Machine Learning Models for Prediction of COVID-19 Infection in North Macedonia

Maja Kukusheva Paneva [1], Cveta Martinovska Bande [2],
Natasha Stojkovikj [2], Dushan Bikov [2]

[1] *Faculty of Electrical Engineering, Goce Delcev University, Krste Misirkov, 10A, 2000 Stip, North Macedonia*
[2] *Faculty of Computer Science, Goce Delcev University, Krste Misirkov, 10A, 2000 Stip, North Macedonia*

***Abstract* –The COVID-19 pandemic, caused by the SARS-CoV-2 virus, has emerged as one of the most significant global crises of this century, with severe health and socio-economic impacts worldwide. Existing research has highlighted the critical role of comorbidities in influencing COVID-19 outcomes, but effective prediction models remain a challenge. This study investigates the potential of machine learning algorithms to predict the outcomes of COVID-19 based on patients' comorbidities. The algorithms K-Nearest Neighbors, Decision Tree, Logistic Regression, and Random Forest are applied to an epidemiological dataset comprising only positive COVID-19 cases, obtained from the Public Health Institute of North Macedonia. Additionally, two ensemble learning techniques, XGBoost and RUSBoost, are used to enhance prediction accuracy. The models achieved high accuracy of 90% across the various algorithms. These findings suggest that machine learning models can be an effective tool for predicting COVID-19 outcomes, especially when comorbidity data is available.**

***Keywords* – Machine learning, classification, ensemble methods, COVID-19 dataset**

DOI: 10.18421/TEM141-15
https://doi.org/10.18421/TEM141-15

Corresponding author: Natasha Stojkovikj,
Faculty of Electrical Engineering, Goce Delcev University, Krste Misirkov, 10A, 2000 Stip, North Macedonia.
Email: natasa.stojkovik@ugd.edu.mk

Received: 09 July 2024.
Revised: 23 October 2024.
Accepted: 24 December 2024.
Published: 27 February 2025.

1. Introduction

The initial outbreak of COVID-19 occurred in Wuhan, Hubei Province, China in late 2019. Recognized for its high contagion, COVID-19 attained pandemic status on March 12th, 2020, declared as such by the World Health Organization (WHO). This declaration came amidst a surge in confirmed cases reported across numerous countries and regions globally. There were about 125 600 confirmed cases across 118 countries. A pandemic denotes the widespread prevalence of an infectious disease, reaching nearly every corner of the world [1], [2].

The symptoms of COVID-19 are fever, dry cough, fatigue, with occasional gastrointestinal symptoms. These symptoms are more severe in older adults with underlying chronic conditions, and many patients also experience shortness of breath, which can resemble flu-like symptoms. The virus is primarily transmitted through direct contact with respiratory droplets from an infected person, particularly through sneezing and coughing [3].

Expert systems and other artificial intelligence techniques are crucial for diagnosing and containing the COVID-19 pandemic. Implementing these non-therapeutic approaches can alleviate significant pressure on healthcare systems by offering optimal diagnostic and predictive methods for managing 2019-nCoV effectively. The vaccination is used as a primary strategy to control the coronavirus (COVID-19) pandemic, also known as SARS-CoV-2.

However, there remains insufficient data on how different clinical and sociodemographic factors impact outcomes related to COVID-19. Despite extensive global research for mortality and morbidity rates and the effects of various sociodemographic and clinical characteristics on COVID-19, gaps in understanding persist [4].

In recent years, in order to predict the spread of infectious disease and due to potential to provide accurate and timely forecast machine learning (ML) models are utilized.

Infectious diseases pose significant public health challenges worldwide, affecting millions of individuals and placing immense strain on healthcare systems. Therefore, developing effective prediction models is crucial for early detection, prevention, and control of these diseases [5].

To forecast the spread of infection, various ML techniques have been utilized. These models leverage epidemiological data, demographic information, and environmental factors to generate predictions [6].

In [7] authors indicate that integrating multimodal data, such as gene expression, clinical characteristics, and comorbidities, significantly improves the accuracy of disease severity prediction. Machine learning models, particularly XGBoost, demonstrated exceptional precision, with 95% accuracy and a 0.99 AUC, in distinguishing severity groups.

The paper [8] indicates that lung-related comorbidities, such as chronic obstructive pulmonary disease and asthma, along with vascular conditions like cardiovascular and cerebrovascular diseases, are most significantly associated with COVID-19 severity.

The paper [9] examines the applications of machine learning (ML) in the fight against COVID-19, emphasizing the importance of various algorithms such as LR, RF, KNN, and SVM, as well as the need for further research using advanced methods like boosting and stacking.

In [10], the prognostic capabilities of various machine learning algorithms, including J48 decision tree, k-NN, MLP, SVM, XGBoost, NB, RF, and LR, were evaluated for predicting COVID-19 mortality using a comprehensive set of features, such as chest computed tomography severity score data, demographics, risk factors, clinical manifestations, and laboratory findings.

The results demonstrated that ML-based predictive models, leveraging routine data, can provide timely and accurate risk stratification for COVID-19 patients. Notably, the RF model, with its extensive set of predictors, effectively identified high-risk patients at the time of admission, potentially improving survival outcomes.

The paper [11] shows that the K-Nearest Neighbors (KNN) classifier with two neighbors achieves the best results in predicting the presence of COVID-19, with an accuracy of 98.37% and a low absolute error, providing valuable support for clinical practice in identifying patients with COVID-19 symptoms.

The authors in [12] conclude that age and comorbidities, such as hypertension, diabetes, and cardiovascular diseases, are significant risk factors for severe cases of COVID-19. These findings underscore the importance of targeted preventive measures, including vaccination programs, to protect vulnerable populations.

The study further emphasizes the shared mechanisms between chronic diseases and infectious conditions, such as inflammation and weakened immune responses, which increase susceptibility to severe outcomes. These insights call for tailored public health strategies to mitigate risks among high-risk groups and reduce the burden of severe COVID-19 cases.

Paper [13] presents the results of the research, which combines machine learning methods using the k-means algorithm for clustering, followed by prediction and mapping of distribution patterns with KNN and ID3, show a 90% accuracy rate in applying to the spread of COVID-19 in Indonesia.

Artificial intelligence techniques, particularly convolutional neural networks and transfer learning, have demonstrated significant potential in diagnosing and monitoring COVID-19 through chest imaging and laboratory data. Systematic reviews reveal the importance of AI in automating medical image analysis, emphasizing its role in addressing the challenges of manual annotation during the pandemic [14].

The paper [15] reviews 1,196 prediction models for COVID-19 from 2020, analyzed using a systematic search across databases like Google Scholar, Web of Science, and Scopus.

This paper explores various algorithms for data classification, such as KNN, Decision Tree, Logistic Regression, and Random Forest. Furthermore, it incorporates two ensemble learning methods: XGBoost and RUSBoost.

By using these advanced techniques and epidemiological data, this paper aims to develop accurate and reliable models for forecasting the spread of COVID-19.

The analysis reveals that the Decision Tree algorithm exhibits slightly superior accuracy compared to other methods, achieving 92%. Similarly, Logistic Regression and the KNN algorithm also demonstrate high accuracy at 91% and 90%, respectively. In contrast, the Random Forest algorithm achieves a lower accuracy of 73%. Furthermore, employing ensemble learning techniques such as XGBoost and RUSBoost yields accuracies of 90% and 87%, respectively.

2. Data

The dataset contains 501 patients in the period from 2020 to 2021. The dataset encompasses 9 features, including 2 demographic variables (age and gender) and clinical indicators: pneumonia, cardiovascular diseases (CVDs), diabetes, chronic kidney disease (CKDs), neuromuscular, liver disease, cancer and the outcome (death – D or recovered - R).

The algorithms are trained and evaluated using an epidemiological dataset comprising only positive COVID-19 cases in North Macedonia, provided by the Public Health Institute of North Macedonia [16].

In this study, the value 1 is assigned to males, while 0 represents females. For comorbidities, a value of 1 indicates the presence of a condition, and 0 indicates its absence. The dataset contains no missing values.

Detailed information about COVID-19 patients, including demographic, clinical, and epidemiological characteristics, is provided in Table 1.

Table 1. Dataset features

No.	Feature	Description
1	Age	1. 0 - 15, 2. 16-30, 3. 31-45, 4. 45-60. 5>60.
2	Sex	0, 1
3	Pneumonia	0, 1
4	CVDs	0, 1
5	Diabetes	0, 1
6	CKDs	0, 1
7	Neuromuscular	0, 1
8	Liver Disease	0, 1
9	Cancer	0, 1
10	Outcome	1-Recover, 0-Death

Table 2. Descriptive statistics

No	Variable	mean	std	min	max
1	Age	44.5	17.47	1	90
2	Gender	0.48	0.5	0	1
3	CVDs	0.2	0.4		
4	Diabetes	0.06	0.23	0	1
5	Pneumonia	0.063	0.244	0	1
6	Liver Disease	0.003	0.063	0	1
7	CKDs	0.015	0.125	0	1
8	Neuromus cular	0.007	0.089	0	1
9	Cancer	0.003	0.063	0	1
10	Outcome	0.93	0.26	0	1

The descriptive statistics of the data set given in Table 2 include the minimum, maximum and mean values, and standard deviation for each feature. In Table 3 a sample of the data set is represented.

Table 3. Sample of the dataset

Variable	0	1	2	3	4
Age	49	65	67	63	77
Gender	0	0	1	0	0
CVDs	1	1	0	0	0
Diabetes	1	0	0	0	1
Pneumonia	0	0	0	0	0
Liver Disease	0	0	0	0	0
CKDs	0	0	0	0	0
Neuromuscu lar	0	0	0	0	0
Cancer	0	0	0	0	0
Outcome	R	D	R	R	D

3. Methodology

This study employs several machine learning algorithms to analyze an epidemiological dataset, including K-Nearest Neighbors (KNN), Decision Tree, Logistic Regression, and Random Forest. Additionally, advanced ensemble learning techniques, such as XGBoost and RUSBoost, are incorporated to enhance prediction accuracy.

The K-Nearest Neighbours (KNN) is a straightforward and intuitive algorithm commonly utilized for both classification and regression tasks. It is nonparametric and instance-based, implying that it does not assume specific distribution for the data and relies on the entire dataset for making predictions [17].

The Decision Tree machine learning algorithm is used to segment learning tasks by recursively splitting the dataset into subsets until each partition becomes homogeneous that contains only a single class. Specific characteristics of the dataset are used for classification [18].

Logistic Regression is a supervised machine learning algorithm used for binary classification to estimate the likelihood of a binary outcome (such as: yes/no, 1/0 or true/false) given input data. By applying a logistic (sigmoid) function, it models how independent variables relate to a categorical dependent variable. This method is favoured in machine learning for its straightforward approach and clear interpretation. Logistic regression can be adapted for multi-class classification by employing methods such as one-versus-rest (OvR) or multinomial logistic regression, allowing it to predict multiple classes rather than just two [19].

Random Forest is a versatile machine learning algorithm used for classification and regression tasks. It builds an ensemble of decision trees and combines the output from these decision trees to produce more accurate predictions. For classification task, the final prediction is determined based on majority vote for a class label, while for regression task, the average of the prediction from all trees is the outcome. This method is popular for its ability to handle complex prediction tasks and to work effectively with large and diverse datasets [20].

XGBoost, short for Extreme Gradient Boosting, stands out as a robust and efficient machine learning algorithm primarily utilized for supervised tasks such as classification and regression. It falls under the gradient boosting family, renowned for its capability to manage intricate data relationships, estimate feature importance, and employ regularization techniques to prevent overfitting. By building an ensemble of weak prediction models, typically decision trees, XGBoost optimizes these models using gradient descent to minimize prediction. Its widespread adoption in machine learning competitions and practical applications is attributed to its reliability and ability to deliver precise predictions [21].

RUSBoost, or Random Under-Sampling Boosting, is a machine learning algorithm designed to tackle class imbalance in datasets, especially in binary classification tasks. It utilizes ensemble learning and sampling techniques to boost classifier performance.

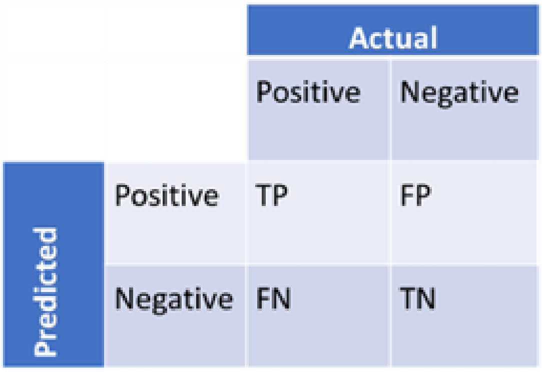
In order to balance the class distribution, RUSBoost starts randomly under-sampling to reduce the size of the majority class. It then trains a base classifier on each balanced subset of data. In subsequent iterations, it emphasizes misclassified instances from the minority class by assigning them higher weights. This iterative approach focuses on enhancing classification accuracy for the underrepresented class. The process continues until either the required number of classifiers is generated or a predefined stopping condition is met.

RUSBoost proves effective in scenarios where class imbalance is problematic, aiming to enhance predictive accuracy by directly addressing skewed class distributions during training [22].

4. Evaluation Metrics

Standard metrics, including accuracy, precision, recall, F1 score and cross-validation using ROC AUC scoring are used to assess model performance.

The confusion matrix (Figure 1) displays predicted and actual values in a structured format. The positive cases that are correctly predicted as positive are denoted as TP, while the positive cases incorrectly predicted as negative are denoted as FN. The negative cases incorrectly predicted as positive are denoted as FP, while the negative cases correctly predicted as negative are denoted as TN.



		Actual	
		Positive	Negative
Predicted	Positive	TP	FP
	Negative	FN	TN

Figure 1. Confusion matrix

Accuracy measures the ratio of correctly predicted instances to the total number of instances in the dataset.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}.$$

Precision, also known as positive predictive value, represents the ratio of true positive predictions to the total number of positive predictions made by the model.

$$Precision = \frac{TP}{TP + FP}.$$

Recall, also referred to as sensitivity or true positive rate, measures the model's capability to correctly identify all relevant instances, answering how many actual positives were predicted correctly by the model.

$$Recall = \frac{TP}{TP + FN}.$$

The F1 score, which is the harmonic mean of precision and recall, offers a balanced assessment of a model's performance:

$$F1 = \frac{2 \cdot P \cdot R}{P + R}.$$

When dealing with imbalanced datasets, accuracy may not adequately reflect performance on minority classes.

Instead, ROC AUC (Figure 2) evaluates how well effectively the model balances sensitivity (true positive rate) and specificity (true negative rate) across various thresholds, with higher values indicating better performance.

$$True\,Positive\,Rate = \frac{TP}{TP + FN},$$

$$False\,Positive\,Rate = \frac{FP}{FP + TN}$$

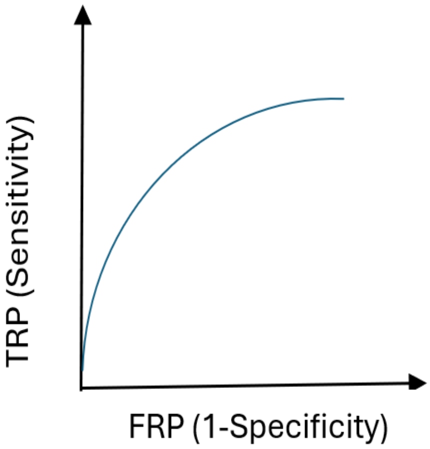


Figure 2. ROC AUC curve

5. Result

The correlations between different features of the epidemiological dataset are represented in Figure 3. Age, cardiovascular diseases, and diabetes are significant factors influencing the disease's adverse outcomes. The data also demonstrates associations between age and cardiovascular disease, as well as between cardiovascular disease and diabetes. Similarly, there are correlations observed between age and pneumonia, and between cardiovascular disease and pneumonia.

Table 4 shows the correlations coefficients between dependent variables and outcome of the disease as an independent variable.

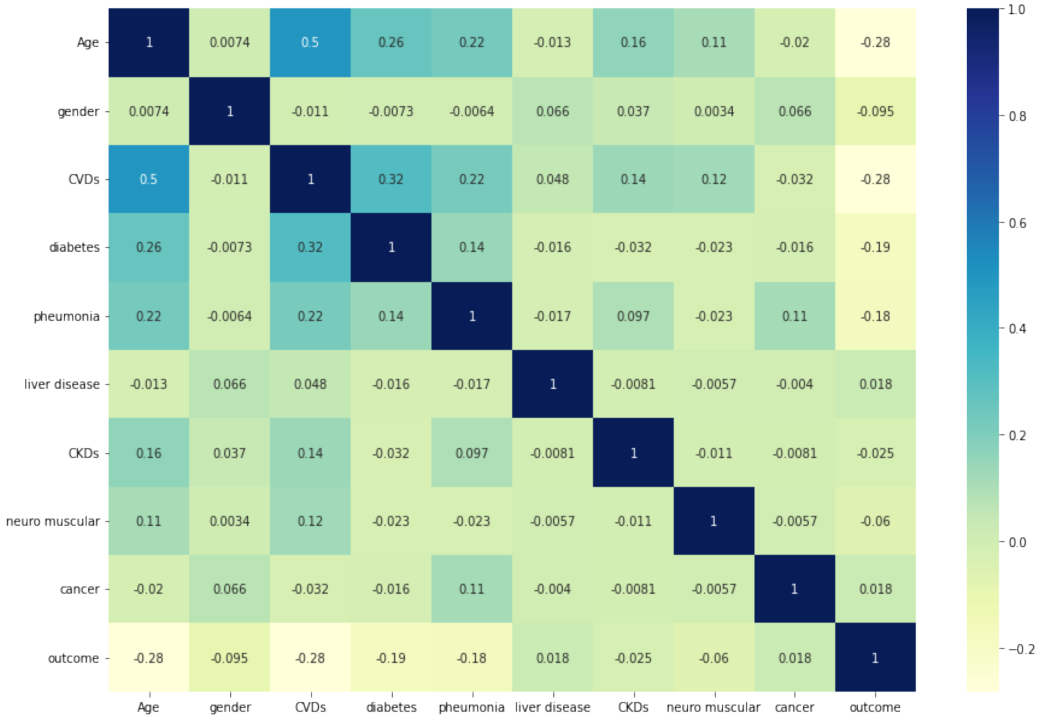


Figure 3. Correlation heatmap for epidemiological dataset

Table 4. Correlation coefficients between features and outcome of the disease

Dependent variable	Correlation coefficient
age	-0.28
gender	-0.095
CVDs	-0.28
diabetes	-0.19
pneumonia	-0.18
liver disease	0.018
CKDs	-0.025
neuro muscular	-0.06
cancer	0.018

Age histograms for the epidemiological dataset for patients that recovered and died from COVID-19 are represented in Figure 4.

Figure 4 shows that COVID-19 mortality was more prevalent among the older population.

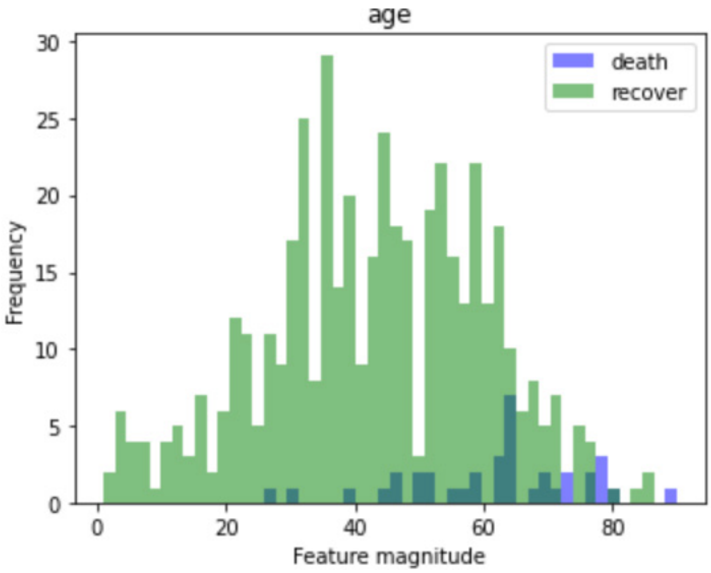


Figure 4. Patient's age distribution who recovered or died from COVID- 19

The evaluation of the used classifiers KNN, Decision Trees, Logistic Regression, Random Forest and ensemble methods XGBoost and RUSBoost are given in Table 5.

Based on the results obtained, the recall and precision of classifiers for the minority class (patients with a deceased outcome) were found to be unsatisfactory. The imbalance in the number of examples between the two classes was addressed using specific techniques:

- RUSBoost was used, which integrated random under sampling of the majority class with boosting techniques, (such as AdaBoost). This technique was specifically designed to handle class imbalances by reducing the majority class in each boosting iteration, thereby emphasizing examples from the minority class.
- Furthermore, the SMOTE algorithm was used for the Random Forest classifier and XGBoost. SMOTE creates synthetic samples were used for the minority class to ensure a balanced representation with the majority class.

Table 5. Evaluation of the classifiers

	precision		recall		F1	
classifier	recover	dead	recover	dead	recover	dead
KNN	0.92	0.33	0.98	0.11	0.95	0.17
DTs	0.95	0.57	0.97	0.44	0.96	0.5
LR	0.92	0.5	0.99	0.11	0.95	0.18
RF	0.92	0.13	0.77	0.33	0.84	0.18
XGBoost	0.94	0.17	0.79	0.44	0.86	0.25
RUSBoost	0.95	0.29	0.87	0.56	0.90	0.39

Using XGBoost, initially an accuracy of 0.90 was achieved. However, the recall for the minority class was zero. Implementing the SMOTE method improved the recall for the minority class, but this enhancement came at the cost of lower accuracy overall (Table 6).

These methods were implemented to mitigate the challenges posed by class imbalance in the dataset, aiming to improve the performance of the classifiers particularly for the underrepresented class of patients with deceased outcomes.

For the KNN classifier, by analyzing the accuracy change across different *k* values, the optimal value of 6 for *k* was determined.

Regarding the Decision Tree classifier, three models were developed using different splitting criteria (Gini, entropy, and log loss), to assess the quality of each split, resulting in comparable outcomes.

In the case of the Random Forest classifier, an accuracy of 0.91 was achieved initially, but encountered a challenge with zero recall for the minority class. When employing the SMOTE method to address class imbalance, although the accuracy decreased (0.73), the issue with minority class recall persisted.

Table 6. Accuracy of the classifiers

classifier	accuracy
KNN	0.90
DTs	0.92
LR	0.91
RF	0.73
XGBoost	0.90
RUSBoost	0.87

Using Logistic Regression, initially an accuracy of 0.91 was achieved using the 'newton-cg' solver.

However, the recall for the minority class was zero under these conditions. By introducing polynomial features with a degree of 2, the recall for the minority class while maintaining the same level of accuracy was improved.

Figure 5 displays the decision tree model constructed from the epidemiological database. The decision tree was obtained using WEKA software [23] and J48 class for generating C4.5 decision tree class. while maintaining the same level of accuracy was improved.

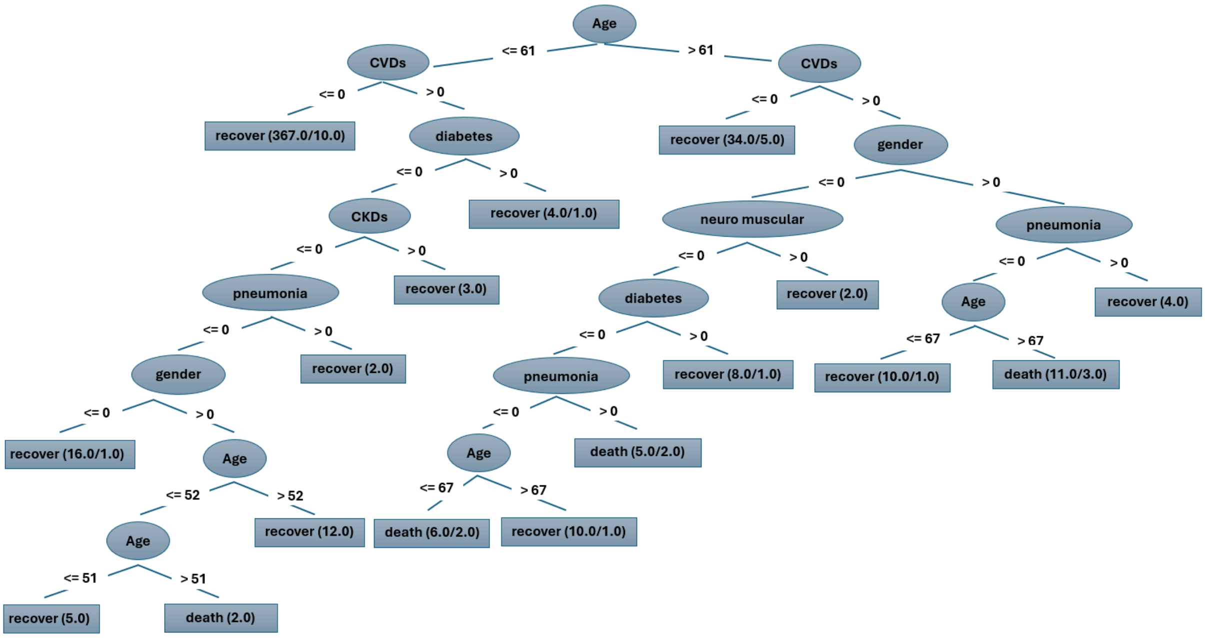


Figure 5. Decision tree for the epidemiology database

6. Conclusion

This study explored the application of multiple machine learning (ML) algorithms, including K-Nearest Neighbors (KNN), Decision Trees, Logistic Regression, Random Forest, and ensemble methods such as XGBoost and RUSBoost, for predicting COVID-19 outcomes in North Macedonia.

By leveraging epidemiological data with demographic and clinical features, the study achieved promising results, particularly in identifying critical risk factors like age and comorbidities (e.g., cardiovascular diseases, diabetes, and pneumonia). Techniques like SMOTE and RUSBoost were utilized to address class imbalances, enhancing the prediction of outcomes for minority classes (e.g., deceased patients).

The findings demonstrate the effectiveness of ML models in predicting disease outcomes with accuracies reaching up to 92%. While the Decision Tree model performed slightly better, KNN, Logistic Regression, and XGBoost also delivered high accuracy. However, challenges with recall for the minority class persist, highlighting the need for further optimization. The study underscores the potential of ML to support healthcare systems by providing timely and accurate risk stratification, enabling more effective allocation of medical resources and tailored interventions.

To improve the performance of the current model in predicting COVID-19 outcomes, future research should focus on several key aspects:

-Developing models that analyze temporal trends to predict disease progression dynamically, enabling proactive medical interventions and personalized healthcare management.

-Incorporating additional clinical and environmental variables, such as vaccination status, genetic predispositions, and socio-economic factors, to improve predictive accuracy.

-Integrating advanced machine learning methods, such as ensemble deep learning models or hybrid approaches that combine rule-based and data-driven methods, to enhance both interpretability and performance of the model.

Acknowledgements

The Authors express their gratitude for the support by Goce Delcev University Stip, North Macedonia through an individual scientific fund.

References:

[1]. Lazarova, L. K. (2022). Mathematical model for predictions of COVID-19 dynamics. *International Journal of Applied Mathematics*, *35*(1), 119-133. Doi: 10.12732/ijam.v35i1.9

[2]. Sugiyanto, S., & Abrori, M. (2020). A mathematical model of the Covid-19 Cases in Indonesia (under and without lockdown enforcement). *Biology, Medicine, & Natural Product Chemistry*, *9*(1), 15-19.

[3]. Doi: 10.14421/biomedich.2020.91.15-19
Lazarova, L. K., Stojanova, A., Stojkovikj, N., Miteva, M., & Ljubenovska, M. (2022). Analysis and prediction of the spread of COVID-19 in North Macedonia. *Asian-European Journal of Mathematics*, *15*(10), 2250237. Doi: 10.1142/S1793557122502370

[4]. Boateng, A., et al. (2023). Analysis of COVID-19 cases and comorbidities using machine learning algorithms: A case study of the Limpopo Province, South Africa. *Scientific African*, 21, e01840. Doi: 10.1016/j.sciaf.2023.e01840

[5]. Yaacob, W. F. W., et al. (2021). Machine learning models for COVID-19 confirmed cases prediction: A meta-analysis approach. *Journal of Physics: Conference Series*, 2084(1). Doi:10.1088/1742-6596/2084/1/012013

[6]. Muhammad L. J. et al. (2021). Supervised Machine Learning Models for Prediction of COVID-19 Infection Using Epidemiology Dataset. *SN Computer Science*, *2*(1). 1–13. Doi: 10.1007/s42979-020-00394-7

[7]. Sethi, S., et al. (2024). A Machine Learning Model for the Prediction of COVID-19 Severity Using RNA-Seq, Clinical, and Co-Morbidity Data. *Diagnostics*, *14*(12), 1284. Doi: 10.3390/diagnostics14121284

[8]. Aktar, S., Talukder, A., Ahamad, M. M., Kamal, A. H. M., Khan, J. R., Protikuzzaman, M., ... & Moni, M. A. (2021). Machine learning approaches to identify patient comorbidities and symptoms that increased risk of mortality in COVID-19. *Diagnostics*, *11*(8), 1383. Doi: 10.3390/diagnostics11081383

[9]. Tiwari, S., Chanak, P., & Singh, S. K. (2022). A review of the machine learning algorithms for COVID-19 case analysis. *IEEE Transactions on Artificial Intelligence*, *4*(1), 44-59.

[10]. Zakariaee, S. S., et al. (2023). Comparing machine learning algorithms to predict COVID-19 mortality using a dataset including chest computed tomography severity score data. *Scientific reports*, *13*(1), 11343. Doi: 10.1038/s41598-023-38133-6

[11]. Puttegowda, K., et al. (2024). Automatic COVID-19 Prediction with Comprehensible Machine Learning Models. *The Open Public Health Journal*, *17*(1). Doi: 10.2174/0118749445286599240311102956

[12]. Yang, J., et al. (2020). Prevalence of comorbidities in the novel Wuhan coronavirus (COVID-19) infection: a systematic review and meta-analysis. *Int J Infect Dis*, *94*(1), 91-95. Doi: 10.1016/j.ijid.2020.03.017

[13]. Arlis, S., & Defit, S. (2021). Machine learning algorithms for predicting the spread of COVID-19 in Indonesia. *TEM Journal*, *10*(2), 970–974. Doi: 10.18421/TEM102-61

[14]. Mano, L. Y., et al. (2023). Machine learning applied to COVID-19: a review of the initial pandemic period. *International Journal of Computational Intelligence Systems*, *16*(1), 73. Doi: 10.1007/s44196-023-00236-3

[15]. Shakeel, S. M., et al. (2021). COVID-19 prediction models: a systematic literature review. *Osong public health and research perspectives*, *12*(4), 215. Doi: 10.24171/j.phrp.2021.0100

[16]. Public Health Institute of North Macedonia. (2021). *Movement of Acute Infectious Diseases in North Macedonia in 2020.* Iph.mk. Retrieved from: https://iph.mk/ [accesed: 09 September 2024]

[17]. Cunningham, P., & Delany, S. (2007). k-Nearest neighbour classifiers – A tutorial. *ACM Computing Surveys, 54*(6), 1-25. Doi: 10.1145/3459665

[18]. Rokach, L., & Maimon, O. (2005). *Decision trees. Data mining and knowledge discovery handbook*, 165-192. Springer New York. Doi: 10.1007/0-387-25465-X_9

[19]. Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica*, *24*(1), 12-18. Doi: 10.11613/BM.2014.003

[20]. Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. *Ensemble machine learning: Methods and applications*, 157-175. Doi:10.1007/978-1-4419-9326-7_5

[21]. Ali, Z. A., et al. (2023). Exploring the power of eXtreme gradient boosting algorithm in machine learning: A review. *Academic Journal of Nawroz University*, *12*(2), 320-334. Doi: 10.25007/ajnu.v12n2a1612

[22]. Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2009). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE transactions on systems, man, and cybernetics-part A: systems and humans*, *40*(1), 185-197.

[23]. WEKA software. (n.d.). *The Weka Workbench.* WEKA. Retrieved from: https://ml.cms.waikato.ac.nz/weka/ [accessed 02 July 2024].