Journal of Digitovation and Information System

http://jdiis.de/index.php/jdiis

A Model for Prioritizing Covid-19 Vaccination Distribution Using Electronic Health Record

Shamsudeen Ademola Sanni[a,*] Phathizwe Vilakazi[b]

[a] *Department of Computer Science, Faculty of Science and Engineering, University of Eswatini*

Abstract

Covid-19 vaccination drives in most African and less developed countries have been very slow due to the limited number of vaccine doses available. In order to optimize the vaccination drive, it becomes imperative to create a priority list for vaccination based on the risk of mortality. The study applied Machine Learning (ML) techniques to develop a model for prioritizing Covid-19 vaccination. The dataset for training and validating the model was collected from the Kaggle datasets repository which consists of medical information for 4711 patients with confirmed Covid-19 infection. Four Machine Learning (ML) models were trained, tested and validated on the 4711 patients' health records. The performances of the models were compared based on their precision, sensitivity, accuracy and area under the curve (AUC) scores. Performance of the four models on the datasets indicated that Multi-Tree XGBoost was the best performing model [precision: (Survival: 0.86, death: 0.74), accuracy: 0.83, AUC: 0.89], followed closely by Random Forest [precision: (survival: (Survival: 0.81, death: 0.83), accuracy: 081, AUC: 0.88]. The multi-Tree XGBoost model was therefore chosen for the model creation. This model can be adopted by health authorities and partners to make informed decisions in the Covid-19 vaccine administration.

Keywords

Priority List, Covid-19, Covid-19 Vaccine, EHRMs, ML Models, Vaccination

Article Information

Received 14 February 2022
Revised 16 August 2022
Accepted 01 September 2022

https://doi.org/10.54433/JDIIS.2022100014
ISSN 2749-5965

1. Background

Coronavirus disease is a zoonotic virus causing many severe respiratory diseases that are closest to SARS-Cov (severe acute respiratory syndrome coronavirus) and MERS-CoV (Middle East respiratory syndrome coronavirus) (Dawood, 2020). Since the advent of the virus on 31 December 2019 in Wuhan City China, the world has been crippled by the Covid-19 pandemic. Economy, education, healthcare, sports and tourism are just a few of many sectors to have been severely affected by the prevalence of the pandemic. The devastating loss of human lives has been recorded and cases are increasing with each subsequent day with wave after wave of fear, anger, and anxiety, as the virus mutates and changes its structure. Covid-19 infections are still rising in 50 countries, with at least 253,206,000 reported infections and 5,350,000 reported deaths as of November 2021 (Reuters, 2021).

Efforts to stem the tide of infection include vaccination drives across the world, which was affected by low vaccine production, distribution and vaccine hesitancy (Dzinamarira et al., 2021; Lim et al., 2022; Mellet & Pepper, 2021) and this calls for proper prioritization of vaccination which is the main problem set out to be solved by this research.

.

*Corresponding author: e-mail addresses: sanniade01@gmail.com (S.A. Sanni)

Since the supply of coronavirus vaccines is very limited in less developed countries, effective administration of vaccines to maximize their effects is paramount. On the 11th of December, 2020, the first vaccine for use known as Pfizer-BioNTech COVID-19 Vaccine was approved by the United States Food and Drug Administration (FDA). With this emergency use authorization, the Pfizer-BioNTech Covid-19 vaccine was distributed and started being used to vaccinate the US population. Soon after the approval of the Pfizer vaccine, other vaccines were soon approved towards the end of 2020 and the beginning of the year 2021. By the end of the year 2021, eight vaccines have been approved by the World Health Organization which include Moderna, Pfizer/BioNTech, Janssen/Johnson & Johnson, Oxford/AstraZeneca, Covishield (Oxford/AstraZeneca formulation), Bharat Biotech/Covaxin, Sinopharm (Beijing), BBIBP-CorV (Vero Cells) and Sinovac/CoronaVac.

Many countries around the world experienced shortages of vaccines during the initial roll-out of vaccines and this problem persists. Countries were faced with a huge predicament of maximizing the effectiveness of the vaccine rolled out to their populations due to the limited number of vaccines that were made available. Most countries prioritize the frontline workers and then the elderly group as they were perceived to be more vulnerable to contracting the virus as opposed to the other demographics. Such an approach meant that many people that would have a higher probability of succumbing to a Covid-19 infection due to other factors besides being frontline workers or elderly were ignored because there is no system available to give priority to those who mostly and timely require the vaccines to stay alive. A system that can predict Covid-19 mortality earlier and faster based on medical records would assist health authorities to administer vaccine doses in order to save lives.

As noted, the administration of vaccines in most countries was based on the job description and age. While this strategy and other regulations have been effective in slowing down Covid-19 infection rates, it could be optimized using machine learning models to create priority lists of vaccine candidates. Available research related to predicting mortality due to CovidD-19 has mainly focused on infected patients. For example, Schwab et al., (2021) created a model that could predict mortality from Covid-19 infected patients with very promising results, ranging from 78.8% with a 95% confidence interval [CI]: 76.0, 84.7%) to 69.4 (95% confidence interval: 51.6, 75.2%) with specificity at sensitivities greater than 95 %. This system was useful in hospitals to assign priority to patients that were at a greater risk of succumbing to Covid-19 infections. In Estiri et al., (2021) age-stratified generalized linear models (GLMs) with component-wise gradient boosting were utilized to predict the probability of death from Covid-19 infections based only on past medical health records. The model developed by Zoabi et al., (2021) predicted the probability of infection using eight binary risk factors amongst them including age, which is known to be an important Covid-19 risk factor and the appearance of initial clinical symptoms of the virus. In Wang et al., (2020), a deep learning algorithm that used CT images to screen for Covid-19 infection was used. The algorithm predicted the presence of the infection based on radiographic changes in CT images. It extracted Covid-19's graphical features in order to provide a clinical diagnosis ahead of a pathogenic clinical test of the infection. Muhammed et al., (2020) designed supervised machine learning models for the prediction of Covid-19 infection from an epidemiology dataset. The models were trained on a tagged dataset from Mexico for Covid-19 infection. The Dataset contained demographic and clinical data as well as the Reverse Transcription Polymerase Chain Reaction (RT-PCR) testing results for Covid-19 infection. The models were trained on 11 risk factors from the data set and these risk factors included features like gender, age and commodities like the presence of pneumonia, diabetes and asthma. The results from the models showed the decision tree to be the best model amongst all the other models which included Logistic Regression, Naïve Bayes, Support Vector Machine and Artificial Neural Network in terms of accuracy of 94.99. Similarly, Gong, et al., (2020) designed a tool for the early prediction of severe Covid-19 disease among patients. The model identified prognostic risk factors of Covid-19 mortality among patients to be older age, higher lactate dehydrogenase (LDH), C-reactive protein (CRP), and many other clinically tested risk factors that correlated with the odds of patients developing severe Covid-19. From these risk factors, an effective prognostic nomogram was developed and validated with high sensitivity and specificity for accurate individualized assessment of patients who develop severe Covid-19. This model identified the turnover of red blood cells to be

involved in severe illness. In the work of Schwab, et al., (2021), a real-time prediction of Covid-19 related mortality using electronic medical records was developed that did not only focus on identifying and using just risk factors but also focused on risk factors that changed over time in the prediction of Covid-19 mortality. From this model, a Covid-19 early warning system (CovEWS) was devised and predicted Covid-19 mortality from 78.8% (95% confidence interval (CI): 76.0%, 84.7%) to 69.4% (95% CI: 57.6%, 75.2%) specificity at sensitivities greater than 95%. This system could be used to serve as an early warning to health practitioners of patients that were at risk of mortality due to Covid-19 so that it could be dedicated to them. In essence, there have been numerous studies dedicated to the prediction of Covid-19 infection and the prognosis of the virus in patients using machine learning techniques. These experiments focused on designing, developing and training predictive models to complement traditional healthcare activities and perform very important tasks faster.

Meanwhile, virology experts have warned that the world might have to live with Covid-19 for years to come and we might likely face a similar pandemic in the future, therefore it becomes pertinent to have a system that informs health authorities on individuals that should be prioritised for vaccination in order to reduce the burden of healthcare workers and resources. We witnessed how hospitals and clinics were overwhelmed during the peak period of the Covid-19 pandemic in many countries, leading to hospitals rejecting patients and health authorities advising infected individuals to adopt home care. This could have been properly managed with a system that could reliably predict patients who are more vulnerable to the disease and therefore would be accorded special priority in terms of treatment and vaccination in order to reduce mortality. These researchers designed a model that can optimize the vaccination process by creating a vaccination priority list supported by the Electronic Health Record Management System (EHMRS) to reduce the risk of death as a result of Covid-19 infection. The study designed and developed an age-stratified predictive model that can be used to predict mortality after a Covid-19 infection using only the routinely collected electronic medical records and the identification and understanding of the most important risk factors across age groups.

2. Subjects and Method

2.1. Research Framework

The research framework is presented in Figure 1

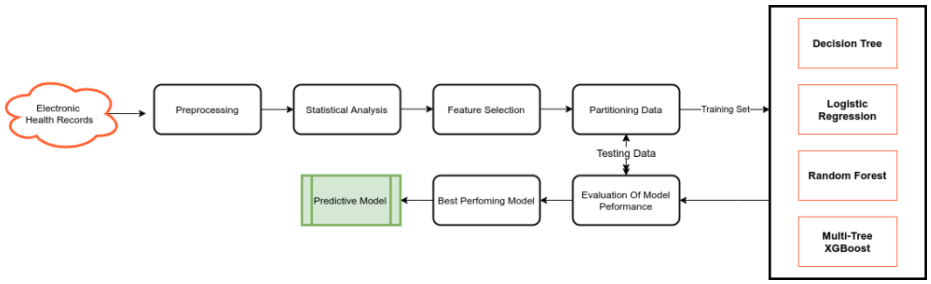

Figure1: Framework for the model development

2.2. Study Design

The study adopts Machine Learning (ML) techniques to create a model for prioritizing Covid-19 vaccine distribution. Datasets for model training and validation were collected from the Kaggle Data Science repository (https://www.kaggle.com/). The data were derived from a healthcare surveillance software package (Clinical Looking Glass [CLG]; Streamline Health, Atlanta, Georgia) and a review of the primary medical records. It relates to Covid-19 patients admitted to a single healthcare system, over a specific period of time, and separated into the first 6 weeks of the pandemic. The data contains a record

of 4711 patients with 85 parameters, out of which only 10 were considered important for testing and validation based on their XGBoost algorithm score. The data file contains information on demographics, comorbidities, admission laboratory values, admission medications, admission supplemental oxygen orders, discharge, and mortality. Some of the variables included in the dataset are the length of hospital stay (LOS), myocardial infarction (MI), peripheral vascular disease (PVD), congestive heart failure (CHF), cardiovascular disease (CVD), dementia (Dement), Chronic obstructive pulmonary disease (COPD), diabetes mellitus simple (DM simple), diabetes mellitus complicated (DM complicated), oxygen saturation (OsSats), mean arterial pressure, in mmHg (MAP), D-dimer, in mg/ml (Ddimer), platelets, in k per mm3 (Plts), international normalized ratio (INR), blood urea nitrogen, in mg/dL (BUN), alanine aminotransferase, in U/liter (AST), white blood cells, in per mm3 (WBC) and interleukin-6, in pg/ml (IL-6).

2.3. Data Analysis

The Multi-Tree XGBoost model was chosen as the model to base the prediction of mortality of coronavirus as it has the best performance [precision: (survival: 0.86, death: 0.74), accuracy: 0.83, AUC: 0.89].

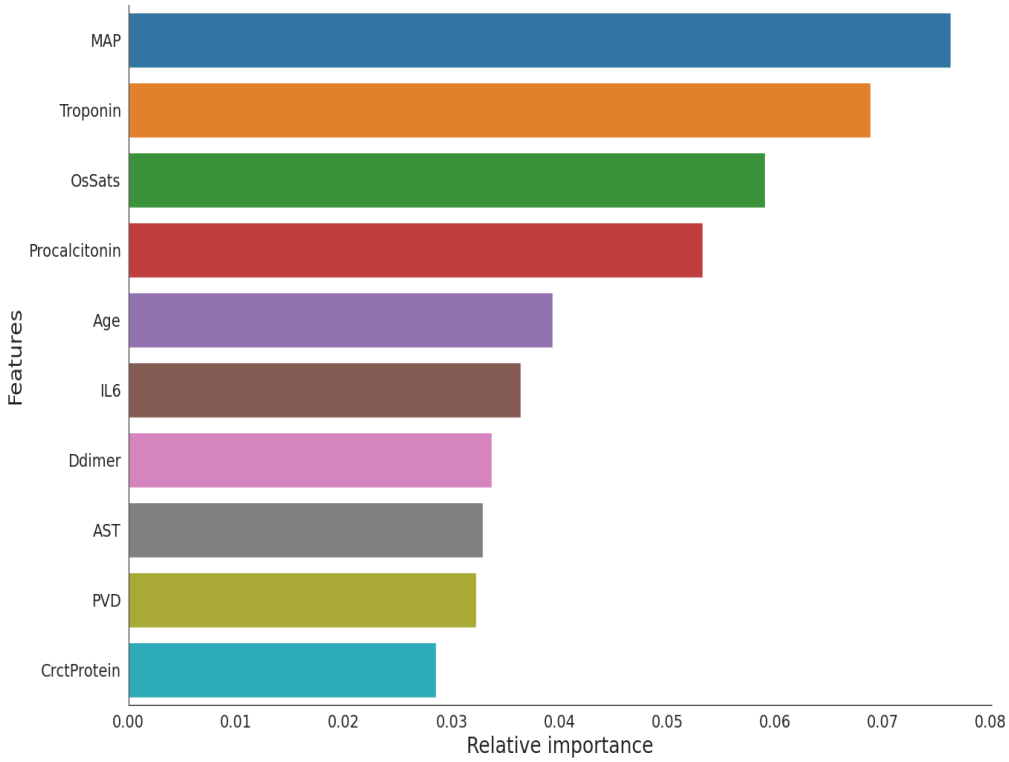

Figure 2: Features of the datasets according to their relative importance in predicting Covid-19 mortality

This model was the best predictive model because of the recursive decision tree structure building from past residuals. Besides, it is capable of identifying those trees that contribute the most to the decision of the predictive model. The features of the datasets according to their relative importance in predicting Covid-19 mortality are presented in Figure 2 and Table 1.

The metrics used in deciding the choice of algorithm to be used in the creation of the prediction models are as follows:

Table 1: Features of the datasets according to their relative importance in predicting Covid-19 mortality

	Col	Xgb
1	MAP (Mean Arterial Pressure)	0.0762
2	Troponin	0.0688
3	OsSats (Oxygen Saturation)	0.05908
4	Procalcitonin	0.0532
5	Age	0.0394
6	IL6 (Laterlenkin-6)	0.0364
7	D-dimer	0.0338
8	AST (alanine aminotransferase)	0.0329
9	PVD (Peripheral Vascular Disease)	0.0323
10	CrctProtein	0.0286

During data pre-processing activity, patient samples that had null values in any of the chosen 10 features were excluded in the training and testing of the machine learning models, while other missing data were replaced by -1 before processing. Classification accuracy, sensitivity/recall, precision and F1 score were the metrics used to assess the models created to find the optimum one for the prediction of mortality. A supervised XGBoost classifier was used as the predictor model. This model has a high performance, and the architecture involved is a recursive tree structure that benefits from great interpretability. PyCharm IDE serves as the development tool for all the data processing, algorithm implementation, model training and model validation. The following libraries were used for the development of the code: a) scikit-learn, b) Pandas, c) NumPy, d) matplotlib, e) seaborn, f) utils_features_selection.

3. Results

The following results were obtained after the training and validation process through the following prediction models

i. Decision Tree training and validation using Confusion Matrix

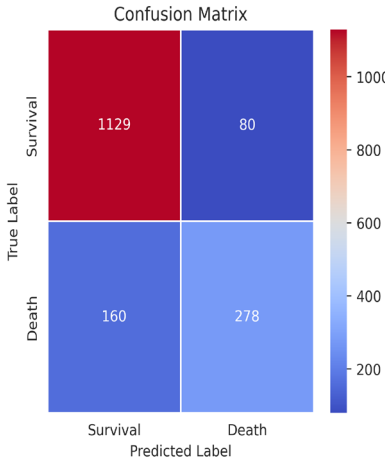

Figure 3: Confusion Matrix for training Decision Tree Model

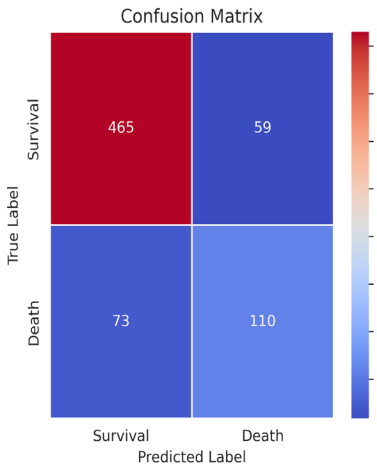

Figure 4: Confusion Matrix for Validation of Decision Tree Model

The confusion matrices depicted in Figure 3 and Figure 4 summarize the performance of the Decision Tree Model, which is a classification algorithm. This gives us a better idea of what our model is getting right and if there is any error.

Table 2: Training performance of the Decision Tree Model

	Precision	Recall	f1-score	Support
Survival	0.88	0.93	0.9	1209
Death	0.78	0.63	0.7	438
Accuracy			0.85	1647
Macro Average	0.83	0.78	0.8	1647
Weighted Average	0.85	0.85	0.85	1647

Table 3: Validation performance of the Decision Tree Model

	Precision	Recall	f1-score	Support
Survival	0.86	0.89	0.88	524
Death	0.65	0.6	0.62	183
Accuracy			0.81	707
Macro Average	0.76	0.74	0.75	707
Weighted Average	0.81	0.81	0.81	707

Table 4: F1-Score and AUC scores for training and validation of the Decision Tree Model

	Training	Standard Deviation	Validation	Standard Deviation	Prediction
F1-Score	0.67	0.02	0.59	0.04	
AUC	0.86	0.01	0.81	0.02	0.88

The figure and Tables 2, 3, and 4 above present the accuracy score and AUC score using the Decision Tree Model.

ii. Logistic Regression training and validation using Confusion Matrix

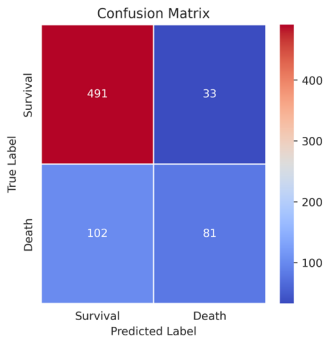

Figure 5: Confusion Matrix for training Logistic Regression Model

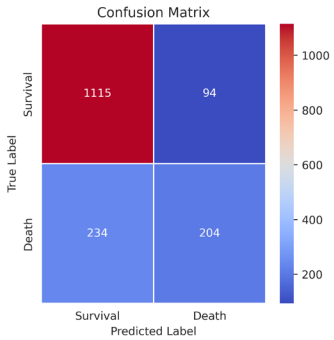

Figure 6: Confusion Matrix for validation Logistic Regression Model

The Logistic Regression depicted in Figure 5 and Figure 6 summarizes the performance of the Logistic Regression Model, which is a classification algorithm. This gives us a better idea of what our model represents.

Table 5: Training performance of the Logistic model

	Precision	**Recall**	**f1-score**	**Support**
Survival	0.83	0.92	0.87	1209
Death	0.68	0.47	0.56	438
Accuracy			0.8	1647
Macro Average	0.76	0.7	0.72	1647
10 Weighted Average	0.79	0.8	0.79	1647

Table 6: Validation performance of the Logistic Regression Model

	Precision	**Recall**	**f1-score**	**Support**
Survival	0.83	0.94	0.88	524
Death	0.71	0.44	0.54	183
Accuracy			0.81	707
Macro Average	0.77	0.69	0.71	707
Weighted Average	0.8	0.81	0.79	707

Table 7: F1-Score and AUC scores for training and validation of the Logistic Regression Model

	Training	Standard Deviation	Validation	Standard Deviation	Prediction
F1-Score	0.54	0.02	0.53	0.04	
AUC	0.80	0.01	0.8	0.02	0.79

The figure and Tables 5, 6, and 7 above present the accuracy score and AUC score using Linear Regression.

iii. Random Forest training and validation using Confusion Matrix

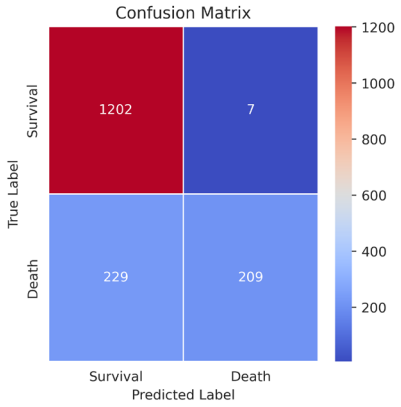

Figure 7: Confusion Matrix for training Random Forest Model

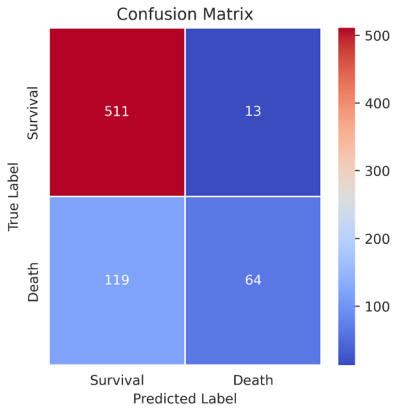

Figure 8: Confusion Matrix for validating Random Forest Model

The Random Forest Model depicted in Figure 7 and Figure 8 summarizes the performance of the Random Forest Model, which is a classification algorithm. This gives us a better idea of the performance of the model.

Table 8: Training performance of the Logistic model

	Precision	Recall	f1-score	Support
Survival	0.84	0.99	0.91	1209
Death	0.97	0.48	0.64	438
Accuracy			0.86	1647
Macro Average	0.9	0.74	0.77	1647
10 Weighted Average	0.87	0.86	0.84	1647

Table 9: Validation performance of the Random Forest Model

	Precision	Recall	f1-score	Support
Survival	0.81	0.98	0.89	524
Death	0.83	0.35	0.49	183
Accuracy			0.81	707
Macro Average	0.82	0.66	0.69	707
Weighted Average	0.82	0.81	0.78	707

Table 10: F1-Score and AUC scores for Training and Validation of the Random

	Training	Standard Deviation	Validation	Standard Deviation	Prediction
F1-Score	0.62	0.01	0.54	0.03	
AUC	0.91	3.02e-3	0.88	0.01	0.88

The figure and Tables 8, 9, and 10 above present the accuracy score and AUC score using Random Forest Model.

iv. Multi-Tree XGBoost training and validation using Confusion Matrix

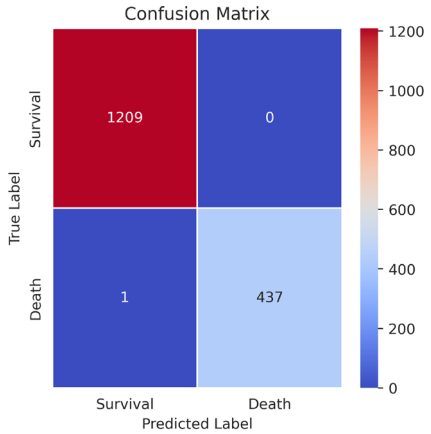

Figure 9: Confusion Matrix for training Multi-Tree XGBoost Model

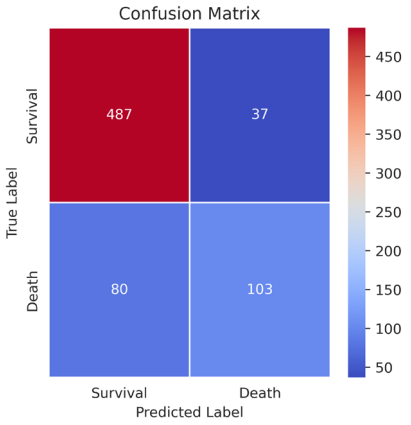

Figure 10: Confusion Matrix for Validation Multi-Tree XGBoost Model

The Multi-Tree XGBoost depicted in Figure 9 and Figure 10 summarizes the performance of the Multi-Tree XGBoost Model, which is a classification algorithm. This gives us a better idea of the performance of the model.

Table 11: Training performance of the Multi-Tree XGBoost model

	Precision	Recall	f1-score	Support
Survival	1	1	1	1209
Death	1	1	1	438
Accuracy			1	1647
Macro Average	1	1	1	1647
10 Weighted Average	1	1	1	1647

Table 12: Validation performance of Multi-tree XGBoost Model

	Precision	Recall	f1-score	Support
Survival	0.86	0.93	0.89	524
Death	0.74	0.56	0.64	183
Accuracy			0.83	707
Macro Average	0.8	0.75	0.77	707
Weighted Average	0.83	0.83	0.73	707

Table 13: F1-Score and AUC scores for Training and Validation of the Multi-Tree Model

	Training	Standard Deviation	Validation	Standard Deviation	Prediction
F1-Score	0.998	1.16e-3	0.7	0.03	
AUC	0.99999	1.26e-0.5	0.89	0.01	0.89

The results presented in tables 11, 12, and 13 and the figures above indicated that Multi-Tree XGBoost was the best performing model in predicting Covid-19 mortality with an accuracy score of 0.83 and AUC score of 0.89. This is due to its recursive decision tree structure building from past residuals. It is capable of identifying those trees that contribute the most to the decision of the prediction model. This model can be adopted by health authorities and partners to make informed decisions in vaccine administration.

4. Discussion and Conclusion

The administration of vaccines since the start of the vaccination drive in most countries is generally based on whether an individual is a frontline worker or an elderly. While this strategy has been effective, it could be optimized using machine learning models to create much better priority lists of vaccine candidates which could be especially very useful in less developed countries with limited vaccine doses. Findings from this research study suggest that the best performing model in predicting Covid-19 mortality is the Multi-Tree XGBoost with an accuracy score of 0.83 and AUC score of 0.89. This model is capable of identifying those trees that contribute the most to the decision of the prediction model and

it could be used to design priorities for Covid-19 vaccination. This model can be adopted by health authorities and partners to make informed decisions in vaccine administration. Many African and less developed countries have to lobby and queue to receive batches of COVID-19 vaccines from the developed nations. It is therefore imperative for health authorities in Africa and less developed countries to have a system of the priority list that can capture the most vulnerable individuals that are likely to succumb to COVID-19 infections and make them a priority for the limited doses of the vaccines. This study demonstrates the use of Machine Learning (ML) models in creating a vaccination priority list to reduce the risk of death as a result of COVID-19 infections. This research designed a model that can optimize the vaccination process by creating a vaccination priority list supported by the Electronic Health Record Management System (EHMRS) to reduce the risk of death as a result of COVID-19 infection. The application of predictive models in COVID-19 pandemic mitigation and control has led to spectacular outcomes and endeavour in this area continues to grow. Africa and other less developed regions of the world suffered from a lack of resources in fighting COVID-19 pandemic and there is no evidence that COVID-19 pandemic will end anytime soon. Therefore, it is important to accelerate vaccination drives to reduce COVID-19 deaths. The simple cost-effective approach recommended by this research could save lives and help many countries in fighting the pandemic and could be adopted for future disease outbreaks. Scientists and virologists have warned that this might not be the last we would see disease outbreaks and pandemics. While COVID-19 is being contained and vaccines are being rolled out, there are still reported cases of high infections around the world. Therefore, this research work is still very relevant and would be useful now and in the future.

.

Reference

Dawood, A. A. (2020). Mutated COVID-19 may foretell a great risk for mankind in the future. New microbes and new infections, 35, 100673.

Dzinamarira, T., Nachipo, B., Phiri, B., and Musuka, G. (2021). Covid-19 vaccine roll-out in South Africa and Zimbabwe: Urgent need to address community preparedness, fears and hesitancy. Vaccines, 9(3), 1–10. https://doi.org/10.3390/vaccines9030250

Estiri, H., Strasser, Z. H., Klann, J. G., Naseri, P., Wagholikar, K. B., and Murphy, S. N. (2021). Predicting COVID-19 mortality with electronic medical records. NPJ digital medicine, 4(1), 1-10.

Government of Eswatini (2021). COVID-19 Information. Retrieved March 10, 2021, from https://sz.usembassy.gov/covid-19-information/

Gong, J., Ou, J., Qiu, X., Jie, Y., Chen, Y., Yuan, L., and Hu, B. (2020). A tool for early prediction of severe coronavirus disease 2019 (COVID-19): a multicenter study using the risk nomogram in Wuhan and Guangdong, China. Clinical infectious diseases, 71(15), 833-840.

Lim, J., Norman, B. A., and Rajgopal, J. (2022). Redesign of vaccine distribution networks. International Transactions in Operational Research, 29(1), 200–225. https://doi.org/10.1111/itor.12758

Mellet, J., and Pepper, M. S. (2021). A covid-19 vaccine: Big strides come with big challenges. Vaccines, 9(1), 1–14. https://doi.org/10.3390/vaccines9010039

Muhammad, L. J., Algehyne, E. A., Usman, S. S., Ahmad, A., Chakraborty, C., and Mohammed, I. A. (2021). Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. SN computer science, 2(1), 1-13.

Reuters (2021) COVID-19 Global tracker, Reuters research. Available at: https://graphics.reuters.com/world-coronavirus-tracker-and-maps/.

Schwab, P., Mehrjou, A., Parbhoo, S., Celi, L. A., Hetzel, J., Hofer, M., and Bauer, S. (2021). Real-time prediction of COVID-19 related mortality using electronic health records. Nature communications, 12(1), 1-16.

Wang, C., Wang, Z., Wang, G., Lau, J. Y. N., Zhang, K., and Li, W. (2021). COVID-19 in early 2021: current status and looking forward. Signal Transduction and Targeted Therapy, 6(1), 1-14.

Wang, S., Kang, B., Ma, J., Zeng, X., Xiao, M., Guo, J., and Xu, B. (2021). A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). European radiology, 31(8), 6096-6104.

Wedlund, L., and Kvedar, J. (2021). New machine learning model predicts who may benefit most from COVID-19 vaccination. NPJ Digital Medicine, 4(1), 1-1.

Zoabi, Y., Deri-Rozov, S., and Shomron, N. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj digital medicine, 4(1), 1-5.