

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

ML algorithm for COVID-19 outcome prediction based on demographic features in North Macedonia

Maja Kukuseva Paneva^{1*}, Natasha Stojkovikj¹, Vasko Kokalanov¹, Vladimir Milićević²

¹ Goce Delcev University, Faculty of Computer Science, Stip, R. N. Macedonia

² University of Kragujevac, Faculty of Mechanical and Civil Engineering in Kraljevo, Serbia

ARTICLE INFO

* Correspondence: maja.kukuseva@ugd.edu.mk

DOI: <https://doi.org/10.46793/ET26.D10KP>

PAGES: D81 – D85

ABSTRACT

This study presents a machine learning approach for modeling and predicting COVID-19 patient outcomes in the Republic North Macedonia based on demographic characteristics. A decision tree-based classification model has been developed using algorithms implemented in WEKA environment. The model constructs the decision tree by selecting random subsets of features at each node and determines the optimal split points based on information gain.

The results indicate that age is the most significant predictor of patient outcome, serving as the root node and primary splitting attributes in the decision tree. Individuals below 60 years of age are classified as recovered, while higher age groups exhibit an increased probability of death. The gender attribute appears as a secondary factor, contributing to more refined predictions within specific age intervals. In addition to its predictive capability, the model offers a high level of interpretability, enabling clear visualization of decision rules and facilitating understanding of the relationship between demographic factors and disease outcomes. These findings highlight the importance of age-based risk stratification and demonstrate the potential of machine learning models as decision support tools in public health management.

KEYWORDS

Machine learning, COVID-19, Decision tree

1. INTRODUCTION

The outbreak of COVID-19 created significant challenges for healthcare systems not just in N. Macedonia, but worldwide due to its rapid transmission among the population. Since its emergence, COVID-19 has affected millions of individuals globally, [1], placing substantial pressure on the healthcare resources and emphasizing the need for efficient tools capable of supporting clinical decision-making patient management.

The integration of machine learning (ML) within healthcare system has fundamentally transformed clinical decision through analysis of heterogeneous datasets, pattern recognition and robust predictive modeling. Current algorithmic frameworks demonstrate high efficiency in diagnostic classification, patient risk stratification, and prognostic forecasting. Notably, during the COVID-19 pandemic, these predictive models served as vital epidemiological and clinical tools, enabling the early identification of high-risk individuals, [2], [3], [4]. Among different machine learning approaches, decision-tree based models, [5] are widely used due to its simplicity, interpretability and ability to represent decision making processes. These models operate through hierarchal framework consisting of root nodes, internal decision nodes, branch nodes and leaf nodes. Classification is executed by recursively partitioning the dataset according to selected attributes and decision rules. The main advantage of decision trees models is that they can handle numerical and categorical data simultaneously without requiring complex preprocessing procedures, and can identify the importance among attributes, allowing researchers to determine the most likely outcome and which individuals will be more affected.

2. METHODOLOGY

In this paper, the effect of demographic features on the outcome of COVID-19 is analyzed. Demographic features considered in this study are age and gender, while the outcome is classified as recovered or death. The data set is obtained

from Institute of Public Health of N. Macedonia and consists of 500 infected individuals. The main objective of the analysis is to determine whether age and gender have a significant impact on the progression and outcome of the disease. For this purpose, statistical methods were applied on the data for processing and analysis. The data set does not have missing value. The gender and outcome are binary values. Within the gender attribute value 1 indicates male, whereas value 0 indicates female individual. Within the outcome attribute, a value of 0 indicates recovered, whereas value of 1 indicates death. Table 1 presents the basic descriptive statistics for the parameters including minimum and maximum values, mean value and standard deviation for the demographic features. The minimum and maximum values provide insight into the age range of the infected individuals, while the mean value represents the average age of infected individuals included in this study. The standard deviation indicates the variability and distribution of data around the mean value. These statistical indicators clarify the population's characteristics, helping determine how demographic features influence the outcome of COVID-19.

Table 1: Descriptive statistics of the dataset

No	Attribute	Minimal value	Maximum value	Mean value	Standard deviation
1	Age	1	90	44.505	17.467
2	Gender	0	1	0.519	0.5
3	Outcome	0	1	0.93	0.26

Table 2 represents a representative sample consisting of data from five infected individuals within the data set.

Table 2: Sample of the dataset for 5 infected individuals

Feature					
Age	65	49	67	63	77
Gender	0	0	1	0	0
Outcome	1	0	0	0	1

This study employs different machine learning (ML) algorithms in order to analyze the impact of demographic features in the outcome of COVID-19. A decision tree-based classification model was developed using Random Tree and Random Forest algorithms implemented in WEKA® environment. The random tree classifier constructs a single tree by selecting a random subset of features at each node and determines an optimal split points based on information gained. On the other hand, the Random Forest constructs a collection of independent trees and then aggregates their predictions by voting among all trees.

In order to evaluate the performance of the developed classification model, several evaluation metrics were utilized, including Accuracy, Precision, Recall, F1 score and confusion matrix. The confusion matrix, as shown in Figure 1, is a tabular representation of comparison of predicted values against actual outcomes with four key metrics. The matrix categorizes the model prediction into correct classifications (True Positive (TP) and True Negative (TN)), and misclassifications (False Positive (FP) and False Negative (FN)).

		Actual Outcome	
Predicted Values		TP	FP
		FN	TN

Figure 1: Confusion matrix

Accuracy is one of the most used metrics used to evaluate the overall performances of a classification model. It presents the ration of correctly identified predicted values both True Positive and True Negatives to the total dataset population and is calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision is a critical metric that measures the ration of the model's true positive predictions to the total number of positive predictions made by the model as:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

In clinical predictive modeling, Recall (or Sensitivity) is a fundamental metric that measures the model's ability to identify all actual positive cases. It is calculated as ratio of correctly predicted positive values (TP) to the total number of actual positive outcomes in the total dataset population (TP and FN):

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

In this study, recall indicates the percentage of deceased individuals that the model successfully detected out of all individuals who actually died due to COVID-19. This means that a high Recall value ensures that the model is highly capable of capturing the critical outcomes, thereby minimizing False Negative (FN).

If the dataset outcome is imbalanced, relying only on overall accuracy can be misleading. That's why F1 score utilizes the calculation of the harmonic mean of Precision and Recall evaluating how effectively the model identifies the critical class (death). F1 score is calculated as:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (4)$$

3. RESULTS AND DISCUSION

The prediction of COVID-19 infected individual outcomes based on demographic features was performed using Random Tree and Random Forest classification algorithms implemented in WEKA® machine learning environment,[8]. The main objective of the analysis was to classify the infected individuals into two outcome categories: Recovered (R) and Death (D) based on the two demographic features, Age and Gender. Future research in [2] considers incorporating additional clinical and laboratory variables, including comorbidities and chronic diseases.

These algorithms were selected because they are widely used in medical data analysis due to their ability to model non-linear relationships while maintaining a high level of interpretability. Furthermore, tree-based classifiers require minimal data preprocessing and can efficiently handle both numerical and categorical variables.

The Random Tree model achieved an overall accuracy of approximately 91%, while the Random Forest model achieved an accuracy of approximately 90.6%. The relatively high accuracy values indicate that both algorithms were able correctly to identify infected individuals who recovered from COVID-19, as indicated by the high precision, recall and F1-score values for the Recovered class, as represented in Table 3.

Table 3: Comparison of performance Metrics of Random Tree and Random Forest Model for Predicting Recovered Individuals

Model	Accuracy (%)	Precision	Recall	F1-score
Random Tree	91.0	0.882	0.910	0.893
Random Forest	90.6	0.879	0.906	0.891

The obtained results show that both machine learning models perform very well in identifying recovered individuals. The Random Tree classifier achieved slightly higher values of Accuracy, Precision, Recall and F1-score compared to Random Forest, although the differences between the two algorithms are relatively small. Both models correctly classified more than 90% of recovered infected individuals, indicating that demographic features alone contain sufficient for recognizing the majority outcome class. The Precision values of 0.882 for Random Tree and 0.879 for Random Forest indicate that the most infected individuals predicted as recovered were correctly classified. Similarly, the high Recall values (0.910 and 0.906) demonstrate that the majority of recovered infected individuals were successfully identified by both models. The F1-score, which combines Precision and Recall into a single metric, further confirms the excellent predictive performance for the Recovered class, with values of 0.89 for both algorithms. Although Random Tree achieved slightly better values in most evaluation metrics, the Random Forest obtained a higher ROC Areal value (0.751) than Random Tree (0.647) indicating a better ability to distinguish between the two outcomes classes. These results suggest that Random Tree and Random Forest can effectively predict recovery outcomes.

Table 4 presents the evaluation results focused on the Death class, which is clinically most important category in this dataset. Unlike overall accuracy, which can be dominated by the majority class (in this case Recovered), these metrics provide a clearer insight into how well the models are able to identify critical and high-risk outcomes. In medical prediction tasks, the class Death is particularly important because false negatives, cases where the individual is incorrectly predicted as Recovered instead of Death, can have a serious consequence. Therefore, metrics such as Recall and F1-score for the class Death are more informative than overall accuracy alone.

Table 4: Comparison of performance Metrics of Random Tree and Random Forest Model for Predicting Deceased Individuals

Model	Precision	Recall	F1-score
Random Tree	0.250	0.108	0.151
Random Forest	0.222	0.108	0.145

Both Random Tree and Random Forest models show very similar performance for the class Death with extremely low Recall values (0.108) and low F1-score value, 0.151 for Random Tree and 0.145 for Random Forest. Random Forest performs slightly better in terms of overall discriminative ability as it achieves higher ROC Area, 0.751 compared to 0.647 for Random Tree.

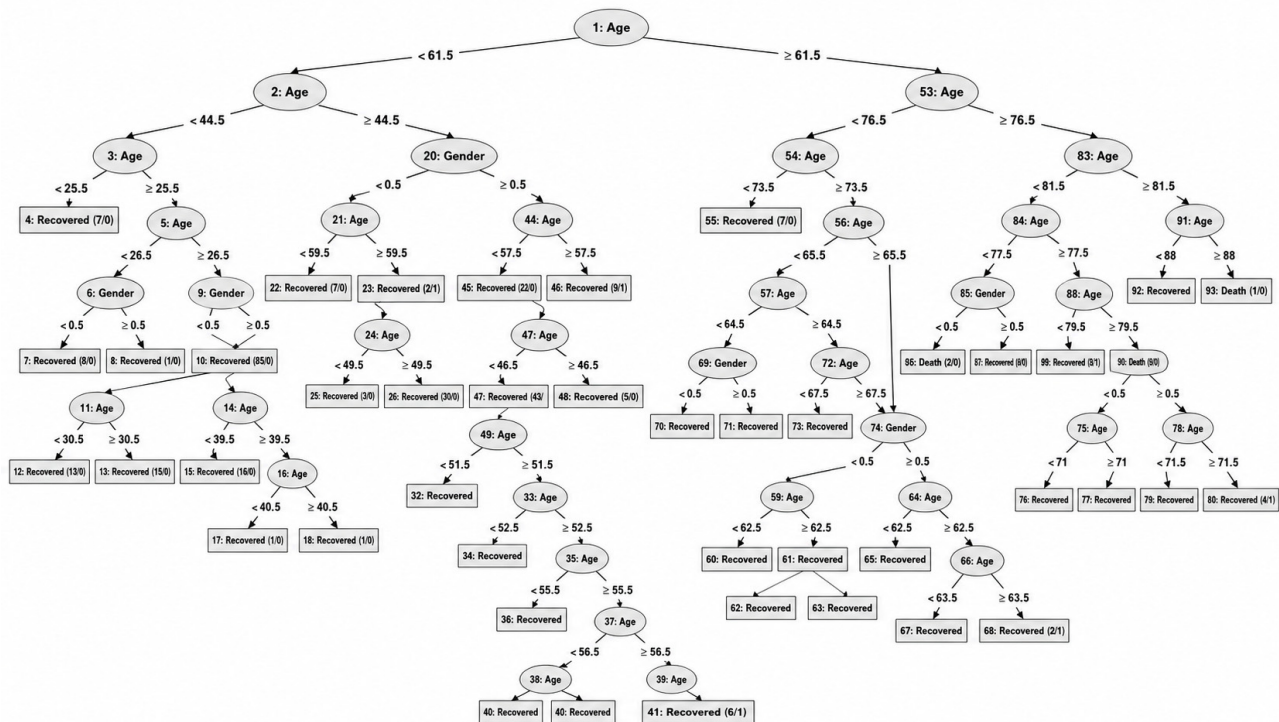


Figure 2. Decision tree generated using Random Tree

Figure 2 represents the decision tree generated using Random Tree algorithm for predicting COVID-19 outcome of infected individual based on demographic features Age and Gender. One of the main advantages of decision tree models is their interpretability, allowing researchers to clearly understand how different attributes contribute to the final prediction. Unlike black-box machine learning models, decision trees provide explicit decision rules that can be directly analyzed and interpreted. The generated tree reveals that Age is the most significant and dominant predictive attribute, as it appears as a root node and in the majority of subsequent decision nodes in the tree. This indicates that the age of infected individual has the strongest impact in determining whether the infected individual is classified as Recovered or Death. This finding is consistent with numerous epidemiological studies that have identified age as one of the strongest predictors of severe COVID-19 outcomes and mortality, [6]-[7]. The initial split of the dataset occurs at an age threshold of 61.5 years, dividing younger and older infected individuals into separate branches within the classification process. For individuals younger than 61.5 years, the model predominantly predicts the outcome Recovered. Sever additional age-based thresholds further refine the classification process. Despite the additional subdivisions, the majority of terminal nodes within this branch corresponds to recovery outcome of the infected individuals, suggesting that younger individuals generally exhibit a lower risk of mortality.

Gender appears as a secondary attribute in several branches of the tree, suggesting that although it contributes to the classification process, its influence is considerably lower compared to Age attribute. This observation indicates that gender contributes additional discriminatory information but possesses lower predictive importance compared to the age attribute. In the branches corresponding to older individuals, above 61.5 years, the tree becomes more complex, indicating a higher variability in outcomes among elderly population. Additional thresholds, such as 64.5, 67.5, 76.5, 79.5 and 81.5 years, are used to distinguish between different risk groups. Notably, the few terminal nodes classified as Death are located within these older age ranges, suggesting an increased mortality risk among elderly individuals within the population.

The decision tree contains a large number of leaf nodes classified as Recovered, while only a small number of leaves correspond to Death. This reflects the class distribution in the dataset, where recovered cases greatly outnumber death cases.

Overall, the decision tree analysis of Figure 2 confirms that age is dominant demographic factor affecting COVID-19 outcomes of infected individuals, while gender has a secondary demographic factor. The generated decision tree successfully captures the increased mortality rate associated with elderly infected individuals. Also, the model provides an interpretable representation of the decision-making process and highlights the increased risk associated with advanced age. Therefore, the random tree model not only achieves classification capability but also provides valuable insight into underlying factors influencing infected individual's prognosis.

4. CONCLUSION

This research investigates the potential of decision tree-based machine learning algorithms to predict COVID-19 outcome of infected individuals in R.N. Macedonia using only demographic features, namely age and gender. The obtained results demonstrate that both Random Tree and Random Forest algorithms achieved high overall classification accuracy of approximately 91%, indicating that demographic information alone provides valuable predictive information for identifying the majority outcome class.

The analysis revealed that age is the most significant predictor of COVID-19 outcome, while gender contributes as a secondary factor in specific decision branches. The generated decision tree clearly illustrates that individuals younger than approximately 60 years are predominantly classified as recovered, whereas the probability of death increases considerably among older age groups. This finding is consistent with previously reported epidemiological research identifying advanced age as one of the strongest risk factors for severe COVID-19 outcomes.

Overall, this research confirms that interpretable machine learning models such as Random Tree and Random Forest can serve as useful decision support tools for providing transparent classification rules and identifying the relative importance of demographic features. Furthermore, future research will focus on improving COVID-19 outcomes by incorporating additional clinical features including chronic disease and infected individuals' symptoms such as fever, shortness of breath, muscular pain, cough etc. The inclusion of these predictors is expected to increase the accuracy and reliability of the developed machine learning models.

REFERENCES

1. WHO COVID-19 dashboard, <https://data.who.int/dashboards/covid19/cases?n=o>
2. M.K. Paneva, C.M. Bande, N. Stojkovikj, D. Bikov, "Machine Learning Models for Prediction of COVID-19 Infection in North Macedonia, TEM Journal, Vol. 14(1), pp. 160-168, <https://www.doi.org/10.18421/TEM141-15> , (2025)
3. D. N. Jayavelu, H. Samaha, S.T. Wimalasena et al. "Machine learning models predict long COVID outcomes based on baseline clinical and immunologic factors". Commun Med 6 (1). <https://doi.org/10.1038/s43856-025-01230-w>, (2026).
4. Y. Cheng, R. Cheng, T. Xu, X. Tan, Y. Bai, "Machine Learning Techniques Applied to COVID-19 Prediction: A Systematic Literature Review", Bioengineering (Basel), Vol. 12(5):514. <https://doi.org/10.3390/bioengineering12050514>, (2025).
5. Z. Mohammadi-Pirouz, K. Hajian-Tilaki, M.S. Haddat-Zavareh, et al. "Development of decision tree classification algorithms in predicting mortality of COVID-19 patients", Int J Emerg Med, Vol 17, 126 <https://doi.org/10.1186/s12245-024-00681-7>, (2024)
6. E.J. Williamson, A.J. Walker, K. Bhaskaran, et al. "Factors associated with COVID-19-related death using OpenSAFELY". Nature 584, 430–436, <https://doi.org/10.1038/s41586-020-2521-4>, (2020).
7. Y. Zhou, Q. Yang, J. Chi et al. "Comorbidities and the risk of severe of fatal outcomes associated with coronavirus disease 2019: A systematic review and meta-analysis", International Journal of Infectious Diseases, Vol. 99, 47-56, <https://doi.org/10.1016/j.ijid.2020.07.029> , (2020).
8. <https://www.weka.io/>