Predicting COVID-19 Mortality Rates using Machine Learning Models

Afiefah binti Zulkifli[1], Safia Adrina binti Mohd Zulkifli[2], Xu Jianhong[3], Mohamed Moubarak Mohamed Misbahou Mkouboi[4]

Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
43600, Bangi, Selangor, Malaysia
P145215@siswa.ukm.edu.my[1], P138679@siswa.ukm.edu.my[2], P139575@siswa.ukm.edu.my[3], P144469@siswa.ukm.edu.my[4]

Abstract— **Coronavirus disease (COVID-19) was a global pandemic outbreak that started in 2020 and ended in 2023. It is an infectious disease that spread thru airborne. COVID-19 infection caused most of people to experience mild to moderate respiratory problem but they recovered without needing any special treatment. However, people in the high-risk group categories which includes pregnant women, elderly and individuals with underlying medical conditions like cancer, diabetes, cardiovascular disease and chronic respiratory disease have higher tendency to develop serious illness that can lead to mortality. Medical supply shortage was a serious issue during the pandemic and it is vital to know which patients should be prioritized to get access to the treatment. This study aims to forecast which factors play a role in determining the mortality rate caused by COVID-19 using machine learning model. The COVID-19 mortality dataset was obtained from Kaggle, an opensource platform providing data from various fields. In this study, three machine learning models were built to predict COVID-19 mortality rates including logistic regression, random forest and neural network. This study found that the neural network model achieved the highest AUC score of 0.96 and a test accuracy of 0.9103, outperforming logistic regression and random forest models. The model's high accuracy and reliability make it a valuable tool for healthcare providers to prioritize patient care and allocate resources effectively, especially during resource-constrained scenarios like a pandemic. This study provides a foundation for further exploration and refinement of predictive models, aiming to contribute to more effective and informed healthcare delivery.**

Keywords—COVID-19, COVID-19 Mortality Prediction; Machine Learning Models; Neural Network; Medical Resource Allocation; High-Risk Patiens

I. Introduction

The coronavirus disease (COVID-19) first emerged in Wuhan, Hubei, China in December 2019. The virus then spread rapidly to neighbouring countries and has impacted the global health, causing significant morbidity and mortality. Medical supply shortage was a serious issue during the pandemic that leads to changes in trade in industrial policy globally. According to Brown (2021), China has increased the imports and decreased exports of their personal protective equipment; thus, removing supplies from the global market [1]. An effective management and timely intervention for COVID-19 patients require accurate risk stratification to prioritize medical resources and provide targeted care. In this context, machine learning (ML) is a powerful tool to analyse large datasets and uncover patterns that can assist in predicting health outcomes.

A predictive model using machine learning algorithms can assist healthcare providers in making informed decisions, thereby enhancing patients' outcomes and optimizing resource allocation during the ongoing pandemic. Throughout the pandemic, a significant challenge for healthcare providers has been the shortage of medical resources and the lack of an effective distribution plan. Given the scarcity of medical supplies, it is crucial to allocate resources to patients with the highest chances of survival. A machine learning model can help healthcare providers prioritize patients based on their likelihood of survival, ensuring that medical resources are used efficiently.

There are many studies that have utilised machine learning to analyse the mortality rate of COVID-19. For instance, Aslam and Biswas (2023) found that various regression machine learning model including XGBoost, Random Forest, and SVM can be employed to predict death cases in the near future during epidemic [2]. Pourhomayoun and Shakibi (2021) used AI model to facilitate medical practitioners prioritize patients with higher chance of surviving when there is medical supply shortage [3]. Yadaw et al. (2020) used machine learning to predict COVID-19 death cases based on five clinical features including age, minimum oxygen saturation during encounter, type of patient encounter, hydroxychloroquine uses and maximum body temperature, as guidance to manage patients better during the pandemic [4]. This study aims to forecast which factors play a role in determining the mortality rate caused by COVID-19 using machine learning model.

II. Literature Review

Predictive models have to be developed quickly in order to help healthcare practitioners manage patient outcomes in the wake of the COVID-19 epidemic. Four studies that use machine learning techniques to predict COVID-19 death rates and severity have been reviewed in this literature review. The

methods used, important findings, and possible implications for clinical decision-making have all been highlighted.

Mohammad Pourhomayoun and Mahdi Shakib [3] has done the study that aims to develop a predictive model using machine learning (ML) algorithms to determine the mortality risk of COVID-19 patients. Utilizing a dataset of over 2.67 million COVID-19 patients from 146 countries, the study applied various ML algorithms including Neural Networks, Support Vector Machine, Random Forest and others with Neural Networks achieving the highest accuracy of 89.98%. The dataset include demographic, physiological and symptomatic data, with several key features such as age, gender, respiratory distress, diabetes and hypertension that are significantly impacting mortality risk. The studies have highlighted the feature selection process and evaluation using 10-fold cross-validation demonstrated the model effectiveness in predicting mortality risk, providing a valuable tool for medical decision-making and patient triage during pandemics.

Zakariaee et al. (2023) [5] has purpose a present study to develop an efficient ML prognostic model based on a more comprehensive dataset including chest CT severity score (CT-SS). The predictive models were developed using eight ML algorithms including the Random Forest (RF), J48 decision tree (J48), support vector machine (SVM), multi-layer perceptron (MLP), k-nearest neighbour (kNN) and others. The performances of the predictive models were evaluated using accuracy, precision, sensitivity, specificity and area under the ROC curve (AUC) metrics. After applying the exclusion criteria with total of 815 positive RT-PCR patients, where 54.86% of the patients were male and the mean age of the study population was 16.76years. The Random Forests algorithm with an accuracy of 97.2%, the sensitivity of 100%, a precision of 94.8%, specificity of 94.5%, F1-score of 97.3% and AUC of 99.9%, had the best performance.

Jamshidi et al. [6] has also done a study with an objective for early prediction of mortality using machine learning based on typical laboratory results and clinical data registered on the day of ICU admission. The researcher is respectively studied 797 patients that are diagnosed with COVID-19 in Iran and the United Kingdom (U.K). Several machines learning algorithms, including Random Forest (RF), logistic regression, support vector machine, artificial neural network and gradient boosting classifier were utilized to build classifications models. As for investigation of model performance, the Random Forest show superior performance on validation sets. The Random Forest model predicts patient outcomes with a 70% sensitivity and 75%specificity.

Emami et al. [7] has introduced research to compare four machine learning algorithms for predicting mortality in COVID-19 machine disease. The data were collected from hospitalized patients with COVID-19 in five hospitals in Tehran (Iran). Database that contained 4120 records, about 25% of which belong to patients who died due to COVID-19. Four machine learning techniques, including random forest (RF), regression logistic (RL), gradient boosting tree (GBT) and support vector machine (SVM) were used in this research modelling. Gradient Boosting Tree model presented higher performance compared to other models followed with Random Forest with accuracy of 70%, sensitivity of 77%, specificity of 69% and the ROC area under the curve 0.857. Random Forest, Regression Logistic and Support Vector machine with the ROC area under the curves 0.836, 0.818 and 0.794 were in the second and third places. In the study, the researcher has highlight that considering the combination of multiple influential factors affecting death COVID-19 can help in early prediction and providing better care plan. In addition, using different modelling on data can be useful for physician in providing appropriate care.

Bottino et al. [8] has developed a predictive model using various machine learning algorithms to assess the mortality risk of COVID-19 patients based on their physiological conditions, symptoms and demographic algorithm. Methods of machine learning algorithms such as neural networks, logistic regression, support vector machine and others has been used. The models' performance was evaluated using metrics like accuracy, AUC-ROC, sensitivity, and specificity, achieving high accuracy of 89.98% with a Neural Network Model.

In summary, these past few studies collectively highlight the effectiveness of various machine learning algorithms in predicting COVID-19 mortality, demonstrating significant advancements in model accuracy and clinical applicability for better patient management and decision-making during the pandemic.

III. Research Methodology

The COVID-19 epidemic caused a critical need for prediction algorithms to help identify high-risk patients. This study uses machine learning algorithms to predict COVID-19 death rates using a large dataset of anonymised patient data. The study aims to give healthcare practitioners actionable insights by using data pre-processing techniques, selecting relevant features, and constructing powerful predictive models. The methodology includes data collection, preparation, pre-processing, feature selection, development of models, evaluation, and visualization to ensure accurate and interpretable results (Figure 1). The results of this research have the potential to greatly help with healthcare planning and resource allocation. Identifying high-risk individuals allows healthcare practitioners to adopt focused therapies that reduce mortality rates.

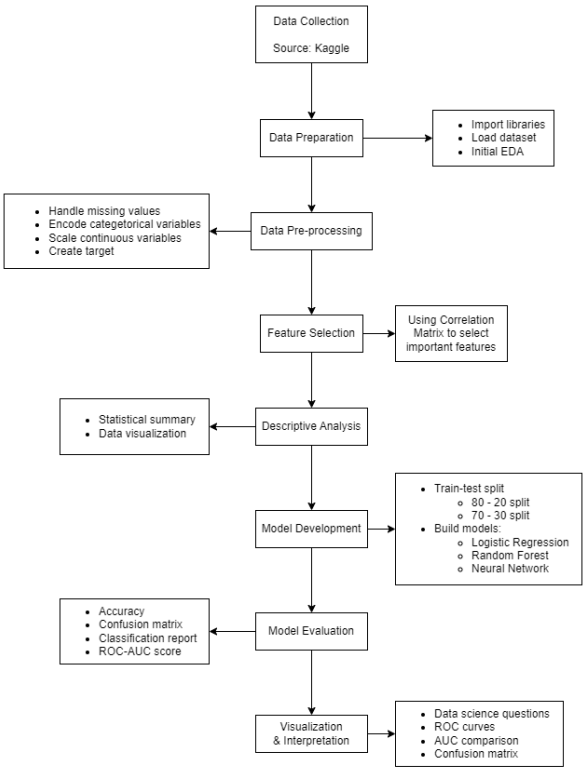


Figure 1. Workflow for COVID-19 Mortality Prediction Using Machine Learning Models

A. *Data Preparation*

The dataset used in this study was collected from the "Covid19-dataset" on Kaggle. This dataset consists of many anonymized patient data, including pre-existing conditions, symptoms, and results. It is composed of 21 different features and data from 1,048,576 unique patients. The major goal is to use this data to predict COVID-19 mortality rates with machine learning algorithms.

B. *Data Pre-processing*

As part of our preliminary analysis, we aim to answer many critical issues through our exploration of the data. Each row in the database represents a unique COVID-19 patient, including detailed information about each case. The columns represent specific features of these patients, such as demographic information, pre-existing conditions, and symptoms. The "DATE_DIED" property generated the target output, the “DEAD” variable. If a patient has passed away, the date of death is mentioned; otherwise, a placeholder value of 9999-99-99 is used. The target variable was set to 1 for deceased patients and 0 for survivors.

Each column contains data categorized into three levels: nominal (e.g., gender, symptoms, pre-existing conditions), ordinal (e.g., symptom severity), and binary (e.g., presence of conditions), with age treated as a continuous variable. Initially, the dataset included missing values represented by 97, 98, and 99, primarily in features such as "ICU," "INTUBED," and "PREGNANT”. Figure 2 shows a depiction of the missing values before processing them.

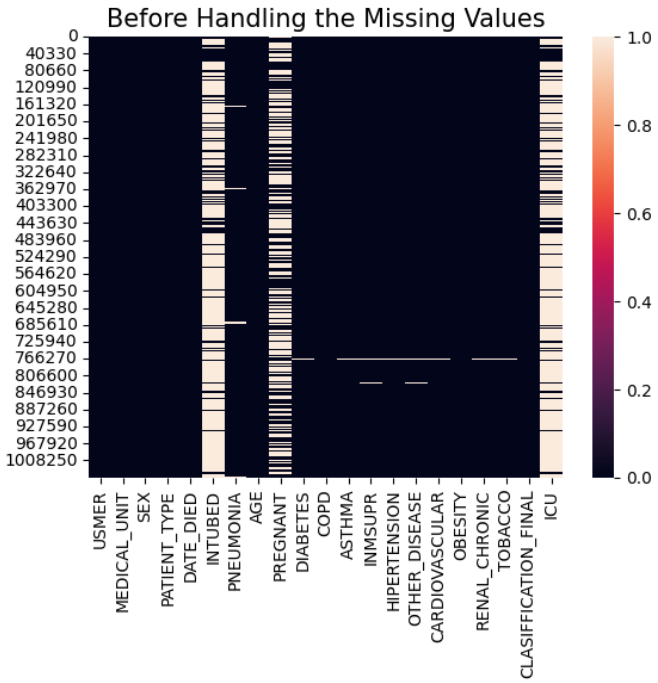


Figure 2. Depiction of the missing values before data pre-processing.

To handle these missing values, certain rules were used:

- Values of 97, associated with males, were substituted with 2, indicating that males cannot become pregnant. This is done by associating "PREGNANT” feature with the “SEX” feature and 97 was associated with the number 2 representing male.

- Non-hospitalized patients were unable to receive ICU and INTUBED therapies, hence 97 values were replaced by 2. From this procedure we noticed that the missing values of (97) are all corresponding to the values of PATIENT_TYPE = 1 which is for non-hospitalized patients, while those of (99) are the missing values of the hospitalized patients. So, we replaced all the values of (97) with (2); since patients who have never been hospitalized couldn't possibly be admitted to the ICU. The same procedure was applied to INTUBED feature as a non-hospitalized patient cannot be connected to the ventilator.

- Values 98 and 99, which could not be reliably imputed, were dropped to maintain data integrity. we cannot replace the female section (98) with (1) because we don't know exactly whether those females were pregnant or not. And we cannot replace the hospitalized patient section (99) with (1) because we don’t know exactly whether a hospitalized patient was

connected to the ventilator or not and whether a patient who was hospitalized was admitted to the ICU or not.

Following the implementation of these rules, the dataset was cleaned of missing values, as shown in Figure 3.

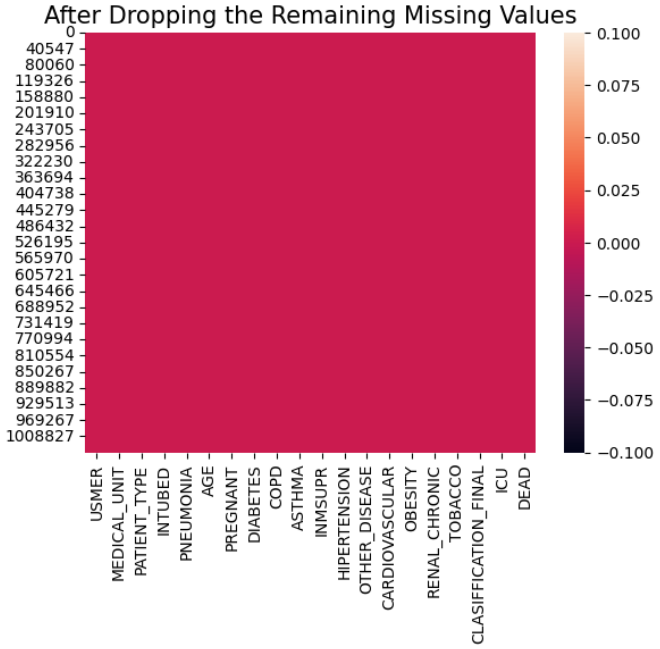


Figure 3. Depiction after handling the missing values.

Continuous variables were scaled as part of the necessary changes. We used "StandardScaler" to scale our dataset because all of the features chosen for modeling were numerical. The scaler was applied to the training data and then transformed both the training and testing datasets. These methods ensured that the data was in a format suited for model training, which improved the prediction models' accuracy and robustness.

Following data preprocessing, a correlation heatmap was created to determine the most relevant features for predicting COVID-19 mortality. This heatmap helped to visualize the correlations between the features and the target variable. Features with a low correlation to the target variable were considered less important for the prediction model and were therefore dropped. Figure 4 shows the initial correlation heatmap, which includes all features.

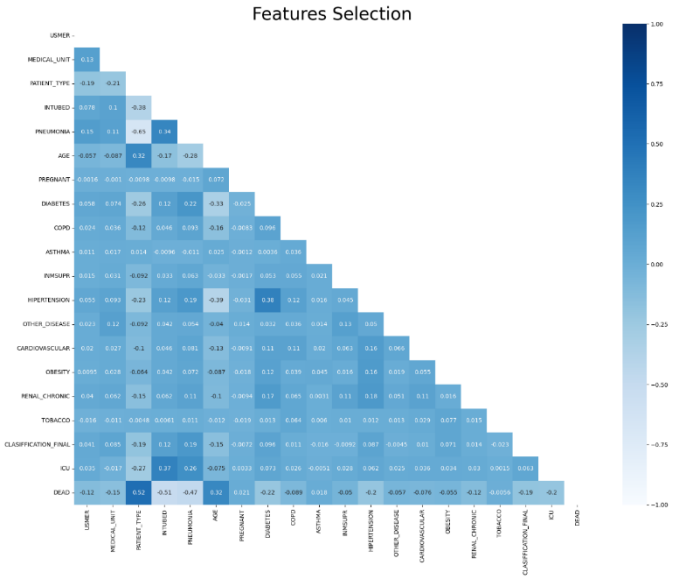


Figure 4. Initial correlation heatmap.

The following features were dropped from the heatmap analysis due to their low association with the target variable: "PREGNANT," "COPD," "ASTHMA," "INMSUPR," "OTHER_DISEASE," "CARDIOVASCULAR," "OBESITY," and "TOBACCO." Figure 5 shows the new dataset with the remaining attributes as an updated correlation heatmap.

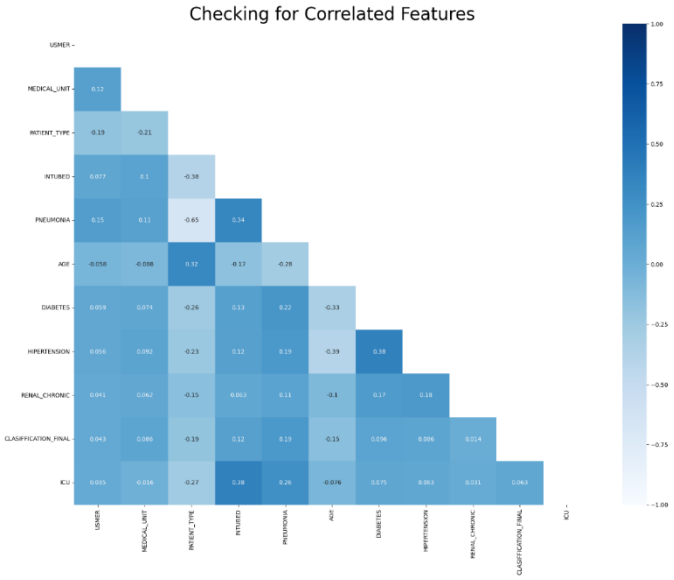


Figure 5. Updated correlation heatmap.

Scaling the features was required to guarantee that all variables contributed equally to model training. Without scaling, features with larger ranges could outperform those with shorter ranges, resulting in biased results. The remaining features were picked because of their stronger association with the target variable, ensuring that the model focused on the most important data for predicting COVID-19 mortality.

C. Descriptive Analysis

After pre-processing the data, a statistical analysis was performed on each attribute to summarize its important attributes. This analysis reveals the central tendency, dispersion, and shape of the distribution of the selected attributes. The descriptive statistics of each attribute and data description of the dataset are shown under the appendix section.

The following figures help us understand the dataset using descriptive analysis.

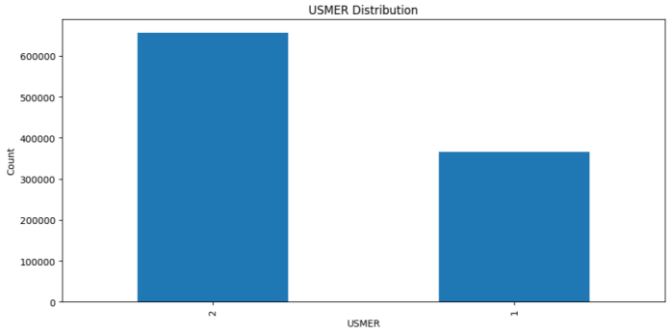


Figure 6. the number of patients treated at USMER facilities.

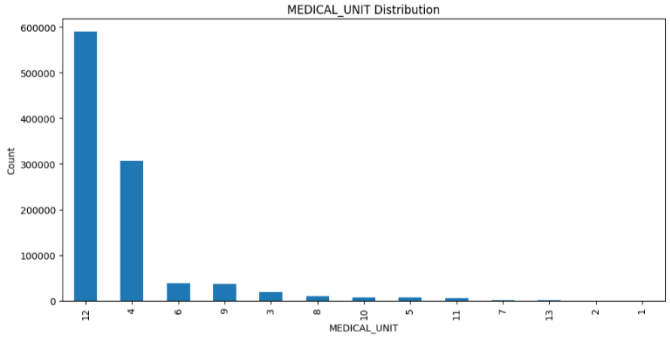


Figure 7. The number of patients treated in various types of medical units.

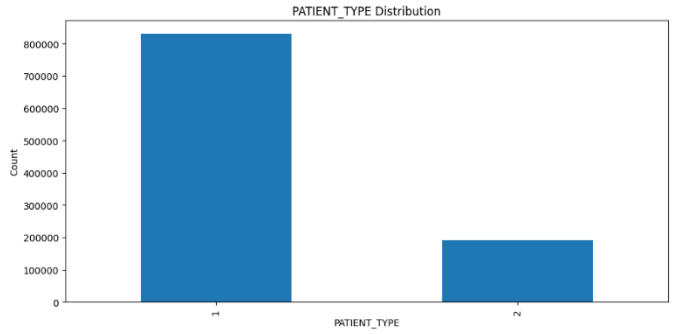


Figure 8. The number of outpatient and inpatient categories.

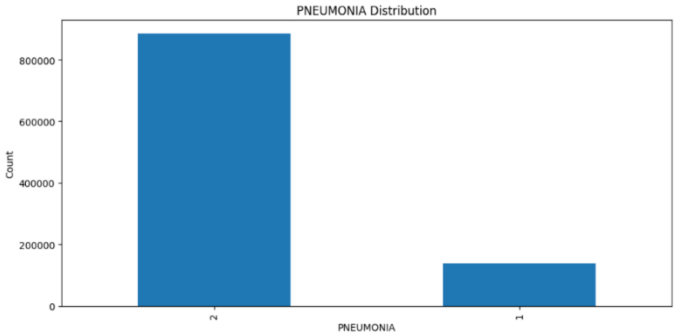


Figure 9. The number of patients with and without pneumonia.

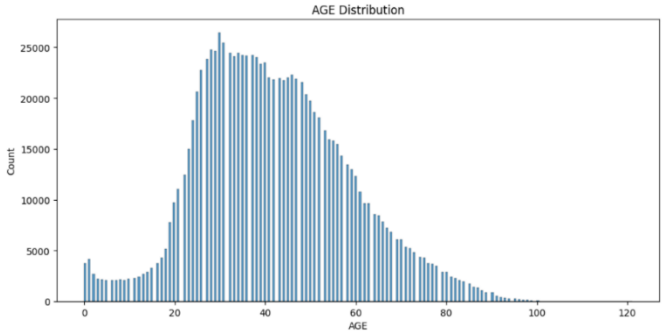


Figure 10. The age distribution.

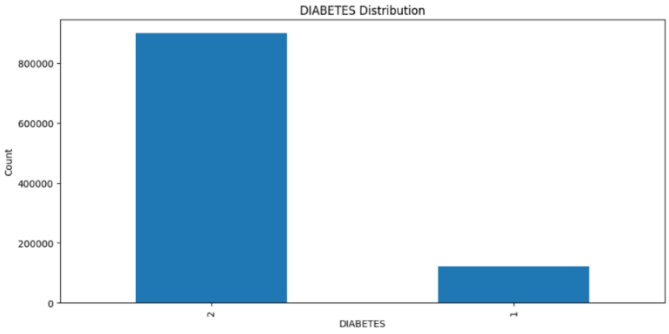


Figure 11. The number of patients with and without diabetes.

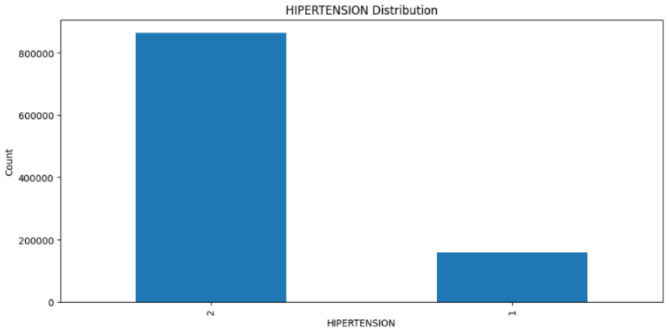


Figure 12. The number of patients with and without hypertension.

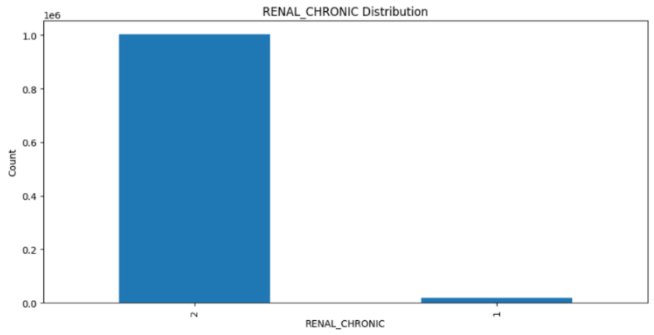


Figure 13. The number of patients with and without chronic renal disease.

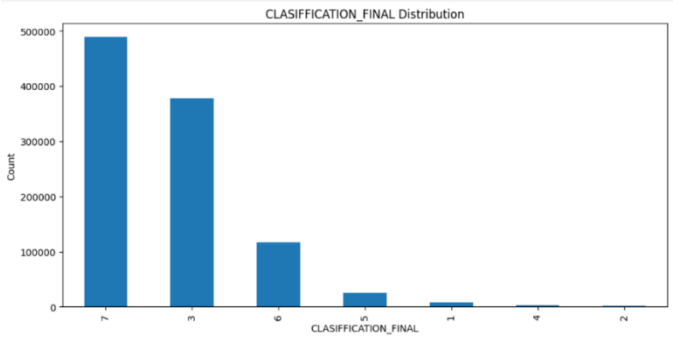


Figure 14. The final COVID-19 severity classification.

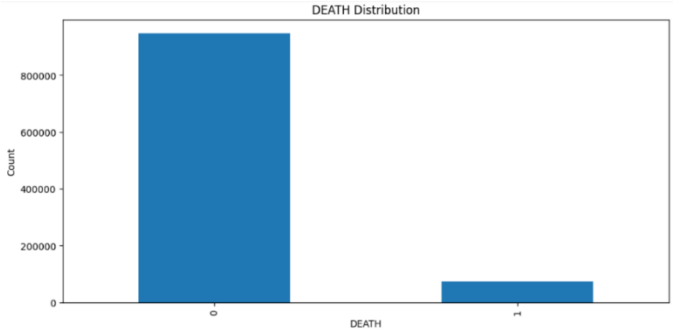


Figure 15. The number of deceased and surviving patients.

The descriptive analysis of the attributes following pre-processing gives a clear picture of the dataset's structure. The statistical measures and visualizations aid in identifying the key patterns and variability of the attributes, ensuring that the dataset is ready for further modeling and analysis.

D. Data Modelling

Modeling the data involves finding suitable machine-learning models, training them on the prepared dataset, and testing their performance. The purpose is to determine which model best predicts COVID-19 death rates. This study used three machine learning models: logistic regression, random forest, and a neural network.

To evaluate the robustness and generalizability of our models, we tested two different data splits: 70-30 and 80-20. The 70-30 split has a larger test set, providing a more robust assessment of model performance, whereas the 80-20 split allows for more training data, potentially enhancing model learning.

Logistic Regression: A linear model for binary classification tasks. It predicts the likelihood that a given input belongs to a specific class. The model is basic, easy to understand, and works well on linearly separable datasets.

Random Forest: Random Forest is an ensemble learning method that builds several decision trees during training and returns the mode of the classes for classification tasks. It is resistant to overfitting and can handle huge datasets of higher dimensionality.

Neural Network (MLP): We tested two alternative Multi-Layer Perceptron (MLP) models with different data splits. For the 70-30 split, we used the Scikit-Learn MLPClassifier, which is efficient and simple to use when creating MLPs with default settings. We used a Keras Sequential model for the 80-20 split, which has two hidden layers, dropout for regularization, and early stopping to avoid overfitting.

Before training our models, we examined the target variable 'DEAD' and discovered that it was imbalanced, with much more 'Alive' than 'Dead' instances. To solve this, we utilized the Synthetic Minority Over-sampling Technique (SMOTE) to balance the classes in the training data. SMOTE generates synthetic samples for the minority class, which results in a more balanced dataset. This step is critical for increasing our models' performance and reliability. The figures below show the distribution of target variables before and after handling the imbalance.

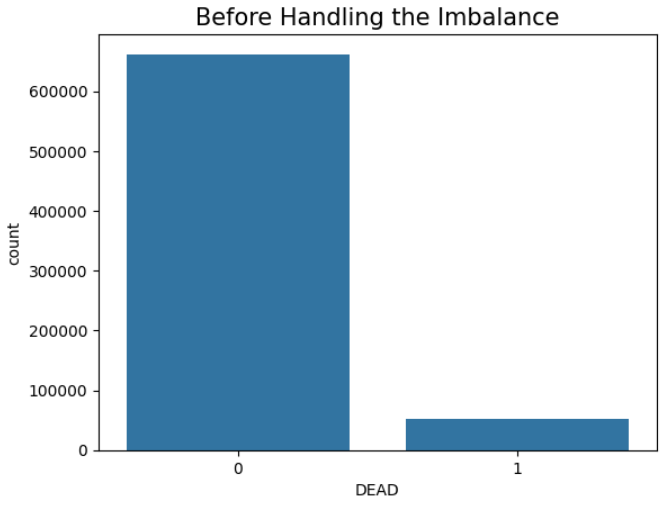


Figure 16. Distribution of target variables before handling the imbalance.

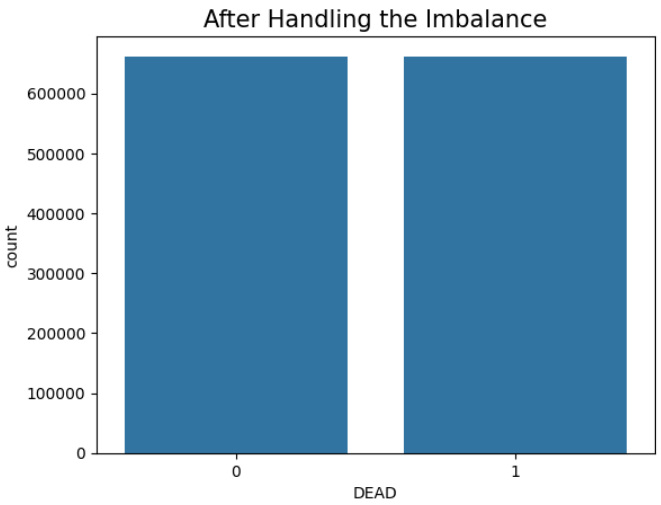


Figure 17. Distribution of target variables after handling the imbalance.

E. 70-30 Split

Initially, 70-30 split was applied, with 70% of the data used for training and 30% for testing. This split ratio provides a large test set, guaranteeing that the models are evaluated on a significant amount of data.

After training the models, their performance was assessed using a variety of measures, including accuracy, confusion matrix, classification report, ROC curve, and AUC score.

The Logistic Regression model had a training accuracy of 0.906 and a testing accuracy of 0.900. The performance will be evaluated using the accuracy, confusion matrix, ROC curve, and classification report.

Table 1. Classification report for logistic regression.

Class	Precision	Recall	F1-Score	Support
Alive	0.99	0.90	0.94	284,090
Dead	0.41	0.91	0.56	21,752
Accuracy	-	-	0.90	305,842
Macro Avg	0.70	0.91	0.75	305,842
Weighted Avg	0.95	0.90	0.92	305,842

The Random Forest model scored 0.953 training accuracy and 0.916 test accuracy. Its robustness and capacity to handle huge datasets make it an appropriate choice for this task. Performance metrics, such as accuracy, confusion matrix, ROC curve, and classification report, will be provided.

Table 2. Classification report for random forest.

Class	Precision	Recall	F1-Score	Support
Alive	0.98	0.93	0.95	284,090
Dead	0.45	0.79	0.57	21,752
Accuracy	-	-	0.92	305,842
Macro Avg	0.72	0.86	0.76	305,842
Weighted Avg	0.95	0.92	0.93	305,842

The Neural Network model (MLP) had a training accuracy of 0.929 and a test accuracy of 0.9001. It had the highest AUC score of 0.964, making it the top-performing model of the three. Detailed performance measures, such as accuracy, confusion matrix, and ROC curve, will be shown.

Table 3. Classification report for MLP.

Class	Precision	Recall	F1-Score	Support
Alive	0.99	0.90	0.94	284,090
Dead	0.41	0.91	0.57	21,752
Accuracy	-	-	0.90	305,842
Macro Avg	0.70	0.91	0.76	305,842
Weighted Avg	0.95	0.90	0.92	305,842

The AUC scores for the various models were as follows: Logistic Regression achieved an AUC score of 0.960, Random Forest achieved 0.940 and the Neural Network (MLP) achieved the highest score of 0.964. Consequently, the best model was identified as the Neural Network (MLP) with an AUC score of 0.964. Each model was evaluated on the first five rows of the test dataset to verify test consistency. The outcomes will be demonstrated through full classification reports and accuracy metrics for each model.

F. 80-20 Split

Next, a 80-20 split was applied, with 80% of the data used for training and 20% for testing. This split ratio provides a large test set, guaranteeing that the models are evaluated on a significant amount of data. After training the models, the performance was evaluated using a variety of measures, including accuracy, confusion matrix, classification report, ROC curve, and AUC score.

The Logistic Regression model had a testing accuracy of 0.940. The performance will be evaluated using the accuracy, confusion matrix, ROC curve, and classification report. As we see, although the accuracy score is the high value which is 94%, it's an imbalance data, the precision value, recall and f1-score of the dead patients are a low value which are 61%, 45% and 52%. To solve the problem, we used undersampling for this case since we have too many patients. Undersampling is a technique to balance uneven datasets by keeping all the data in the minority class and decreasing the size of the majority class. After undersampling, the test accuracy score for logistic regression become 0.900.

Table 4. Classification report for logistic regression.

Class	Precision	Recall	F1-Score	Support
Alive	0.96	0.98	0.97	189,597
Dead	0.61	0.45	0.52	14,799
Accuracy	-	-	0.94	204,396
Macro Avg	0.78	0.71	0.74	204,396
Weighted Avg	0.93	0.94	0.94	204,396

Table 5. Classification report for logistic regression after undersampling.

Class	Precision	Recall	F1-Score	Support
Alive	0.92	0.89	0.90	14,889
Dead	0.89	0.92	0.91	14,974
Accuracy	-	-	0.90	29,863
Macro Avg	0.90	0.90	0.90	29,863
Weighted Avg	0.90	0.90	0.90	29,863

The Random Forest model test accuracy scored is 0.900. Its robustness and capacity to handle huge datasets make it an appropriate choice for this task. Performance metrics, such as accuracy, confusion matrix, ROC curve, and classification report, will be provided.

Table 6. Classification report for random forest.

Class	Precision	Recall	F1-Score	Support
Alive	0.93	0.87	0.90	14,889
Dead	0.88	0.93	0.90	14,974
Accuracy	-	-	0.90	29,863
Macro Avg	0.90	0.90	0.90	29,863
Weighted Avg	0.90	0.90	0.90	29,863

The Neural Network model (MLP) had a test accuracy of 0.9103. It had the highest AUC score of 0.96, making it the top-performing model of the three. Detailed performance measures, such as accuracy, confusion matrix, and ROC curve, will be shown.

Table 7. Classification report for Keras sequential model.

Class	Precision	Recall	F1-Score	Support
Alive	0.95	0.86	0.91	14,889
Dead	0.88	0.96	0.91	14,974
Accuracy	-	-	0.91	29,863
Macro Avg	0.91	0.91	0.91	29,863
Weighted Avg	0.91	0.91	0.91	29,863

The AUC scores for the various models were as follows: Logistic Regression achieved an AUC score of 0.95, Random Forest achieved 0.94 and the Neural Network (MLP) achieved the highest score of 0.96. Consequently, the best model was identified as the Neural Network (MLP) with an AUC score of 0.96.

G. Data Visualisation

The main objective of data visualization is to allow the reader to quickly process the data, which includes potential trends, relationships, and more. As a result, images are treated very seriously, and they are optimized for maximum effectiveness. To effectively present the outcomes of the model evaluations, the following charts and figures will be displayed.

- Confusion matrices for all models (70-30 split).

These will show the performance of each model in terms of true positive, true negative, false positive, and false negative predictions showed in the 70-30 Split in Figure 18(a, b, c).

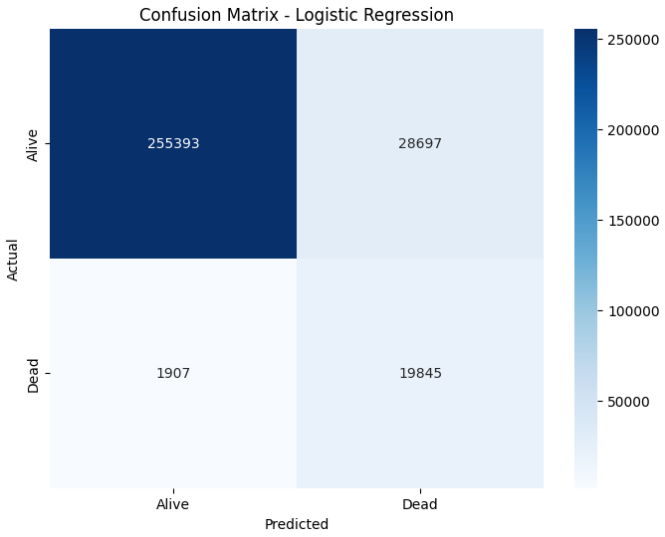


Figure 18.a. Confusion matrix for logistic regression.

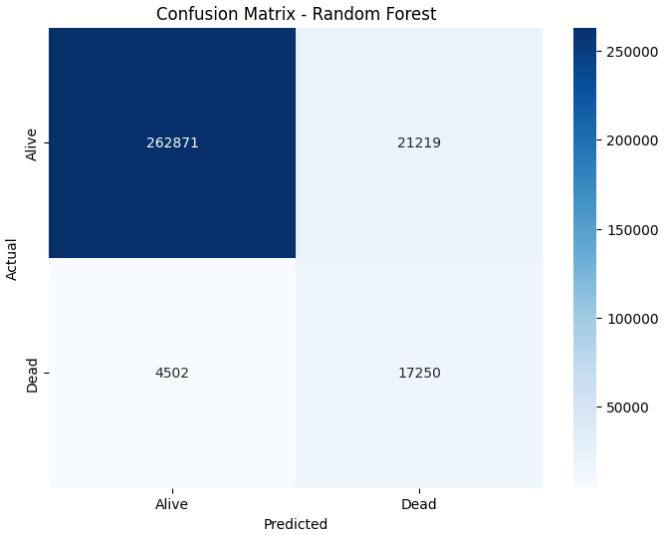


Figure 18.b. Confusion matrix for random forest.

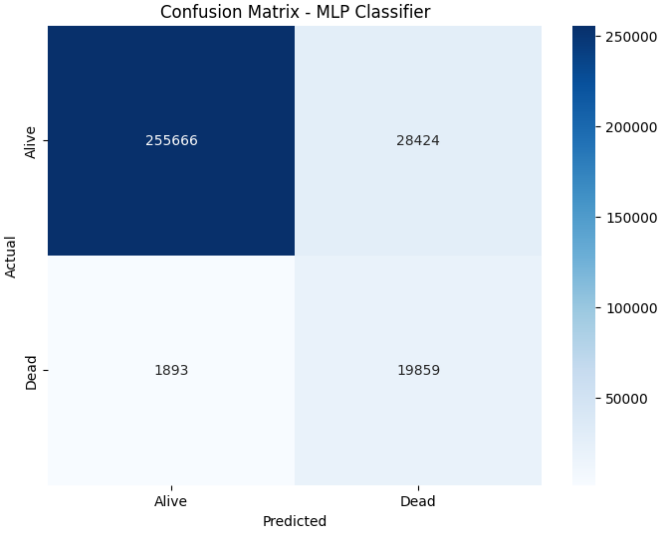


Figure 18.c. Confusion matrix for MLP classifier.

- ROC curves for all models (70-30 split).

These curves will show the trade-off between true and false positive rates for each of the models, and their performance in the 70-30 split showed in Figure 19(a, b, c).

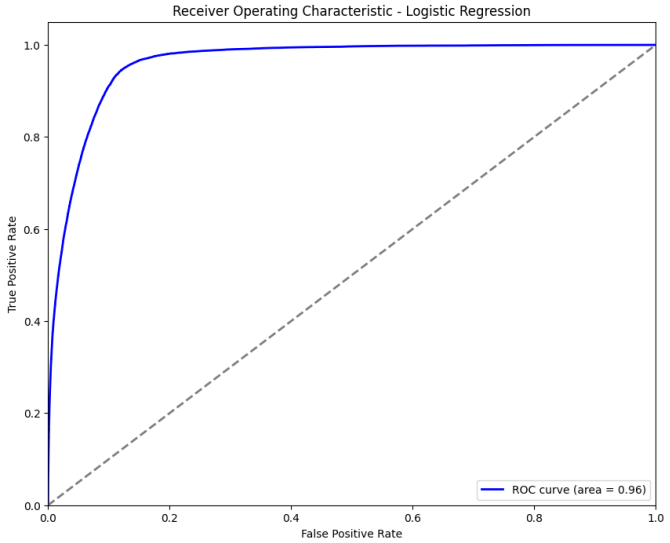


Figure 19.a. ROC for logistic regression.

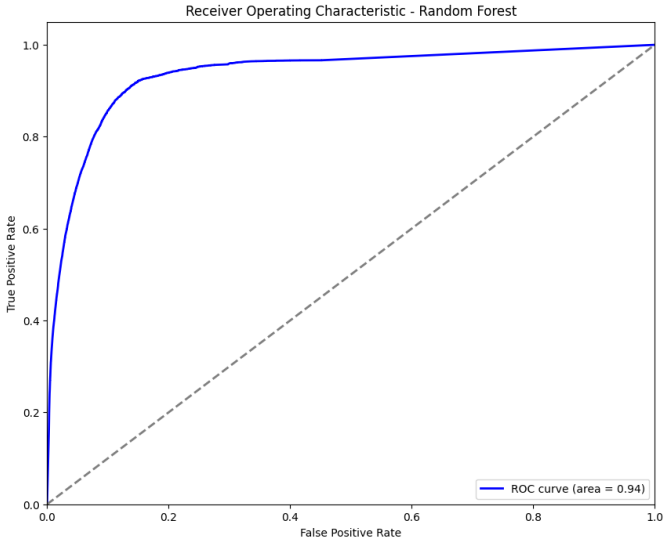


Figure 19.b. ROC for random forest.

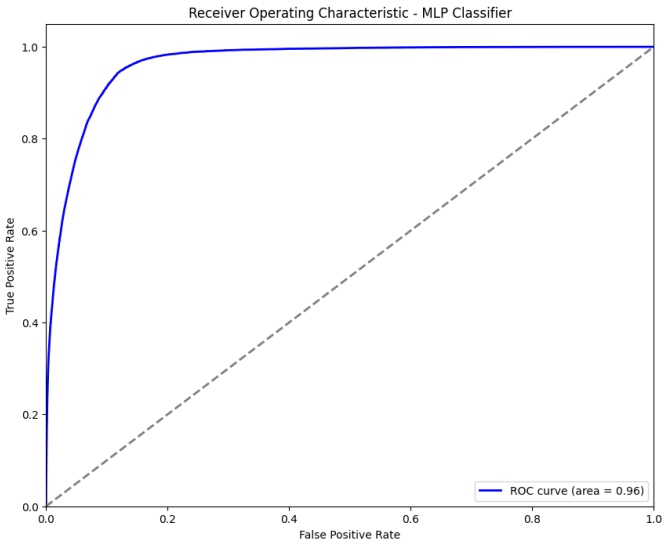


Figure 19.c. ROC for MLP Classifier

- Bar plot showing AUC curves for all models (70-30 split).

This bar plot will compare the AUC scores of the Logistic Regression, Random Forest, and Neural Network models, highlighting the best-performing one in the 70-30 Split showed in Figure 20.

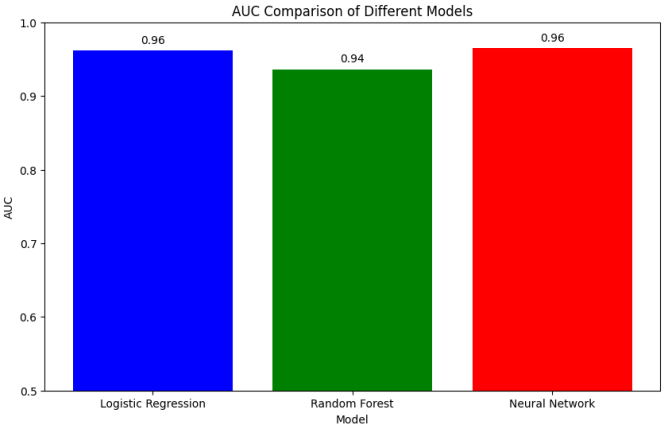


Figure 20. AUC comparison different models.

- Confusion Matrices for all models (80-20 split).

These will show the performance of each model in terms of true positive, true negative, false positive, and false negative predictions in the 80-20 Split showed in Figure 21(a, b, c).

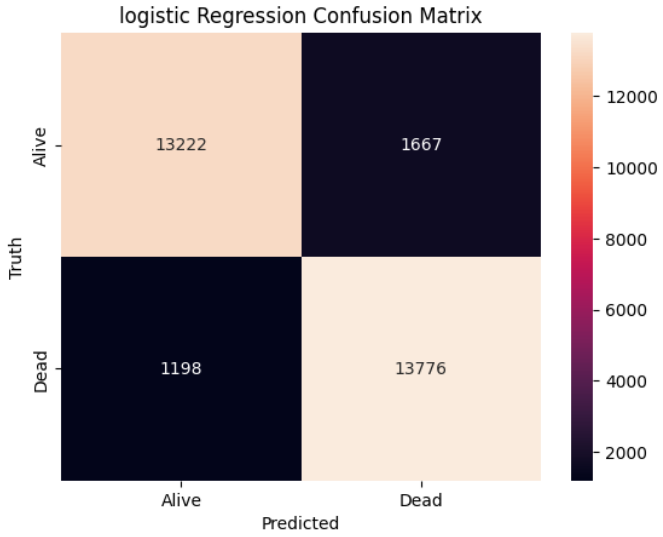


Figure 21.a. Confusion matrix for logistic regression

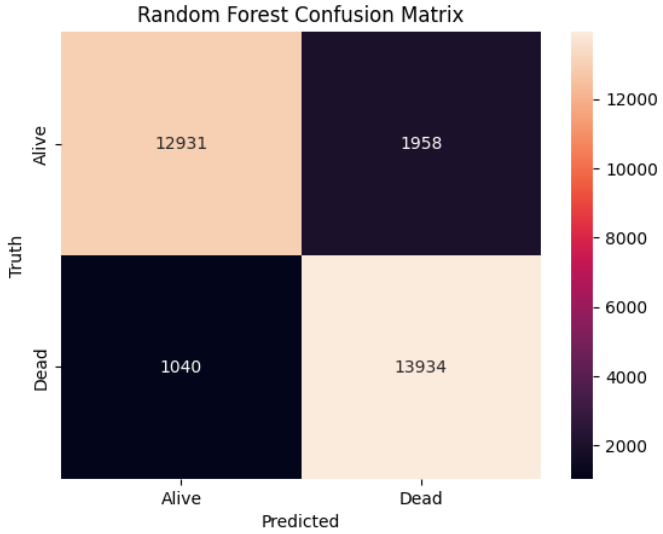


Figure 21.b. Confusion matrix for random forest.

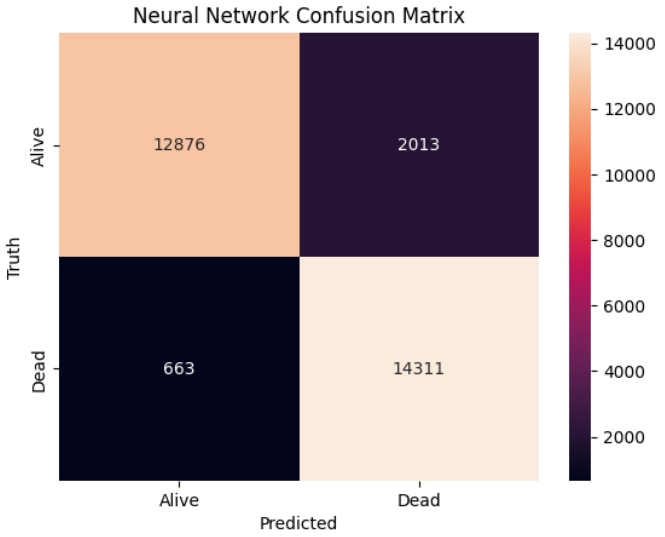


Figure 21.c. Confusion matrix for neural network.

- ROC Curves for all models (80-20 split).

These curves will show the trade-off between true and false positive rates for each of the models, and their performance in the 80-20 Split showed in Figure 22.

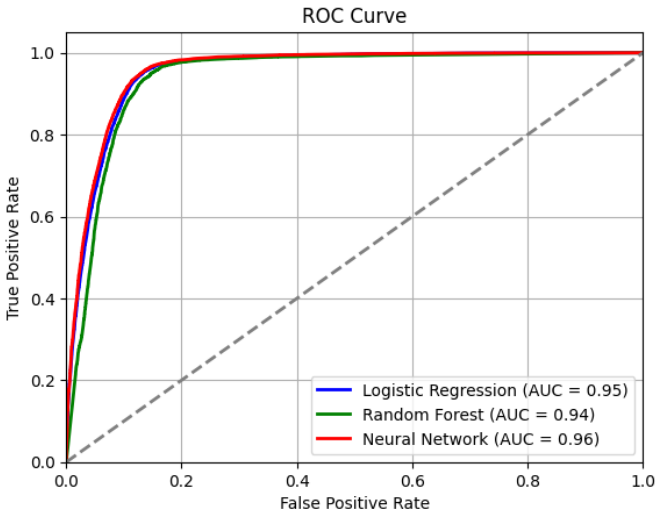
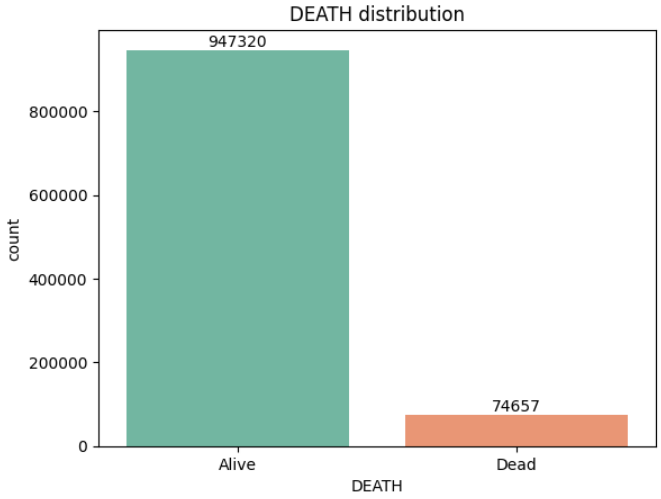
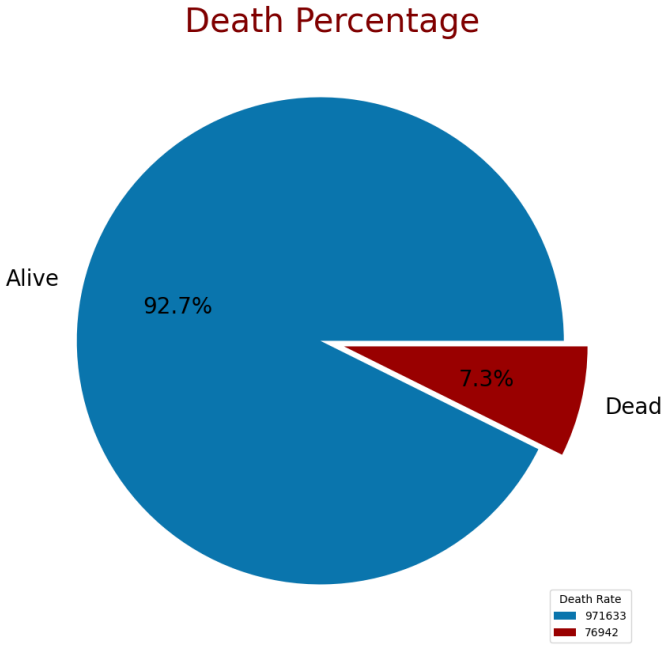


Figure 22. ROC Curves for all models.

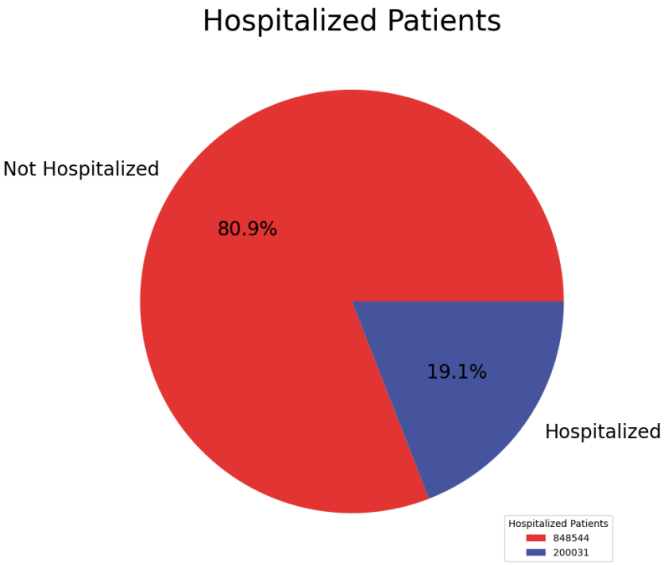
- Data Science question:
  - What is the overall distribution of the 'DEAD' variable in the dataset?



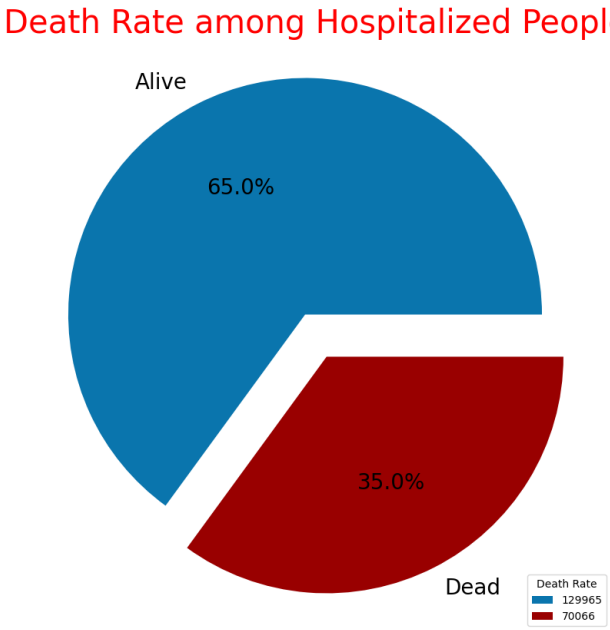
  - What percentage of patients in the dataset are alive vs. dead?



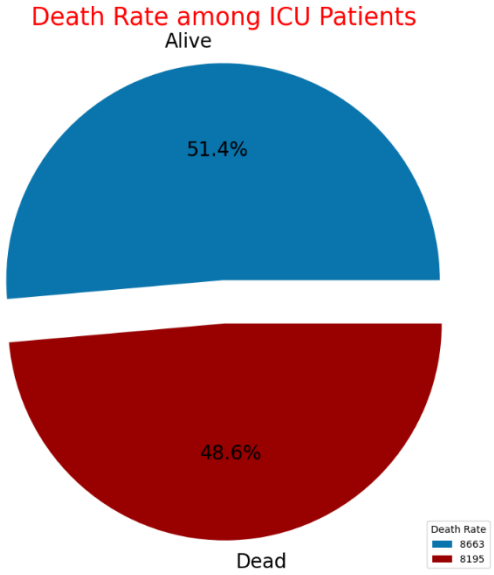
- What is the distribution of hospitalized vs. non-hospitalized patients in the dataset?



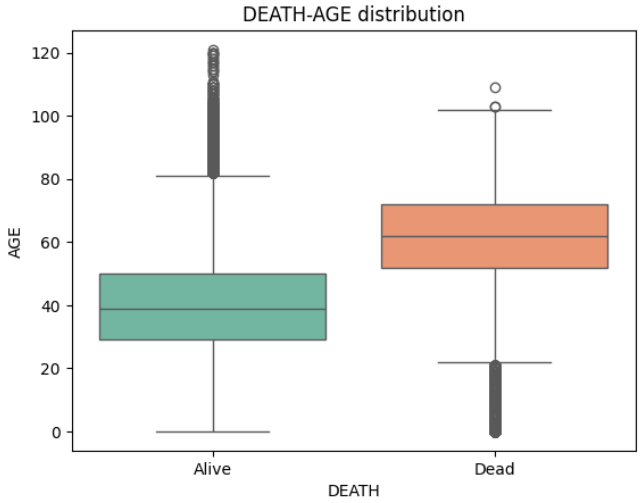
- What is the death rate among hospitalized patients?



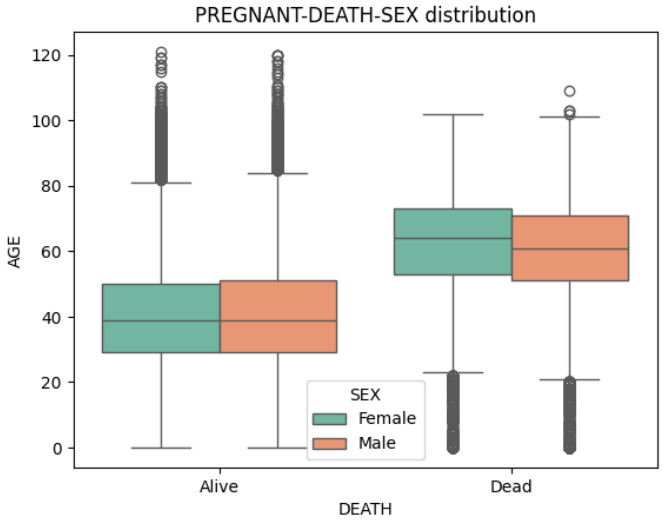
- What is the death rate among ICU patients?



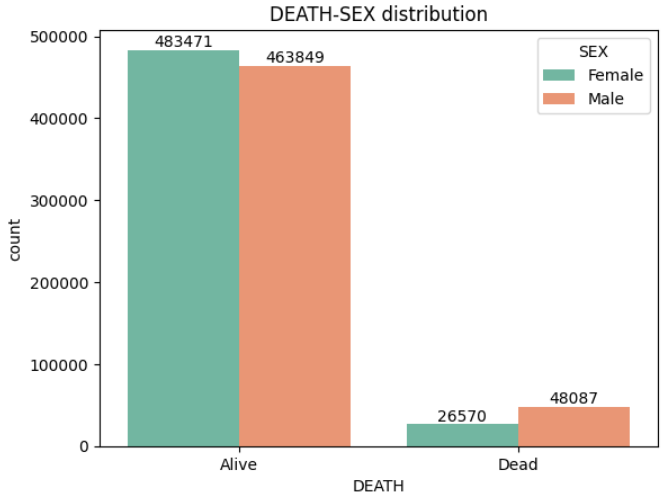
- How does age distribution differ between dead and alive patients?



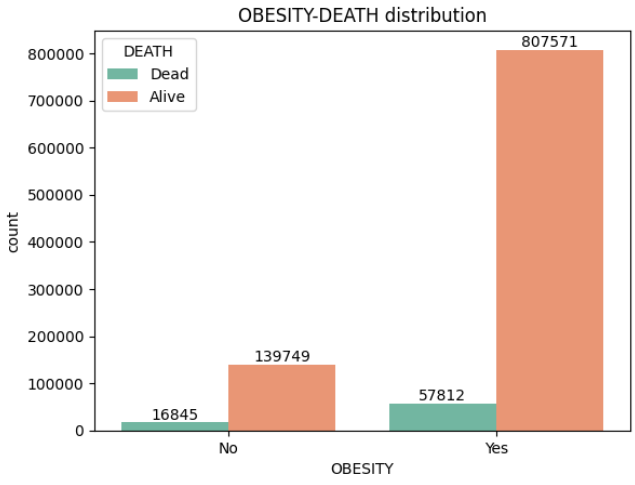
- How does the death rate vary by gender and age?



- What is the gender distribution among dead and alive patients?



- How does obesity relate to the death rate in the dataset?



There are nine key data science questions regarding the COVID-19 dataset. The first is the overall distribution of the 'Dead Variable', analysing the proportion of patients who are alive versus those who are deceased. The second question addresses the percentage of alive versus dead patients, determining the survival rates. The third question explores the distribution of hospitalized versus non-hospitalized patients. The fourth question calculates the death rate among hospitalized patients, while the fifth assesses the mortality rate among ICU patients. The sixth question examines age distribution by survival status, comparing the ages of deceased and surviving patients. The seventh question investigates the death rate by gender and age, examining variations in mortality rates across different demographics. The eighth question looks at the gender distribution among deceased and surviving patients. Lastly, the ninth question explores the relationship between obesity and mortality rate. This comprehensive analysis has provided insights into key patterns and relationships within the dataset, setting the stage for subsequent modelling and predictive analysis.

IV. Result and Discussion

The dataset was first divided using an 80-20 and a 70-30 split into training and testing sets. Because there are differences in the amounts of training and test data available, each model's performance is predicated on these divides. The differences in results and accuracy among the three different models which is Logistic Regression, Random Forest and Neural Network are due to variations in data distribution. Primarily, models like Logistic Regression showed high accuracy due to class imbalance but had low precision and recall for the minority class "dead". To address this imbalance with undersampling improved the performance metrics, particularly for Logistic Regression and Random Forest. The Neural Network model, then with its ability to capture complex patterns, achieved the highest AUC score and accuracy. Evaluation metrics such as precision, recall and F1-score provide a more detailed perspective of each model's effectiveness, highlighting the importance of balanced data handling and appropriate metric selection. In this report we compare three machine learning techniques which is Logistic Regression, Random Forest and Neural Network. In the 70-30 split, 70% of the data is used for training and 30% for testing. Thus, the Logistic Regression achieved a testing accuracy of 90%, Random Forest 91.6% and Neural Network 90.01%, with the Neural Network having the highest AUC score of 0.964. In the 80-20 split, 80% of the data is used for training and 20% for testing. After undersampling to address data imbalance, the Logistic Regression model achieve a testing accuracy of 90%, Random Forest 90%, and Neural Network 91.03%, with the Neural Network again having the highest AUC score of 0.96. Thus, both splits were used with the Neural Network consistently performing the best in terms of AUC score.

In addition, the final algorithm chosen was the Neural Network model, selected for its superior performance metrics. On a 70-30 train-test split, it achieved the highest AUC score

of 0.96 and a test accuracy of 0.9103, outperforming both Logistic Regression with AUC score of 0.95, and test accuracy of 0.940 and Random Forest with AUC score of 0.94 and test accuracy of 0.916. On an 80-20 split, the Neural Network also excelled, demonstrating robust performance across various metrics. The model's ability to effectively handle complex data patterns and generalize well to unseen data made it most suitable choice for predicting COVID-19 mortality in this study. Based on the Neural Network model's superior performance in predicting COVID-19 mortality, it achieved the highest AUC score of 0.96 and a test accuracy of 0.9103 in the 70-30 train-test split, indicating robust predictive capability. The model's precision, recall and F1-scores across different classes were also impressive, reflecting its effectiveness in distinguishing between alive and decease patients. These results are justified by the Neural Network's capacity to capture more complex patterns and interaction within the dataset, which includes different features such as demographic information, pre-existing conditions and symptoms. The Neural Network model was selected as the final model because it often outperformed Random Forest and Logistic Regression in key performance criteria. The result mentioned earlier highlights the capability of complex machine learning methods to improve prediction accuracy and offer significant perspectives for healthcare decision-making within an outbreak.

V. Conclusion

This study demonstrates the potential of machine learning models in predicting COVID-19 mortality rates, with the Neural Network model standing out as the best performer. The Neural Network achieved the highest AUC score of 0.96 and a test accuracy 0.9103, indicating its strong ability to identify patterns in the data, which includes demographic information, pre-existing conditions, and symptoms.

The main advantage of using the Neural Network model is its high accuracy and reliability, which are crucial for making precise predictions. This can help healthcare providers prioritize machine learning models can analyse large datasets and reveal important insights that can improve healthcare strategies and decision-making. However, the limitations in this study is the dataset had an imbalance in the classes, which may affect the model's performance. While undersampling was used to address this, other methods like synthetic data generation could be explored to improve results. Also, the study focuses on short-term mortality predictions and does not consider long-term outcomes or the effects of new virus variants.

The use of machine learning models for predicting COVID-19 mortality rates can have several applications. These models can be integrated into clinical decision support systems to help healthcare professionals identify high-risk patients and optimize resources. The insights from these models can also guide public health policies and interventions. Future research could expand this work by predicting long-term COVID-19 outcomes, using real-time data, and adapting models to new virus variants.

In summary, while there are challenges to address, machine learning models like Neural Networks show great promise for improving predictive accuracy and supporting crucial healthcare decisions during COVID-19 and beyond. This study lays the groundwork for further research and refinement of predictive models to enhance healthcare delivery.

Acknowledgment

We would like to extend our heartfelt gratitude to everyone who has supported and guided us throughout the journey of completing this group project. First and foremost, we wish to acknowledge the invaluable contributions of our team members:

- Afiefah Zulkifli: 24, from Malaysia, a fresh graduate with a Bachelor of Food Technology from Universiti Putra Malaysia (UPM). Her dedication and fresh perspective were instrumental in our project's success.
- Safia Adrina: 25, from Malaysia, who previously worked as a data analyst at Doctor On Call and holds a Bachelor of Computer Science from Universiti Malaysia Pahang (UMP). Her analytical skills and practical experience greatly enhanced our work.
- Mohamad Moubarak: 24, from Comoros Island, a fresh graduate with a Bachelor of Computer Science from Universiti Islam Antarabangsa (UIA). His innovative ideas and enthusiasm were vital to our progress.
- Xujian Hong: 29, from China, who previously worked as an Android Developer. His technical expertise and

problem-solving abilities significantly contributed to our project.

We all met and became friends while pursuing our Master's in Data Science at Universiti Kebangsaan Malaysia. This experience has not only enriched our academic knowledge but also strengthened our bonds. We hope to pass this Master's degree together and achieve our best dreams in the future. Lastly, we express our sincere appreciation to our professors, mentors, and families for their unwavering support and encouragement. Without their guidance, this project would not have been possible. Thank you all for being a part of this journey.

REFERENCES

[1] Bown, C. P. (2021). How COVID-19 medical supply shortages led to extraordinary trade and industrial policy. *Asian Economic Policy Review*, *17*(1), 114–135. https://doi.org/10.1111/aepr.12359c

[2] Aslam, H., & Biswas, S. (2023). Analysis of COVID-19 death cases using machine learning. *SN Computer Science/SN Computer Science*, *4*(4). https://doi.org/10.1007/s42979-023-01835-9

[3] Pourhomayoun, M., & Shakibi, M. (2021). Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health*, *20*, 100178. https://doi.org/10.1016/j.smhl.2020.100178

[4] Yadaw, A. S., Li, Y., Bose, S., Iyengar, R., Bunyavanich, S., & Pandey, G. (2020). Clinical predictors of COVID-19 mortality. *medRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2020.05.19.20103036

[5] Zakariaee, S.S., Naderi N., Ebrahimi, M., & Kazemi-Arpanahi, H. (2023). Comparing machine learning algorithms to predict COVID-19 mortality using a dataset including chest computed tomography severity score data. *Sci Rep* 13, 11343 (2023). **https://www.nature.com/articles/s41598-023-38133-6**

[6] Jamshidi E, Asgary A, Tavakoli N, Zali A, Setareh S, Esmaily H, Jamaldini SH, Daaee A, Babajani A, Sendani Kashi MA, Jamshidi M, Jamal Rahi S, Mansouri N. (2022). Using Machine Learning to Predict Mortality for COVID-19 Patients on Day 0 in the ICU. Front Digit Health. 2022 Jan 13;3:681608. doi:10.3389/fdgth.2021.681608. PMID: 35098205; PMCID: PMC8792458. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8792458/

[7] Emami, H., Rabiei, R., Sohrabei, S., & Atashi, A. (2023) Predicting the mortality of patients with Covid-19: A machine learning approach. https://doi.org/10.1002/hsr2.1162

[8] Bottino F, Tagliente E, Pasquini L, Napoli AD, Lucignani M, Figà-Talamanca L, Napolitano A. COVID Mortality Prediction with Machine Learning Methods: A Systematic Review and Critical Appraisal. *Journal of Personalized Medicine*. 2021; 11(9):893. https://www.mdpi.com/2075-4426/11/9/893

APPENDIX

Table 8. Descriptive Statistics of Each Attribute

Attribute	Count	Mean	Std	Min	25%	50%	75%	Max
USMER	1.021977e+06	1.642009e+00	4.794098e-01	1.000000e+00	1.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
MEDICAL_UNIT	1.021977e+06	8.987361e+00	3.724244e+00	1.000000e+00	4.000000e+00	1.200000e+01	1.200000e+01	1.300000e+01
PATIENT_TYPE	1.021977e+06	1.187472e+00	3.902901e-01	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	2.000000e+00
PNEUMONIA	1.021977e+06	1.865591e+00	3.410915e-01	1.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
AGE	1.021977e+06	4.189535e+01	1.674953e+01	0.000000e+00	3.000000e+01	4.000000e+01	5.300000e+01	1.210000e+02
DIABETES	1.021977e+06	1.880455e+00	3.244286e-01	1.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
HIPERTENSION	1.021977e+06	1.844163e+00	3.627011e-01	1.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
RENAL_CHRONIC	1.021977e+06	1.982074e+00	1.326828e-01	1.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
CLASIFFICATION_FINAL	1.021977e+06	5.291434e+00	1.889546e+00	1.000000e+00	3.000000e+00	6.000000e+00	7.000000e+00	7.000000e+00
DEAD	1.021977e+06	7.305155e-02	2.602212e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00

Table 9. Data Description of the Dataset

Attribute	Description	Data Type
SEX	1 for female and 2 for male	Nominal
AGE	Age of the patient	Continuous
CLASSIFICATION	Covid test findings. Values 1-3 mean that the patient was diagnosed with covid in different degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive.	Ordinal
PATIENT TYPE	Type of care the patient received in the unit. 1 for returned home and 2 for hospitalization.	Nominal
PNEUMONIA	whether the patient already have air sacs inflammation or not.	Binary
PREGNANCY	whether the patient is pregnant or not.	Binary
DIABETES	whether the patient has diabetes or not.	Binary
COPD	Indicates whether the patient has Chronic obstructive pulmonary disease or not.	Binary
ASTHMA	whether the patient has asthma or not.	Binary
INMSUPR	whether the patient is immunosuppressed or not.	Binary
HYPERTENSION	whether the patient has hypertension or not.	Binary
CARDIOVASCULAR	whether the patient has heart or blood vessels related disease.	Binary
RENAL CHRONIC	whether the patient has chronic renal disease or not.	Binary
OTHER DISEASE	whether the patient has other disease or not.	Binary
OBESITY	whether the patient is obese or not.	Binary
TOBACCO	whether the patient is a tobacco user.	Binary

USMR	Indicates whether the patient treated medical units of the first, second or third level.	Nominal
MEDICAL UNIT	type of institution of the National Health System that provided the care.	Nominal
INTUBED	whether the patient was connected to the ventilator.	Binary
ICU	Indicates whether the patient had been admitted to an Intensive Care Unit.	Binary
DATE DIED	If the patient died indicate the date of death, and 9999-99-99 otherwise.	Date
DEAD	1 if the patient died and 0 if the patient is alive	Binary