Machine Learning and Viral Pandemics: COVID 19

Noemie Lamontagne

June 2023

1 Abstract

Viral pandemics will be a recurring issue in the future. Machine Learning may have a place in managing these pandemics through clinical aid or public awareness. A predictive model could provide efficient patient management or could be used as a public information tool for awareness. However, predictive models to be used in a clinical setting should be highly accurate, and held to a higher standard than those to be used as a non-diagnostic information tool for the public. The objective of our research is to develop a predictive model as a public information tool for awareness. This predictive model could come in the form of a website or application.

Two datasets with COVID 19 patient information including patient preconditions and instances of death were combined to form a larger dataset with more samples and a larger variety of preconditions. Logistic Regression (LR), Neural Network (NN), XGBoost, and RandomForest models were trained with and without Synthetic Minority Oversampling Technique (SMOTE) as an attempt to compensate for a class imbalance. The models were then compared using confusion matrices to obtain the number of true positives, false positives, true negatives and false negatives. Using these values, the accuracy, precision, recall, specificity and F1 score of each model were calculated and compared to find the most accurate predictive model.

According to the F1 score, the Logistic Regression model with SMOTE was found to be the most accurate model while the Neural Network model was the least accurate. The LR model was quickly implemented into an experimental online public information tool using Gradio.

In future pandemics, with more balanced and widely available clinical data, similar and finer information tools could be created to inform the public.

2 Introduction

Viral pandemics occur periodically and pandemics such as Covid-19 will arise once again in the future. Machine Learning can help manage pandemics through patient management and public awareness.

A predictive model could provide efficient patient management. Pandemics often result in a surge of patients who need urgent care. Therefore, medical personnel, resources and hospital occupancy need to be used wisely. An accurate predictive model could help predict which patients need the most urgent care. This would save medical personnel time during triage and redirect medical resources to those who need them the most ahead of time. However, it is expected that a predictive model must be highly accurate in order to be used in a clinical setting.

A predictive model could increase public awareness. During pandemics there is confusion or doubt about the severity of the situation. An objective view on an individual's potential situation

if they were to contract the virus could help clear up the confusion, relieve some people of stress or inform people of the high risk they are taking or subjecting to other people. Predictive models could be created for and used by the public to help inform them of their personal mortality risk when associated with the virus or disease. This predictive model could come in the form of a website or application. The public could enter any of their preconditions and the model would return their risk of mortality with a percentage error if they were to hypothetically contract the virus. This predictive model would not be used in a clinical setting, its predictions would not be diagnostic and would not need to achieve clinical accuracy standards.

The objective of our research is to develop a predictive model as a public information tool for awareness. To achieve the objective, two types of predictive models were trained on the same dataset and compared. The main dataset was created by combining two datasets that contained clinical covid patient information such as their preconditions and death or survival. Inside the main dataset, certain columns were omitted and the data was normalized. The training data was then enhanced using SMOTE and was used to train Logistic Regression models and Neural Network (NN) models. The accuracy of Logistic Regression models and the NN models were then compared using confusion matrices.

The data within the training dataset shows an unbalanced amount of negative and positive values within the Death column, the positive Death values being under-represented. To avoid biased prediction and decreased accuracy, some predictive models will be enhanced using SMOTE which will artificially increase the data for a column that is under-represented.

3 Dataset

The main dataset was created by combining two separate datasets which contained similar clinical information. Both Datasets were acquired from Kaggle. They will henceforth be referred to as Kaggle Dataset 1 and Kaggle Dataset 2 [1][2].

3.1 Kaggle Dataset 1[1]

The entire dataset from kaggle had 1048575 original samples with 21 columns. The 21 columns were 'USMER', 'MEDICAL UNIT', 'SEX', 'PATIENT TYPE', 'DATE DIED', 'INTUBED', 'PNEUMONIA', 'AGE', 'PREGNANT', 'DIABETES', 'COPD', 'ASTHMA', 'INMSUPR', 'HIPERTENSION', 'OTHER DISEASE', 'CARDIOVASCULAR', 'OBESITY', 'RENAL CHRONIC', 'TOBACCO', 'CLASIFFICATION FINAL', 'ICU'.

The columns ICU, PATIENT TYPE, INTUBED, PREGNANT, USMR, OTHER DISEASE, CLASSIFICATION FINAL, TOBACCO, and MEDICAL UNIT were excluded. Missing data values were replaced with NaN values. The kaggle dataset then had a remaining 1048575 samples and 12 columns.

The column titles ASTHMA, DIABETES, INMSUPR, CARDIOVASCULAR, RENAL CHRONIC, and DATE DIED were changed respectively to Asthma, Diabetes, Immunodeficiency, Heart disease, Kidney disease, and Death.

Within the column Death, values were changed so that 1 indicates patient death and 0 indicates patient recovery. Within column SEX, values were changed so that 1 indicates female and 0 indicates male. Throughout the rest of the dataset, 1 indicates the presence of a precondition and 0 indicates the lack of a precondition.

The purpose of the manipulations above was to resolve any issues or inconsistencies present in the datasets, in order to align their structures and facilitate a smooth merging process.

3.2 Kaggle Dataset 2[2]

The entire Kaggle Dataset 2 had 319 samples with 39 columns. The 39 columns were 'CBC/CRP', 'Diabetes', 'Asthma', 'Heart disease', 'kidney disease', 'Respiratory disease', 'Cancer', 'Corticosteroids', 'HEM', ' Immunodeficiency', 'Liver disease', 'Rheumatological disease', ' Chest pain', 'Fever', 'Trembling or Shakes', 'Weakness', 'Sweating', 'Sore throat', 'dyspnea', 'Dry cough', 'Cough with sputum', 'Fatigue, whole body hurts', 'Anosmia', 'Ageusia', 'Anorexie', 'Eczema', 'Vertigo', 'Nausea/Diarrhea', and 'Death'.

The columns 'Traveling in past 3 months ago', 'Connection with a suspected (covid-19) person', 'The Infected person (covid-19) in family', 'blood pressure', 'Chest pain: Diagnosis of stroke or heart disease', 'Tobacco', ' transplant', 'Conjunctivitis (Pink eye)', 'Blindness and Tunnel vision', and 'HIV' were excluded. The kaggle dataset then had a remaining 319 samples and 29 columns.

3.3 Combined/Main Dataset

The combined dataset has 1048894 samples with 35 columns. It was created by concatenating the kaggle and research paper datasets. The combined dataset has unknown or NaN values from the kaggle datasets as a result of the concatenation. These unknown values were replaced by the mean value of their respective column.

The 35 columns are 'CBC/CRP', 'Diabetes', 'Asthma', 'Heart disease', 'kidney disease', 'Respiratory disease', 'Cancer', 'Corticosteroids', 'HEM', ' Immunodeficiency', 'Liver disease', 'Rheumatological disease', ' Chest pain', 'Fever', 'Trembling or Shakes', 'Weakness', 'Sweating', 'Sore throat', 'dyspnea', 'Dry cough', 'Cough with sputum', 'Fatigue, whole body hurts', 'Anosmia', 'Ageusia', 'Anorexie', 'Eczema', 'Vertigo', 'Nausea/Diarrhea', 'Death', 'AGE', 'SEX', 'PNEUMONIA', 'COPD', 'HIPERTENSION', and 'OBESITY'. The capital and lower-case column titles group the columns by dataset of origin. The capitalized column titles originally belonged to Kaggle dataset 1 and the lower-case column titles originally belonged to Kaggle dataset 2.

3.3.1 Combined Dataset Visualizations

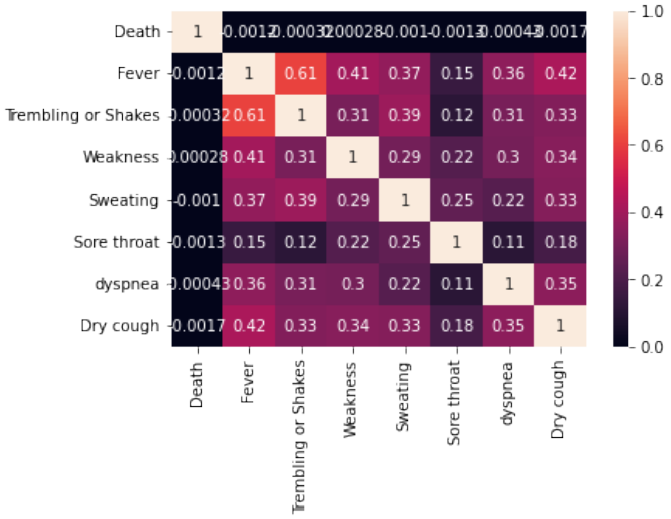

Figure 1: Correlation heatmap comparing the correlation between columns. Specifically between the correlation between the column 'Death' and the other precondition columns. Used in omitted column selection.

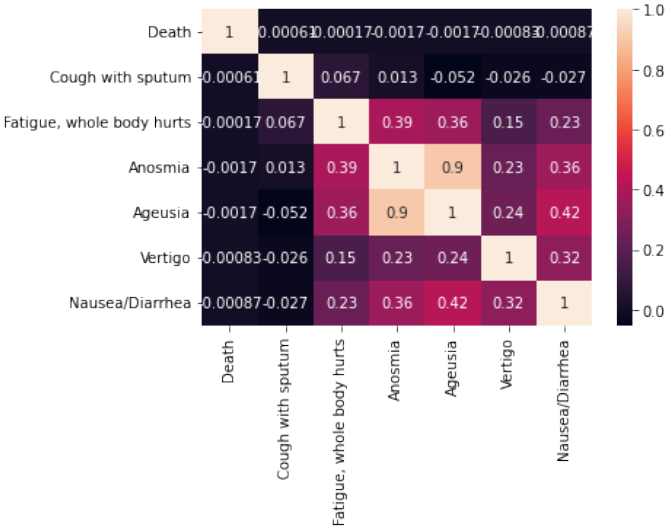

Figure 2: Correlation heatmap comparing the correlation between columns. Specifically between the correlation between the column 'Death' and the other precondition columns. Used in omitted column selection.

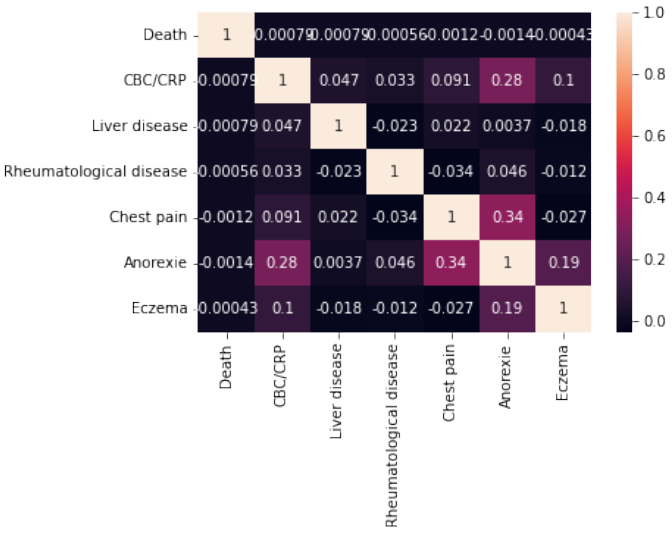

Figure 3: Correlation heatmap comparing the correlation between columns. Specifically between the correlation between the column 'Death' and the other precondition columns. Used in omitted column selection.

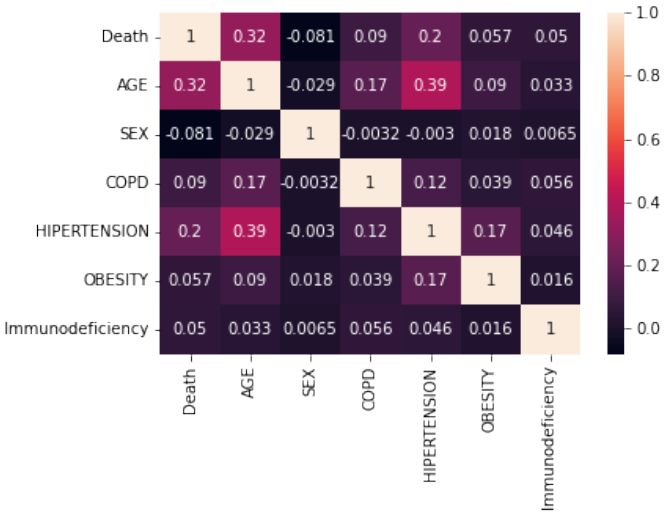

Figure 4: Correlation heatmap comparing the correlation between columns. Specifically between the correlation between the column 'Death' and the other precondition columns. Used in omitted column selection.

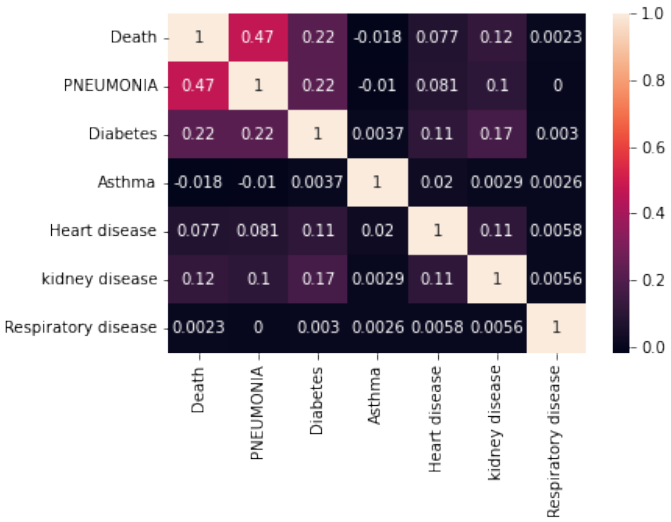

Figure 5: Correlation heatmap comparing the correlation between columns. Specifically between the correlation between the column 'Death' and the other precondition columns. Used in omitted column selection.

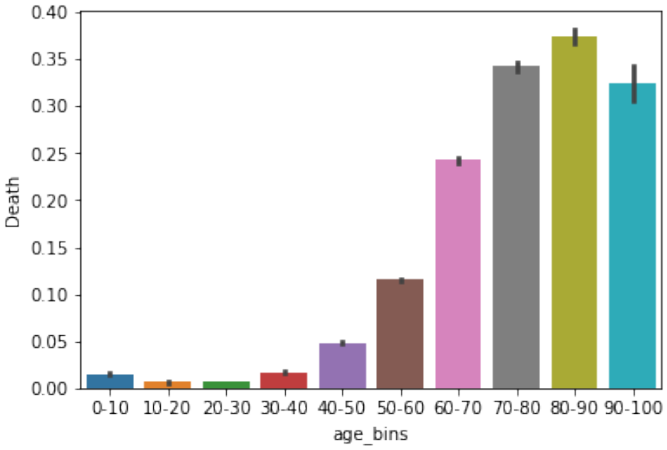

Figure 6: Barplot comparing AGE to Death. As age increases the prevalence of death increases.

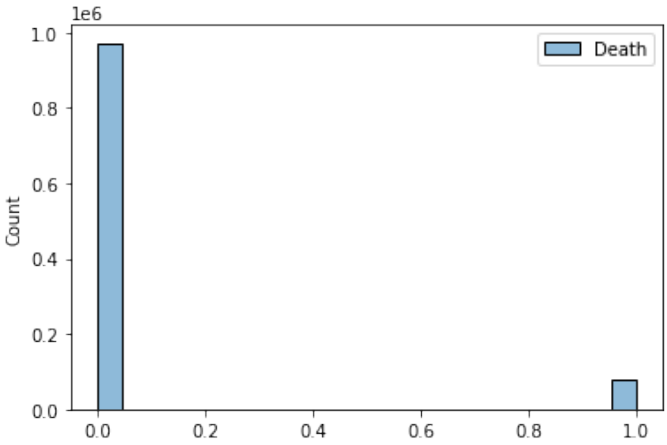

Figure 7: Death histogram counting the amount of negative, and positive Death values. There is a much higher count of negative death values within the dataset which will cause a bias during training. This will decrease the accuracy of positive value predictions.

In all, the following columns from the original datasets were removed:

ICU, PATIENT TYPE, INTUBED, PREGNANT, USMR, OTHER DISEASE, CLASSIFICATION FINAL, TOBACCO, MEDICAL UNIT, Traveling in past 3 months ago, Connection with a suspected (covid-19) person, The Infected person (covid-19) in family, blood pressure, Chest pain: Diagnosis of stroke or heart disease, Tobacco, transplant, Conjunctivitis (Pink eye), Blindness and Tunnel vision, and HIV.

The ICU column indicates whether the patient had been admitted to an Intensive Care Unit. The PATIENT TYPE column indicates the type of care the patient received in the unit. The INTUBED column indicates whether the patient was connected to the ventilator. The USMR column indicates whether the patient treated medical units of the first, second or third level. The CLASSIFICATION FINAL column indicates 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. The MEDICAL UNIT column indicates the type of institution of the National Health System that provided the care. The above described columns were excluded because the predictive model to be trained is meant to predict the mortality risk of a person before they reach the hospital.

The TOBACCO column indicates whether the patient is a tobacco user or not. The column 'Traveling in past 3 months ago' indicates whether or not the patient has traveled in the last 3 months. The column 'Connection with a suspected (covid-19) person' indicates a patient's contact with a suspected infected person. The column 'The Infected person (covid 19) in family' indicates the presence of an infected person in a patient's close family. The column 'blood pressure' indicates a patient's blood pressure. The column 'Chest pain: Diagnosis of stroke or heart disease' indicates whether the patient was diagnosed with stroke or heart disease. The column 'Tobacco' indicates whether the patient was a tobacco user. The column 'transplant' indicates whether or not the patient underwent a transplant. The column 'Conjunctivitis (Pink eye)' column indicates whether the patient had pink eye. The column 'Blindness and Tunnel vision' indicates whether the patient had tunnel vision or blindness. The column 'HIV' indicated whether the patient was HIV positive. The above described columns were excluded because of a too little amount of positive values (1) or a negative correlation, little correlation or a suspect or an outlying correlation with the column,

Death, according to heat correlation maps.

The PREGNANT column indicates whether the patient is pregnant or not. The PREGNANT column was excluded because of a high amount of NaN and negative (0) values. The OTHER DISEASE column indicates whether the patient has other diseases or not. The OTHER DISEASE column was excluded because it was too vague.

3.4 Training/Testing Split

The main dataset was split so that a fraction of the dataset would be used during training and the other fraction would be used during testing. 80 percent of the main dataset was used to train the predictive models, while 20 percent was used to test the predictive models.

4 Models and Methodology

Logistic Regression (LR) models were trained using SMOTE enhanced data and without. Confusion matrices were used to compare accuracies of the LR models. Neural Network (NN) models with different architectures were trained using SMOTE enhanced data and without. Confusion matrices were used to compare the accuracies of the NN models. XGBoost models were trained using SMOTE enhanced data and without. Confusion matrices were used to compare accuracies of the XGBoost models. RandomForest models were trained using SMOTE enhanced data and without. Confusion matrices were used to compare accuracies of the RandomForest models. The confusion matrices from the LR, NN, XGBoost and RandomForest models trained with and without SMOTE enhanced data were compared to find the predictive model with the highest accuracy.

4.1 NN Model

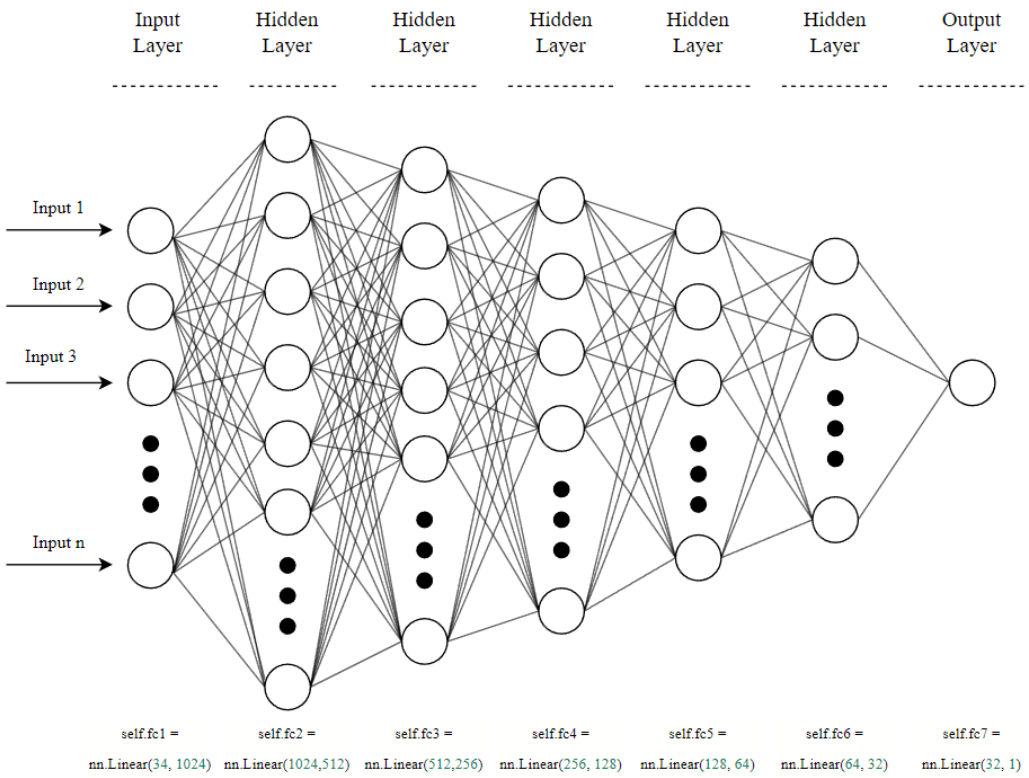

Figure 8: Diagram of NN model architecture

The NN model has 7 fully connected linear layers:

- Input layer:
  - Expects 34 input features
  - Passes on 1024 output features to the next hidden layer
  - ReLU activation
- Hidden layer 1:
  - Expects 1024 features
  - Passes on 512 output features
  - ReLU activation
- Hidden layer 2:
  - Expects 512 features
  - Passes on 256 output features

  - ReLU activation
- Hidden layer 3:
  - Expects 256 features
  - Passes on 128 output features
  - ReLU activation
- Hidden layer 4:
  - Expects 128 features
  - Passes on 64 output features
  - ReLU activation
- Hidden layer 5:
  - Expects 64 features
  - Passes on 32 output features
  - ReLU activation
- Output layer:
  - Expects 32 features
  - Produces 1 feature or output
  - Sigmoid activation: squashes the output into a range between 0 and 1. The output represents the probability of belonging to the positive class.

Dropout is applied after each hidden layer. It randomly drops out a fraction of the neurons during training to prevent overfitting. A dropout rate of 0.2 (20 percent) is used. Dropout is only used during training not evaluation.

4.2 XGBoost model

Table 1: Table of XGBoost parameters

Parameters:	Max Depth	Learning Rate (eta)	Objective	Number of Classes	Number of Rounds
Values:	3	0.3	multi:softprob	2	20

4.3 SMOTE

Synthetic Minority Oversampling Technique (SMOTE) is a statistical technique for increasing the number of cases in your dataset in a balanced way. SMOTE was used to balance the two classes (positive and negative) within the dataset. The SMOTE sampling strategy, 'minority' , creates synthetic samples for the minority class (negative) and was used to try and prevent bias during training.

Confusion matrices were used to compare the NN, LR, RandomForest and XGBoost models.

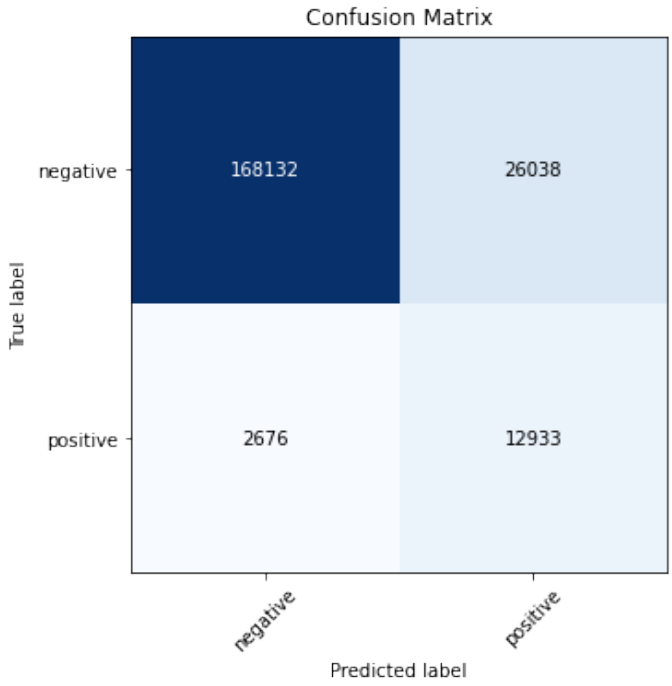

Figure 9: Example of a confusion matrix for the LR model

5 Results and Discussion

The false positives (FP), true positives (TP), false negatives (FN), and true negatives (TN) for each model were obtained from their respective confusion matrix. Those values were then used to calculate the Accuracy, Precision, Recall, Specificity and F1 score for each model.

Table 2: The number of FP, TP, FN and TN predictions from each model

Models	**FP**	**TP**	**FN**	**TN**
LR	3765	5115	10494	190405
LR + SMOTE	26038	12933	2676	168132
NN	13837	953	626	184584
NN + SMOTE	10889	3550	2464	183097
XGBoost	3629	4825	9710	181836
XGBoost + SMOTE	29960	12738	1912	155390
RandomForest	4263	4796	9688	181253
RandomForest + SMOTE	26157	11635	3013	159195

Table 3: Accuracy, Precision, Recall, Specificity (in percentage) and F1 score for each model

Models	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**Specificity (%)**	**F1 Score**
LR	93.203	57.601	32.770	98.061	0.42
LR + SMOTE	86.312	33.186	82.856	86.590	0.47
NN	92.769	6.444	60.355	93.026	0.116
NN + SMOTE	93.324	24.586	59.029	94.387	0.35
XGBoost	93.331	57.074	33.196	98.043	0.42
XGBoost + SMOTE	84.064	29.833	86.949	83.836	0.44
RandomForest	93.025	52.942	33.112	97.702	0.41
RandomForest + SMOTE	85.415	30.79	79.431	85.888	0.44

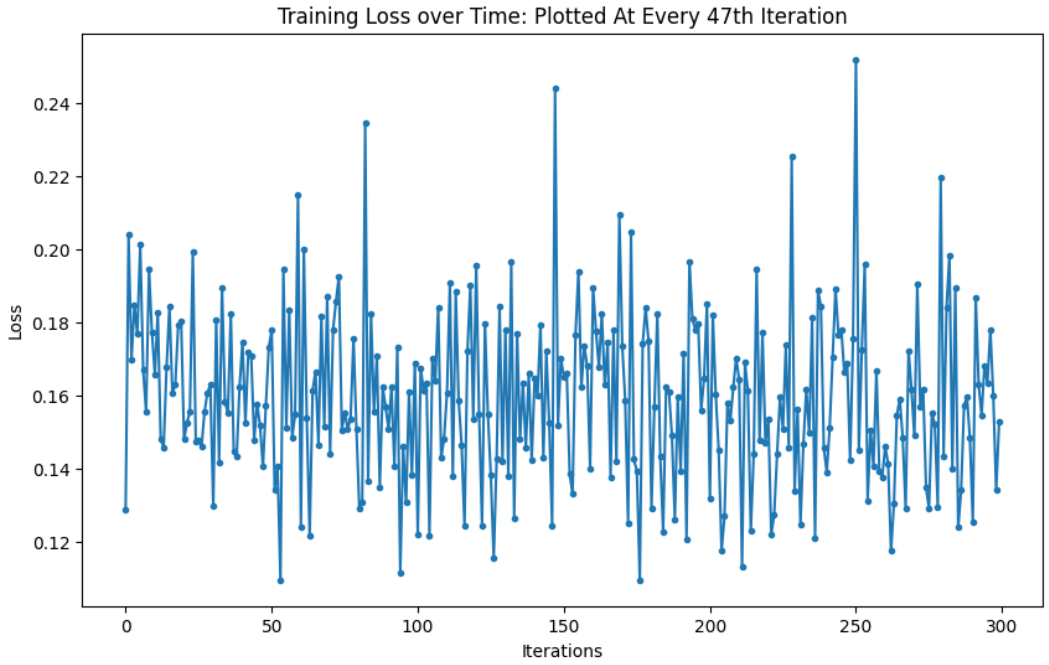

Figure 10: NN Loss Function

The NN model did not perform as expected. The most complex model came to be the least accurate as it was outperformed by other simpler predictive models. As seen in figure 10, the model's loss during training did not steadily decrease over iterations. Instead the loss was unstable and sporadic which indicates that its predictions did not improve significantly over time during training.

Due to an imbalanced class, all the models are biased towards a negative prediction which results in a low recall percentage as the models do not predict positive cases.

Using SMOTE, the recall percentage improves as there are more true positive predictions, however the precision percentage worsens as the amount of false positive predictions also increases.

The exception is the Neural Network Model whose recall percentage decreased and precision percentage increased with SMOTE.

According to the F1 Score, using SMOTE to enhance data results in a slightly more accurate model, the Logistic Regression model with SMOTE being the most accurate.

6 Conclusion

According to the F1 score, the Logistic Regression model with SMOTE was found to be the most accurate model while the Neural Network model was the least accurate. The underperformance of the neural network (NN) model in predicting COVID mortality, when compared to other models, illustrates that simplicity, on occasion, is key. However, NN models do vary greatly based on their architecture. NN models with different architectures that weren't developed during this research could perform better. Despite improvement with SMOTE, the models remain biased towards negative predictions due to the lack of COVID deaths within the dataset.

Using Gradio, the LR model was implemented into an experimental online public information tool.

7 References

[1] Nizri, Meir. "Covid-19 Dataset." Kaggle, 13 Nov. 2022, `https://www.kaggle.com/datasets/meirnizri/covid19-dataset?resource=download`.

[2] Sharifrazi, Danial. "Covid-19 Numeric Dataset." Kaggle, 8 Feb. 2021, `https://www.kaggle.com/datasets/danialsharifrazi/covid19-numeric-dataset?select=dataset2.xlsx`.