\sigma x_1 = \left(1 - \left(1 - T_{x_1}\right)^{\sigma}, 1 - \left(1 - I_{x_1}\right)^{\sigma}, 1 - \left(1 - F_{x_1}\right)^{\sigma}\right) \quad (4)$$

The power of SVNNS can be computed as:

$$x_1^{\sigma} = \left(T_{x_1}^{\sigma}, I_{x_1}^{\sigma}, F_{x_1}^{\sigma}\right) \quad (5)$$

The AHP is a common method for comparing many variables and options by giving weights based on a basic premise of arbitrary computation. When applied to decision-making, evaluation, and prioritizing concerns, AHP provides a powerful model that aids in handling and creating a model hierarchical. Basic measures may be used in a variety of surveys and tests to provide comparison measurements.

2.2 Machine Learning

By iteratively selecting the variable of explanation that yields the "best" divide at every node in the tree, Decision Trees (DT) build regression or classification algorithms that forecast the final result of the response parameter. They are straightforward and easy to grasp, but overfitting, which causes a reduction in generalization, is DT's primary drawback[19], [20].

Random Forest (RF) was presented as a solution to DT's overfitting issue. RF is an ensemble of numerous independent decision trees that work together to create a forecast. Because of the bootstrap aggregation method used to construct the trees, there is no correlation between them. This combined forecast is more precise than any of the separate trees. RF is not as intuitive to comprehend as DTs due to the fact that they are an ensemble of trees, but they do provide a straightforward method of gauging the relative significance of every explanatory factor.

The information gain of an attribute with respect to a category is quantified, as the name implies. On a foundation of information gain numbers, the node is divided, and the tree is built. In addition to dividing the node/attribute with the most information first, the DT method optimizes the information gain ratio. The gain in knowledge may be expressed quantitatively as:

DOI: https://doi.org/10.54216/IJNS.210211
Received: February 19, 2023 Revised: May 01, 2023 Accepted: June 06, 2023

$$IG = Entropy - [(WA) * Entropy(f)] \quad (6)$$

Where WA refers to the weighted average, IG refers to the information gain, and f refers to every feature.

$$Entropy = -P(Y)\log_2 P(Y) - P(N)\log_2 P(N) \quad (7)$$

Where Y refers to yes, and N refers to No

A decision tree's degree of impureness or purity may be quantified using the Gini index. The decision tree method gives more weight to traits with a low Gini index than those with a high index.

The following equation may be used to get the Gini index:

$$GI = 1 - \sum_j p_j^2 \quad (8)$$

Where GI refers to the GINI Index

Support Vector Machines are based on the assumption that linearly separable patterns may be best represented by a hyperplane. Patterns that cannot be disentangled using linear modifications of the original data may be mapped onto a new space employing a kernel function[21], [22]. Although kernel based-SVM also addresses overfitting issues, it is more complex than RF in terms of understanding and parameter tuning[23], [24].

3. Application

This section presented the application of the neutrosophic AHP method and machine learning methods. First, this study used the dataset in Kaggle in CRC. The dataset contains categorical text data and missing values. So, this study made some preprocessing steps for the dataset before working on it. First, this study drops all the missing values. Then encode the dataset. The gender is encoded into 0,1. Then the location also is encoded into four numbers. Also, the dukes stage is encoded. Table 1,2 shows the sample of the dataset before and after applying the preprocessing steps.

Table 1: Sample of CRC dataset before applying the preprocessing steps.

	Unnamed: 0	ID_REF	Age (in years)	Dukes Stage	Gender	Location	DFS (in months)	DFS event	Adj_Radio
0	0	GSM877126	62.0	A	Male	Left	108.0	0.0	1.0
1	1	GSM877127	77.0	B	Male	Left	40.0	1.0	1.0
2	2	GSM877128	66.0	C	Female	Left	49.0	0.0	1.0
3	3	GSM877129	72.0	D	Female	Left	45.0	0.0	1.0
4	4	GSM877130	75.0	C	Male	Left	40.0	0.0	0.0

DOI: https://doi.org/10.54216/IJNS.210211
Received: February 19, 2023 Revised: May 01, 2023 Accepted: June 06, 2023

Table 2: Sample of CRC dataset after applying the preprocessing steps.

	Unnamed: 0	ID_REF	Age (in years)	Dukes Stage	Gender	Location	DFS (in months)	DFS event	Adj_Radio
0	**0**	0	GSM877126	62.0	0	1	1	108.0	0.0
1	**1**	1	GSM877127	77.0	1	1	1	40.0	1.0
2	**2**	2	GSM877128	66.0	2	0	1	49.0	0.0
3	**3**	3	GSM877129	72.0	3	0	1	45.0	0.0
4	**4**	4	GSM877130	75.0	2	1	1	40.0	0.0

The dataset has nine features. Table 3 shows some descriptive statistics on the dataset. The descriptive statistics include mean, standard deviation, count, maximum, minimum, 25%,75%, and 50% of the dataset.

Table 3: Some descriptive statistics on the CRC dataset.

	Unnamed: 0	Age (in years)	Dukes Stage	Gender	Location	DFS (in months)	DFS event	Adj_Radio	Adj_Chem
count	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000	62.000000
mean	30.500000	61.112903	1.451613	0.774194	1.967742	41.774194	0.596774	0.532258	0.483871
std	18.041619	9.581582	1.081550	0.421526	0.904773	26.288076	0.494550	0.503032	0.503819
min	0.000000	28.000000	0.000000	0.000000	0.000000	4.000000	0.000000	0.000000	0.000000
25%	15.250000	56.250000	0.250000	1.000000	1.000000	19.500000	0.000000	0.000000	0.000000
50%	30.500000	62.000000	2.000000	1.000000	2.000000	38.000000	1.000000	1.000000	0.000000
75%	45.750000	67.000000	2.000000	1.000000	3.000000	57.500000	1.000000	1.000000	1.000000
max	61.000000	78.000000	3.000000	1.000000	3.000000	108.000000	1.000000	1.000000	1.000000

DOI: https://doi.org/10.54216/IJNS.210211
Received: February 19, 2023 Revised: May 01, 2023 Accepted: June 06, 2023

This study applies the steps of the neutrosophic AHP method to seven datasets. The pairwise comparison is built from the CRC dataset. Then normalize this matrix. Then compute the weights of features. The weights of features are organized as: F1: 0.073448033, F2: 0.167018815, F3:0.227387061, F4:0.159975853, F5:0.115705805, F6:0.108662843, F7:0.14780159

Then used the machine learning methods to apply some analysis methods. The histogram and box plot of some features like age and DFS are shown in Figures 2 and 3.

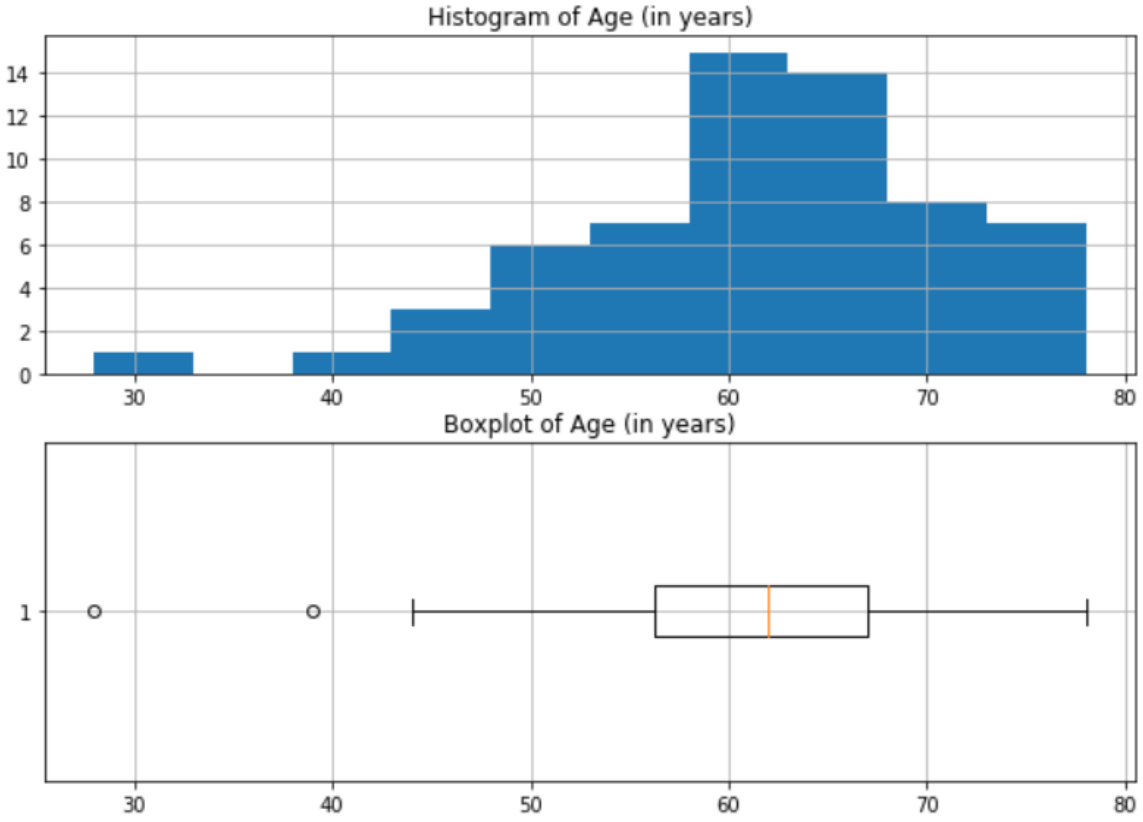

Figure 2: The histogram of age feature.

DOI: https://doi.org/10.54216/IJNS.210211
Received: February 19, 2023 Revised: May 01, 2023 Accepted: June 06, 2023

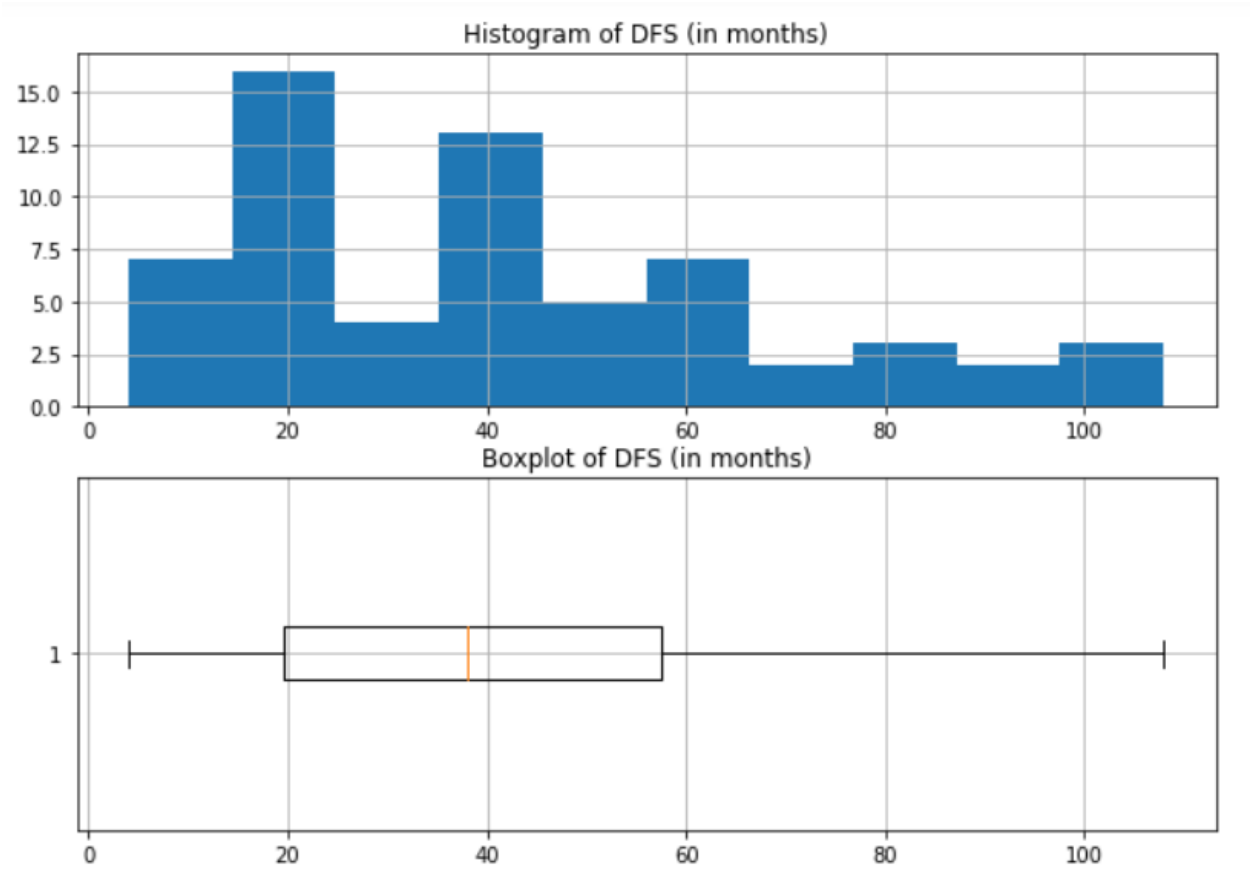

Figure 3: The histogram of DFS feature.

This study visualizes the Pearson and spearman correlation between age and DFS as shown in Figure 4.

DOI: https://doi.org/10.54216/IJNS.210211
Received: February 19, 2023 Revised: May 01, 2023 Accepted: June 06, 2023

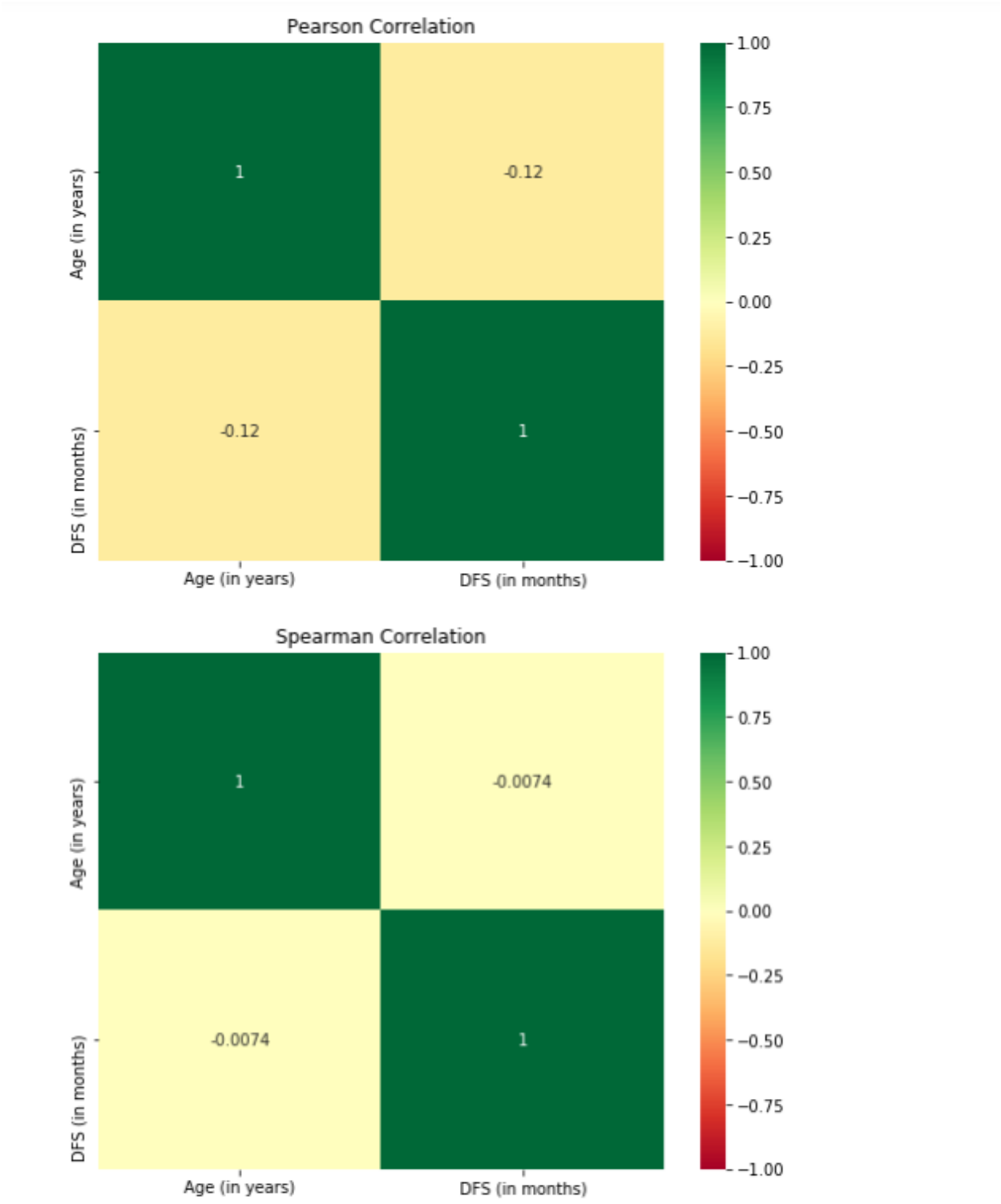

Figure 4: The Pearson and Spearman correlation between age and DFS.

DOI: https://doi.org/10.54216/IJNS.210211
Received: February 19, 2023 Revised: May 01, 2023 Accepted: June 06, 2023

Apply the decision tree and support vector machine of the dataset sets. This study extracts the confusion matrix from the decision tree and support vector machine as shown in Figures 5 and 6.

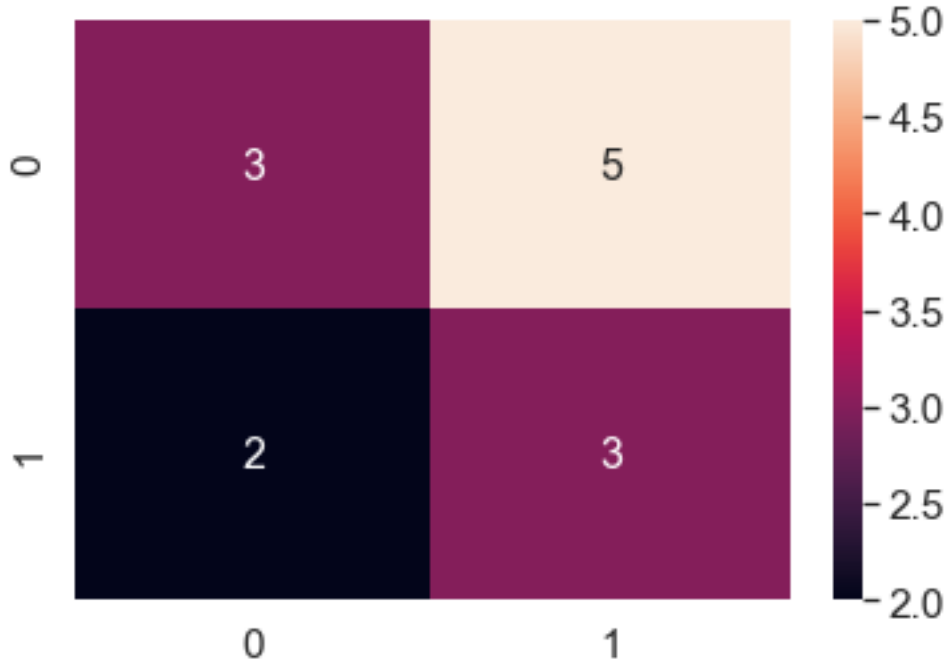

Figure 5: The confusing matrix by the decision tree.

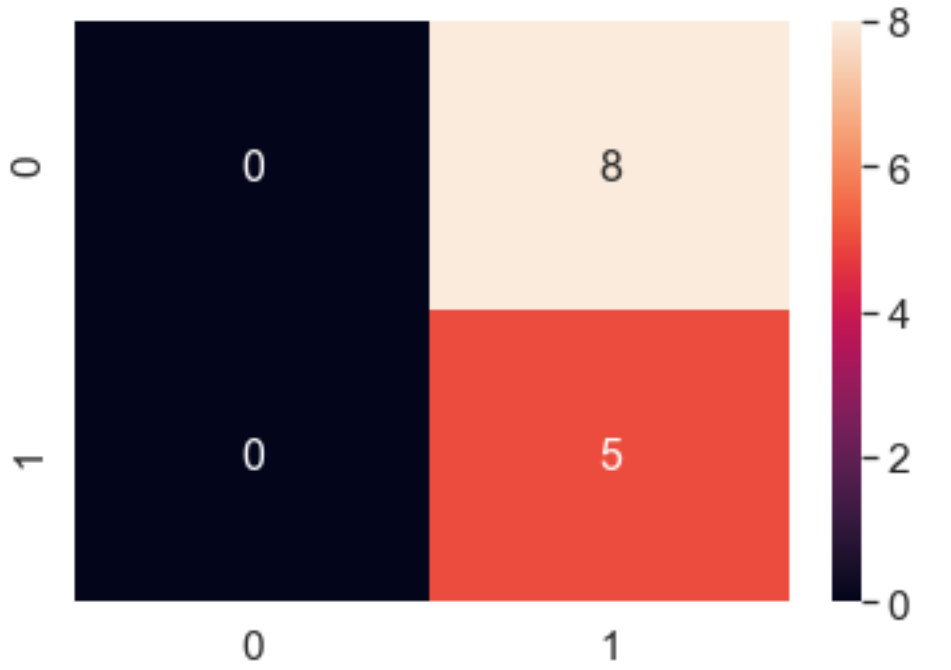

Figure 6: The confusion matrix by the support vector machine.

DOI: https://doi.org/10.54216/IJNS.210211
Received: February 19, 2023 Revised: May 01, 2023 Accepted: June 06, 2023

4. Conclusion

This study used the combination of neutrosophic sets and MCDM with machine learning methods. The neutrosophic sets are used to overcome the uncertainty in the dataset. The neutrosophic set is generalized from the fuzzy sets. We used the single-valued neutrosophic numbers to evaluate the features. The SVNSs are a kind type of neutrosophic set. SVNSs have three values include truth, indeterminacy, and falsity membership degrees. This study used the neutrosophic AHP to obtain the weights of the features. This study used the dataset from Kaggle. The dataset has nine features like age, gender, and DFS. Then the machine learning methods to analyze the dataset. The decision tree and support vector machine are the method of machine learning introduced in this paper. The idea of the decision tree is to split the dataset into a tree. The support vector machine used linear and nonlinear to predict the dataset. In the future study, other algorithms like logistic regression, and random forest used in this dataset. Increase the size of the dataset in the future. The other neutrosophic sets will be used to overcome the vague information.

References

[1] M. O. Kennion, S. Maitland, and M. R. Brady, "Machine learning as a new horizon for colorectal cancer risk prediction? A Systematic Review," *Heal. Sci. Rev.*, p. 100041, 2022.

[2] Y. Konishi *et al.*, "Development and evaluation of a colorectal cancer screening method using machine learning-based gut microbiota analysis," *Cancer Med.*, vol. 11, no. 16, pp. 3194–3206, 2022.

[3] C. Yang *et al.*, "Plasma lipid-based machine learning models provides a potential diagnostic tool for colorectal cancer patients," *Clin. Chim. Acta*, vol. 536, pp. 191–199, 2022.

[4] V. Bejan, E.-N. Dragoi, S. Curteanu, V. Scripcariu, and B. Filip, "The Prediction of Peritoneal Carcinomatosis in Patients with Colorectal Cancer Using Machine Learning," in *Healthcare*, MDPI, 2022, p. 1425.

[5] Z. Liu *et al.*, "Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer," *Nat. Commun.*, vol. 13, no. 1, p. 816, 2022.

[6] A. K. Waljee *et al.*, "Artificial intelligence and machine learning for early detection and diagnosis of colorectal cancer in sub-Saharan Africa," *Gut*, vol. 71, no. 7, pp. 1259–1265, 2022.

[7] V. Pereira and U. Bamel, "Charting the managerial and theoretical evolutionary path of AHP using thematic and systematic review: a decadal (2012–2021) study," *Ann. Oper. Res.*, pp. 1–17, 2022.

[8] Z. Liu *et al.*, "Integrative analysis from multi-center studies identities a consensus machine learning-derived lncRNA signature for stage II/III colorectal cancer," *EBioMedicine*, vol. 75, p. 103750, 2022.

[9] M. Yang *et al.*, "A multi-omics machine learning framework in predicting the survival of colorectal cancer patients," *Comput. Biol. Med.*, vol. 146, p. 105516, 2022.

[10] Z. Zhang, L. Huang, J. Li, and P. Wang, "Bioinformatics analysis reveals immune prognostic markers for overall survival of colorectal cancer patients: a novel machine learning survival predictive system," *BMC Bioinformatics*, vol. 23, no. 1, p. 124, 2022.

[11] Y. Fu, A. R. J. Downey, L. Yuan, T. Zhang, A. Pratt, and Y. Balogun, "Machine learning algorithms for defect detection in metal laser-based additive manufacturing: A review," *J. Manuf. Process.*, vol. 75, pp. 693–710, 2022.

[12] W. Xia, Y. Jiang, X. Chen, and R. Zhao, "Application of machine learning algorithms in municipal solid waste management: A mini review," *Waste Manag. Res.*, vol. 40, no. 6, pp. 609–624, 2022.

[13] 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)

DOI: https://doi.org/10.54216/IJNS.210211
Received: February 19, 2023 Revised: May 01, 2023 Accepted: June 06, 2023

[14] S. Panchal and A. K. Shrivastava, "Landslide hazard assessment using analytic hierarchy process (AHP): A case study of National Highway 5 in India," *Ain Shams Eng. J.*, vol. 13, no. 3, p. 101626, 2022.

[15] 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)

[16] Jesus Estupiñan Rcardo , Maikel Leyva Vázquez, Neutrosophic Multicriteria Methods for the Selection of Sustainable Alternative Materials in Concrete Design, American Journal of Business and Operations Research, Vol. 6 , No. 2 , (2022) : 28-38 (Doi : https://doi.org/10.54216/AJBOR.060203)

[17] Hadeer Mahmoud , Ahmed Abdelhafeez, Spherical Fuzzy Multi-Criteria Decision-Making Approach for Risk Assessment of Natech, Neutrosophic and Information Fusion, Vol. 2 , No. 1 , (2023) : 59-68 (Doi : https://doi.org/10.54216/NIF.020105)

[18] 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)

[19] Y. K. Wan, C. Hendra, P. N. Pratanwanich, and J. Göke, "Beyond sequencing: machine learning algorithms extract biology hidden in Nanopore signal data," *Trends Genet.*, vol. 38, no. 3, pp. 246–257, 2022.

[20] N. Nasir *et al.*, "Water quality classification using machine learning algorithms," *J. Water Process Eng.*, vol. 48, p. 102920, 2022.

[21] V. R. Allugunti, "Breast cancer detection based on thermographic images using machine learning and deep learning algorithms," *Int. J. Eng. Comput. Sci.*, vol. 4, no. 1, pp. 49–56, 2022.

[22] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using Machine learning algorithms," *Mater. Today Proc.*, vol. 80, pp. 3682–3685, 2023

[23] Gómez, Gustavo Adolfo Álvarez, Maikel Yelandi Leyva Vázquez, Jesús Estupiñán Ricardo. "Application of Neutrosophy to the Analysis of Open Government, its Implementation and Contribution to the Ecuadorian Judicial System." Neutrosophic Sets and Systems vol 52, pp.215-224., 2022.

[24] Ricardo, J. E., A. J. R. Fernández, M. Y. L. Vázquez. "Compensatory Fuzzy Logic with Single Valued Neutrosophic Numbers in the Analysis of University Strategic Management." International Journal of Neutrosophic Science, Vol. 18, pp. 151-159, 2022.

DOI: https://doi.org/10.54216/IJNS.210211
Received: February 19, 2023 Revised: May 01, 2023 Accepted: June 06, 2023