

# Prediction model for covid-19 patients

---

FINAL PROJECT FOR MACHINE LEARNING 64064-003

Tammy Liu  
WLIU16@KENT.EDU

## Table of Contents

<b>1. PROJECT GOAL .....</b>	<b>2</b>
<b>2. OVERVIEW OF THE DATA: .....</b>	<b>2</b>
<b>DATA SOURCE.....</b>	<b>2</b>
<b>DATA EXPLORATION .....</b>	<b>2</b>
<b>DATA PREPARATION.....</b>	<b>4</b>
<b>DATA PARTITION: .....</b>	<b>5</b>
<b>3. DETAILS OF MODELLING STRATEGY .....</b>	<b>6</b>
<b>LOGISTIC REGRESSION MODEL .....</b>	<b>6</b>
<b>KNN MODEL .....</b>	<b>7</b>
<b>NAÏVE BAYES MODEL .....</b>	<b>8</b>
<b>DECISION TREE MODEL .....</b>	<b>8</b>
<b>4. ESTIMATION OF MODEL'S PERFORMANCE .....</b>	<b>10</b>
<b>MODEL ACCURACY.....</b>	<b>10</b>
<b>MODEL PERFORMANCE .....</b>	<b>10</b>
<b>PERFORMANCE METRICS .....</b>	<b>12</b>
<b>5. INSIGHTS AND CONCLUSION.....</b>	<b>12</b>
<b>APPENDIX.....</b>	<b>13</b>

## 1. PROJECT GOAL

ABC Clinic is a non-profit medical center that provides clinical and hospital care in Northeast Ohio. It is a leader in fighting COVID-19 and is seeking consultation to better understand which COVID-positive patients have higher death rates. With the help of the ABC Clinic's historical data, we aim to develop a model to predict and identify COVID-19 patients' chance of survival. It is out of the study's scope to look at patients' survival rate who have not been identified as COVID-19 positive.

The task of our team is to apply analytics and help the management of the hospital gain understanding on covid-positive patients' survival rate based on their historic data.

## 2. OVERVIEW OF THE DATA:

### DATA SOURCE

The data is downloaded from Kaggle at COVID-19 dataset, original provided by the Mexican government (Kaggle COVID-19 Dataset, n.d.).

### DATA EXPLORATION

The data contains over 1 million observations of patient data with their symptoms, status and medical history. There are a total of 21 variables (Figure 1) including patient general information (sex, age), symptoms (classification of covid, type of patient care received, whether or not has used ventilator or ICU), and medical history ( pregnancy, pneumonia, diabetes, COPD, Asthma, inmsupr, hypertension, cardiovascular, chronic renal, other disease, obesity, tobacco). The data also has information if the patient has died.

summary(Covid)						
***						
SEX	AGE	CLASIFFICATION_FINAL	PATIENT_TYPE	PNEUMONIA	PREGNANT	DIABETES
Min. :1.000	Min. : 0.00	Min. :1.000	Min. :1.000	Min. : 1.000	Min. : 1.00	Min. : 1.000
1st Qu.:1.000	1st Qu.: 30.00	1st Qu.:3.000	1st Qu.:1.000	1st Qu.: 2.000	1st Qu.: 2.00	1st Qu.: 2.000
Median :1.000	Median : 40.00	Median :6.000	Median :1.000	Median : 2.000	Median :97.00	Median : 2.000
Mean :1.499	Mean : 41.79	Mean :5.306	Mean :1.191	Mean : 3.347	Mean :49.77	Mean : 2.186
3rd Qu.:2.000	3rd Qu.: 53.00	3rd Qu.:7.000	3rd Qu.:1.000	3rd Qu.: 2.000	3rd Qu.:97.00	3rd Qu.: 2.000
Max. :2.000	Max. :121.00	Max. :7.000	Max. :2.000	Max. :99.000	Max. :98.00	Max. :98.000
COPD	ASTHMA	INMSUPR	HIPERTENSION	CARDIOVASCULAR	RENAL_CHRONIC	OTHER_DISEASE
Min. : 1.000	Min. : 1.000	Min. : 1.000	Min. : 1.000	Min. : 1.000	Min. : 1.000	Min. : 1.000
1st Qu.: 2.000	1st Qu.: 2.000	1st Qu.: 2.000	1st Qu.: 2.000	1st Qu.: 2.000	1st Qu.: 2.000	1st Qu.: 2.000
Median : 2.000	Median : 2.000	Median : 2.000	Median : 2.000	Median : 2.000	Median : 2.000	Median : 2.000
Mean : 2.261	Mean : 2.243	Mean : 2.298	Mean : 2.129	Mean : 2.262	Mean : 2.257	Mean : 2.435
3rd Qu.: 2.000	3rd Qu.: 2.000	3rd Qu.: 2.000	3rd Qu.: 2.000	3rd Qu.: 2.000	3rd Qu.: 2.000	3rd Qu.: 2.000
Max. :98.000	Max. :98.000	Max. :98.000	Max. :98.000	Max. :98.000	Max. :98.000	Max. :98.000
OBESENITY	TOBACCO	USMER	MEDICAL_UNIT	INTUBED	ICU	DATE_DIED
Min. : 1.000	Min. : 1.000	Min. :1.000	Min. : 1.000	Min. : 1.00	Min. : 1.00	Length:1048575
1st Qu.: 2.000	1st Qu.: 2.000	1st Qu.:1.000	1st Qu.: 4.000	1st Qu.:97.00	1st Qu.:97.00	Class :character
Median : 2.000	Median : 2.000	Median :2.000	Median :12.000	Median :97.00	Median :97.00	Mode :character
Mean : 2.125	Mean : 2.214	Mean :1.632	Mean : 8.981	Mean :79.52	Mean :79.55	
3rd Qu.: 2.000	3rd Qu.: 2.000	3rd Qu.:2.000	3rd Qu.:12.000	3rd Qu.:97.00	3rd Qu.:97.00	
Max. :98.000	Max. :98.000	Max. :2.000	Max. :13.000	Max. :99.00	Max. :99.00	

Figure 1 Covid-19 data summary

Across the dataset, **1 is used for positive** and **2 for negative**.

There are a few early observations from the dataset:

- The death rate is about 13.8% (54236 death cases) of all COVID patients (391917 cases)
- Positive male (sex=2) death rate is almost double of the death rate of female (sex =1) (Figure 2).
- The death for pregnancy cases is very few and pregnancy is not a focus for our study.
- High percentage of data are missing for INTUBED and ICU. As INTUBED and ICU are only used in extreme conditions, let's assume all the missing data as negative.
- The rest of NAs in each column are normal (<1% are missing) and will be imputed with median values.

DEATH	
1	2
54236	337743

SEX	
1	2
18959	35277
163531	174212

PREGNANT	
1	2
65	18853
2689	159500

*Figure 2 Distribution of Death, by sex and by pregnancy*

## DATA PREPARATION

- Convert all binary data: Since most of the columns use 1 for “Yes” and 2 for “No”, convert CLASSIFICATION and DEATH columns to binary
- Filter only COVID positive cases
- Convert other forms to NA: convert 97, 98, 99 to NAs
- Change attribute: Convert DEATH from date to factor
- Remove unnecessary variables: removed pregnancy, USMER, MEDICAL\_UNIT, DEATH\_DATE, CLASIFICATION\_FINAL
- Delete NA records for age
- Impute the NAs with medians: We replaced NAs with median values for 13 columns (Figure 3)
- By this step, there are 391979 observations of 17 variables left

### **\*\*Data Preparation\*\***

```
```\r}
Covid <- Covid[-c(3, 6, 17, 18, 21, 22)] #delete pregnant, USMER, MEDICAL_UNIT etc
#391979 obs. of 17 variables
Covid <- Covid%>%
  filter(!is.na(AGE)) #391853 obs. of 17 variables

#impute missing values with mean
Covid[, c(4:16)] <- Covid[, c(4:16)]%>%
  mutate_if(is.numeric, function(x) ifelse(is.na(x), median(x, na.rm = T), x))

#Change data attribute from character to factor, the data is coded as 1 as no and 2 as yes
Covid$DEATH <- as.factor(Covid$DEATH) |
```\r}
```

Figure 3 Data preparation

### DATA PARTITION:

Once the data has no missing values, we partition the data using the CARET package in R in training and test sets. The partition index is set at “0.70,” which refers to the training set of 70% of the whole dataset and the test set of 30% (Figure 4).

### **\*\*Data partition\*\***

```
```\r}
#Partition the given training data into 70% training data and 30% testing data
set.seed(100)
index_train <- createDataPartition(Covid$DEATH, p=0.7, list= F)
Covid_train <- Covid[index_train, ]
Covid_test <- Covid[-index_train, ]
```\r}
```

Figure 4 Data Partition

### 3. DETAILS OF MODELLING STRATEGY

We developed four models (Logistic regression, KNN, Naïve bayes and Decision Tree) (**Error! Reference source not found.**) based on the 70% of the training data to determine the most accurate model for predicting the death of COVID-19 patients

#### LOGISTIC REGRESSION MODEL

A logistic regression model is considered since the data's target variable is categorical. When predicting a binomial attribute, a linear regression model is less optimal since its performance likelihood can be negative or more than 1. Logistic regression, ranging between 0 and 1, is the desired outcome for this model.

From the model output (Figure 5), we can interpret information similar to what we have observed before. Among all the factors, the red ones tend to show a higher death rates and blue ones are correlated to lower death rates in COVID-positive patients. Correlation is not causation and we do not conclude red factors lead to a higher death rate.

- SEX: **male**, **female**
- AGE: **old**, **young**
- PATIENT\_TYPE: **home care**, **hospitalized**
- **PNEUMONIA**: if the patient already has air sacs inflammation
- **DIABETES**: if the patient already has diabetes
- **COPD**: if the patient has Chronic obstructive pulmonary disease
- **ASTHMA**: : if the patient has asthma
- **INMSUPR**: if the patient is immunosuppressed

- **HIPERTENSION**: if the patient has hypertension
- **CARDIOVASCULAR**: the patient has heart or blood vessels related disease
- **RENAL\_CHRONIC**: if the patient has chronic renal disease
- **OTHER\_DISEASE**: if the patient has other disease
- **OBESITY**: if the patient is obese
- **TOBACCO**: if the patient is a tobacco user
- **INTUBED**: if the patient has been connected to the ventilator
- **ICU**: if the patient has been admitted to an intensive care unit

```

**Run logistic regression model**
```{r}
set.seed(1)
log_model <- glm(DEATH~., data = Covid_train, family = 'binomial')
log_model
```

```

Call: glm(formula = DEATH ~ ., family = "binomial", data = Covid\_train)

Coefficients:

|             |              |                |               |               |          |          |          |
|-------------|--------------|----------------|---------------|---------------|----------|----------|----------|
| (Intercept) | SEX          | AGE            | PATIENT_TYPE  | PNEUMONIA     | DIABETES | COPD     | ASTHMA   |
| -0.70032    | -0.41051     | -0.05244       | -1.98410      | 1.19472       | 0.29444  | 0.13309  | -0.07068 |
| INMSUPR     | HIPERTENSION | CARDIOVASCULAR | RENAL_CHRONIC | OTHER_DISEASE | OBESITY  | TOBACCO  | INTUBED  |
| 0.29018     | 0.12659      | -0.04596       | 0.73844       | 0.30944       | 0.23432  | -0.13487 | 2.50672  |
| ICU         |              |                |               |               |          |          |          |
| -0.54479    |              |                |               |               |          |          |          |

Degrees of Freedom: 274297 Total (i.e. Null); 274281 Residual  
Null Deviance: 220400  
Residual Deviance: 109100 AIC: 109100

Figure 5 Logistic Regression Model

## KNN MODEL

KNN model failed to run as there are too many ties in the result which KNN can't deal with. This means that there are many similar data points and it is hard for KNN to tell which are the nearest since many points have the same distance. We will forgo KNN model for this use case.



## NAÏVE BAYES MODEL

Naïve Bayes model can show the conditional probabilities of each variable (Figure 6).

```
**Run NB model**
``{r, include = FALSE}
library(e1071)
set.seed(3)
nb_model <- naiveBayes(DEATH~., data = Covid_train)
nb_model
````
```

Naive Bayes Classifier for Discrete Predictors

Call:

naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:

|   |  |           |           |
|---|--|-----------|-----------|
| Y |  | 1         | 2         |
|   |  | 0.1382657 | 0.8617343 |

Conditional probabilities:

|   |     |          |           |
|---|-----|----------|-----------|
|   | SEX |          |           |
| Y |     | [,1]     | [,2]      |
| 1 |     | 1.651426 | 0.4765249 |
| 2 |     | 1.516191 | 0.4997389 |

*Figure 6 Naive Bayes model*

## DECISION TREE MODEL

From the DT model, we can tell that covid-positive patients that have higher death rate are those (Figure 7, Figure 8)

- 1) who have been hospitalized and connected to a ventilator (6%)
- 2) who have been hospitalized, age  $\geq 71$ , and has past medical history of pneumonia (3%)

**\*\*Run Decision Tree\*\***

```

{r}
set.seed(4)
#agnes or hclust object does not work with later prediction
dt_model <- rpart(DEATH~., data = Covid_train, method = "class") #class for binary
rpart.plot(dt_model, extra = 110, main = "Dendrogram of rpart")

```

Figure 7 Decision Tree model

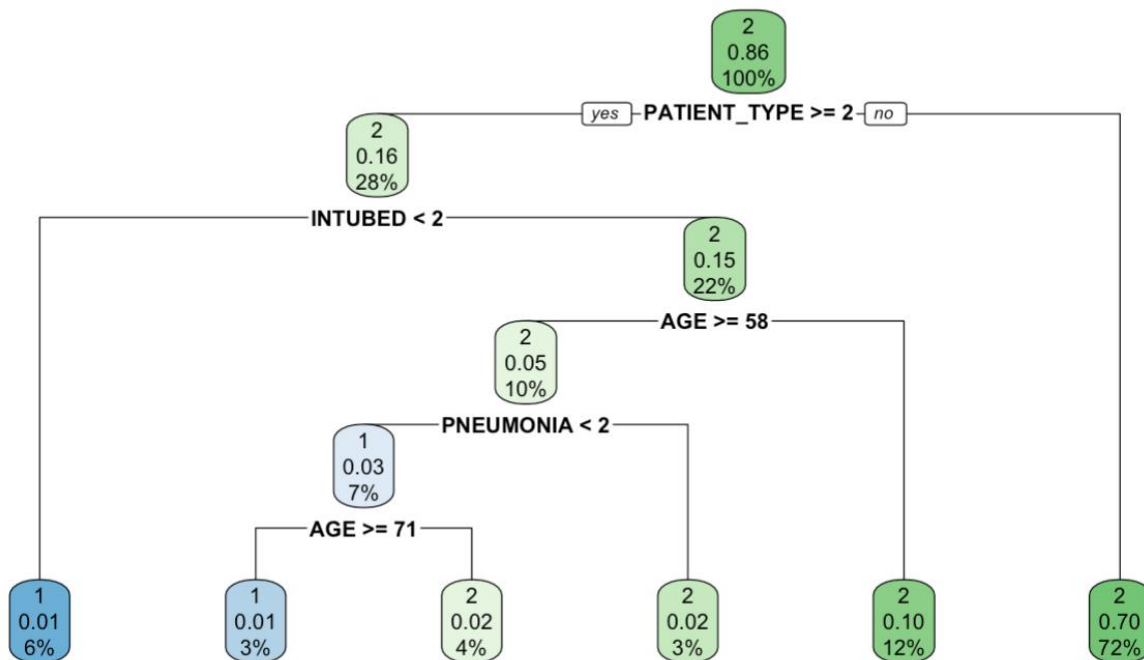


Figure 8 DT Dendrogram

## 4. ESTIMATION OF MODEL'S PERFORMANCE

### MODEL ACCURACY

- We apply the model on the test data to calculate the ROC curve to examine the algorithm's effectiveness and determine the best threshold based on our tolerance for false negatives and desire for true positives. The area under the curve determines the model accuracy where a perfect classifier would be 1. Therefore, the higher the AUC, the more confident we are in our model's predictive ability

### MODEL PERFORMANCE

- Confusion matrix are created to balance the trade-off between false positive and false negative. It reveals information on sensitivity (how good the model can detect a positive patient) and specificity (how good the model can detect a negative patient). We need to choose a model that balance the sensitivity and specificity while also have a  $AUC > 0.85$ . If sensitivity is similar to specificity, we emphasize more on reaching a good sensitivity as it is more important to reduce false negative than false positive.
- Below is example of ROC plot (Figure 9) and confusion matrix (Figure 10) from Logistic Regression.

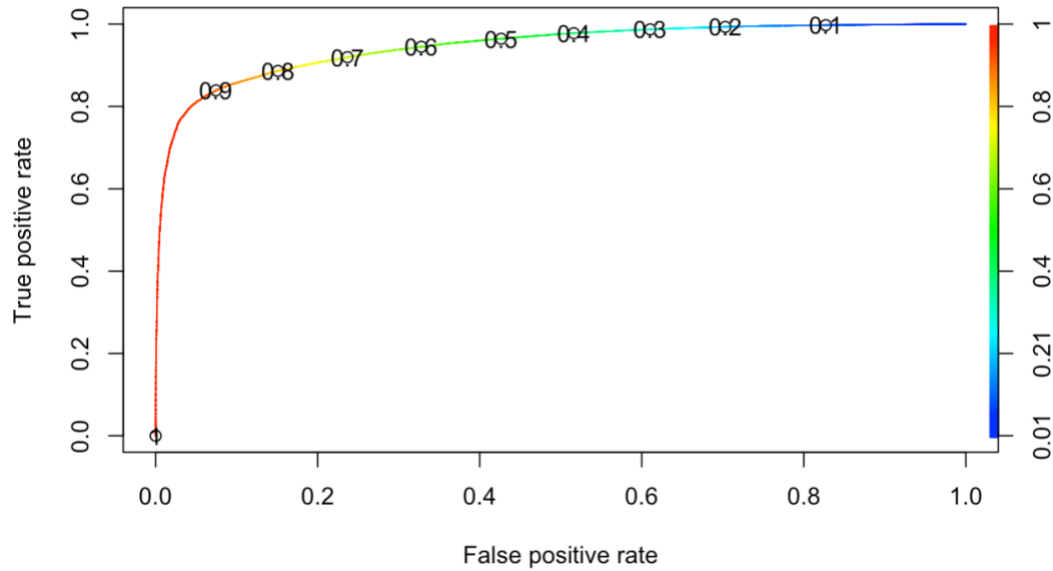


Figure 9 ROC Curve for Logistic Regression

#### Confusion Matrix and Statistics

|            | Reference |        |
|------------|-----------|--------|
| Prediction | 1         | 2      |
| 1          | 13442     | 101037 |
| 2          | 2812      | 264    |

Accuracy : 0.1166  
 95% CI : (0.1148, 0.1184)  
 No Information Rate : 0.8617  
 P-Value [Acc > NIR] : 1  
  
 Kappa : -0.0482  
  
 McNemar's Test P-Value : <2e-16  
  
 Sensitivity : 0.826996  
 Specificity : 0.002606  
 Pos Pred Value : 0.117419  
 Neg Pred Value : 0.085826  
 Prevalence : 0.138267  
 Detection Rate : 0.114346  
 Detection Prevalence : 0.973834  
 Balanced Accuracy : 0.414801  
  
 'Positive' Class : 1

Figure 10 Confusion matrix for logistic regression

## PERFORMANCE METRICS

Performance of the four models are compared against each other (see . ).

| Metric\Model               | Logistic Regression | Naïve Bayes | Decision Tree |
|----------------------------|---------------------|-------------|---------------|
| <b>ROC value</b>           | 0.95                | 0.92        | 0.91          |
| <b>True Positive (TP)</b>  | 13442               | 13857       | 14752         |
| <b>True Negative (TN)</b>  | 264                 | 85337       | 82584         |
| <b>False Positive (FP)</b> | 101037              | 15964       | 18717         |
| <b>False Negative (FN)</b> | 2812                | 2397        | 1502          |
| <b>Miscalculations</b>     | 103849              | 18361       | 20219         |
| <b>Accuracy</b>            | 11.66%              | 84.38%      | 82.80%        |
| <b>Sensitivity</b>         | 82.70%              | 85.25%      | 90.76%        |
| <b>Specificity</b>         | 0.26%               | 84.24%      | 81.52%        |

Table 1 Performance metric by model

Based on TABLE1, we will choose decision tree model as the best model to predict patient survival situation. Decision Tree model has the best sensitivity rate while not compromising too much on specificity therefore it is the best model to use. It also has good capability reducing number of false negatives among all three models.

## 5. INSIGHTS AND CONCLUSION

We recommend the Decision Tree model to ABC Clinic to use on future patient data and identify if the patient have a higher death rate. We aim to help ABC Clinic predict COVID patients survival situation beforehand so they take into account the additional information when treating the patient in order to increase patient survival rate. The model intend to reduce false negatives (correctly identify all positive patients) but produce more false positives (some patients with good survival will be marked with higher death rate).

If we are provided with the cost for treating each false positive and false negative patient, the total cost of saving for the hospital could also be calculated.

## APPENDIX

### Works Cited

*Kaggle COVID-19 Dataset*. (n.d.). Retrieved from  
<https://www.kaggle.com/datasets/meirnazri/covid19-dataset>

|                                                               |    |
|---------------------------------------------------------------|----|
| Figure 1 Covid-19 data summary .....                          | 3  |
| Figure 2 Distribution of Death, by sex and by pregnancy ..... | 4  |
| Figure 3 Data preparation .....                               | 5  |
| Figure 4 Data Partition .....                                 | 5  |
| Figure 5 Logistic Regression Model .....                      | 7  |
| Figure 6 Naive Bayes model .....                              | 8  |
| Figure 7 Decision Tree model .....                            | 9  |
| Figure 8 DT Dendrogram .....                                  | 9  |
| Figure 9 ROC Curve for Logistic Regression .....              | 11 |
| Figure 10 Confusion matrix for logistic regression .....      | 11 |
| Table 1 Performance metric by model .....                     | 12 |