Big Data and Cloud Computing

CMP 4011

Phase 2

Team 5		
Name	**Section**	**BN**
Andrew Boshra	**1**	**17**
Peter Micheal	**1**	**21**
Taher Mohamed Ahmed	**1**	**37**
Omar Mohamed Ahmed	**2**	**6**

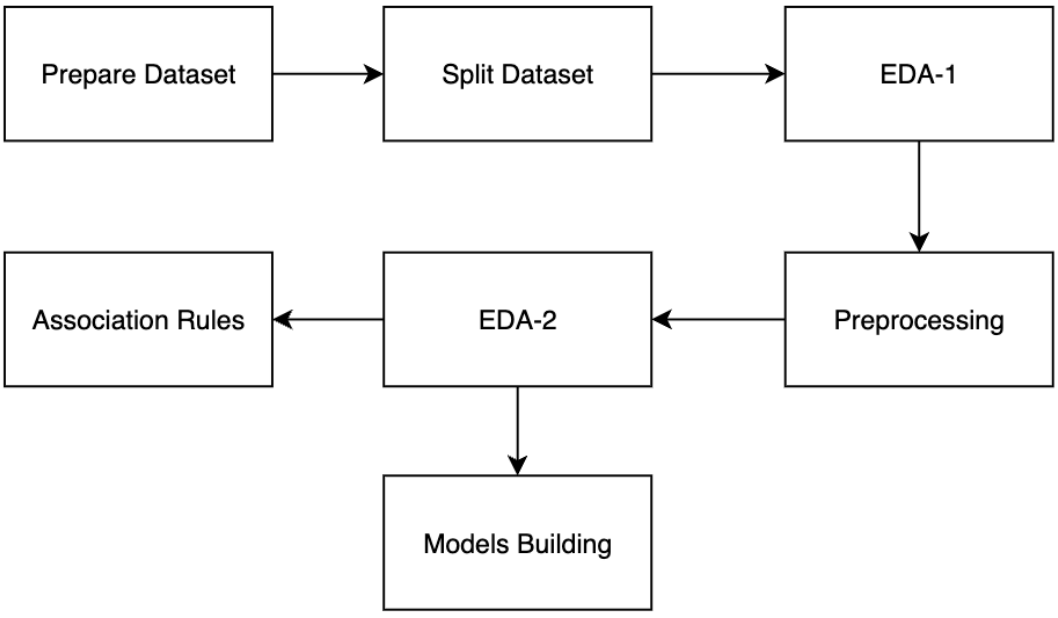
Problem statement

This project aims to determine the mortality status of patients, whether they have passed away or are still alive, in the context of the COVID-19 pandemic. The project utilizes a coronavirus dataset to classify patients based on their symptoms and investigates the association rules between these symptoms, other routines, and diseases. The main objective is to develop a predictive model that accurately identifies patients at risk of mortality and enables early interventions to mitigate the spread of the virus. Furthermore, the study of association rules aids in identifying the co-occurrence of symptoms and underlying diseases, facilitating early diagnosis and effective disease management.

Dataset

Link	https://www.kaggle.com/datasets/meirnizri/covid19-dataset
Number of features	21 unique features
Number of records	1,048,576 unique patients
Size	58.5 MB
Other notes	The dataset was provided by the Mexican government

Project Pipeline



Analysis and Solution of the Problem

- **Prepare dataset**

  ⇒ Convert the values 97,98, and 99 in the binary columns to nulls.

  ⇒ Convert the death date into a binary value to indicate whether the patient died or not.

- **Split dataset**

  ⇒ Split the dataset into 80% train and 20% test.

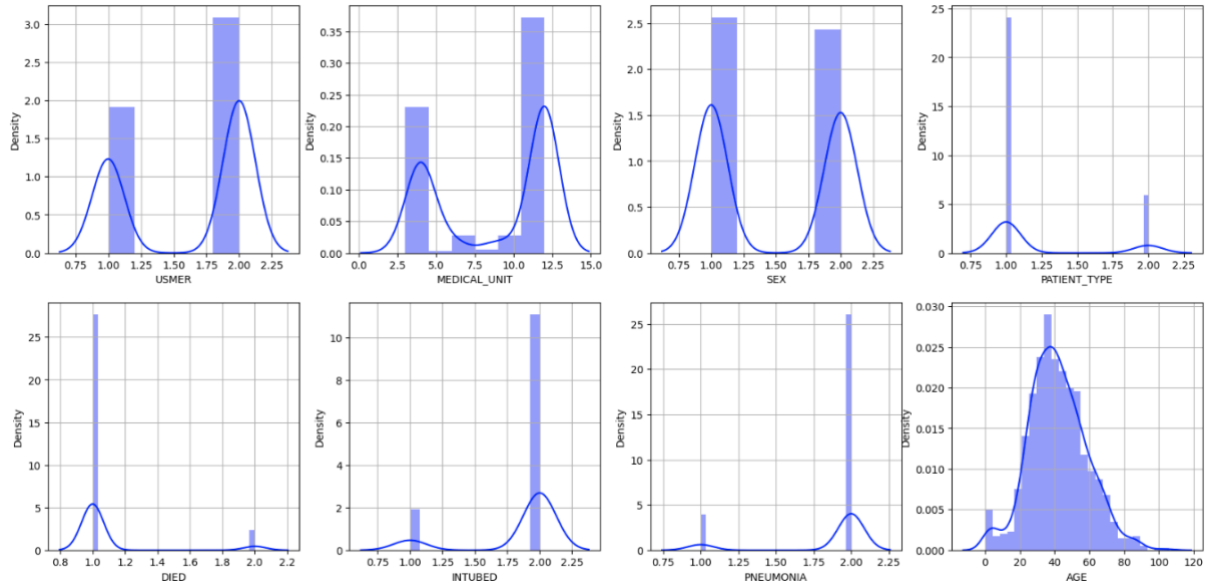
- **EDA-1**

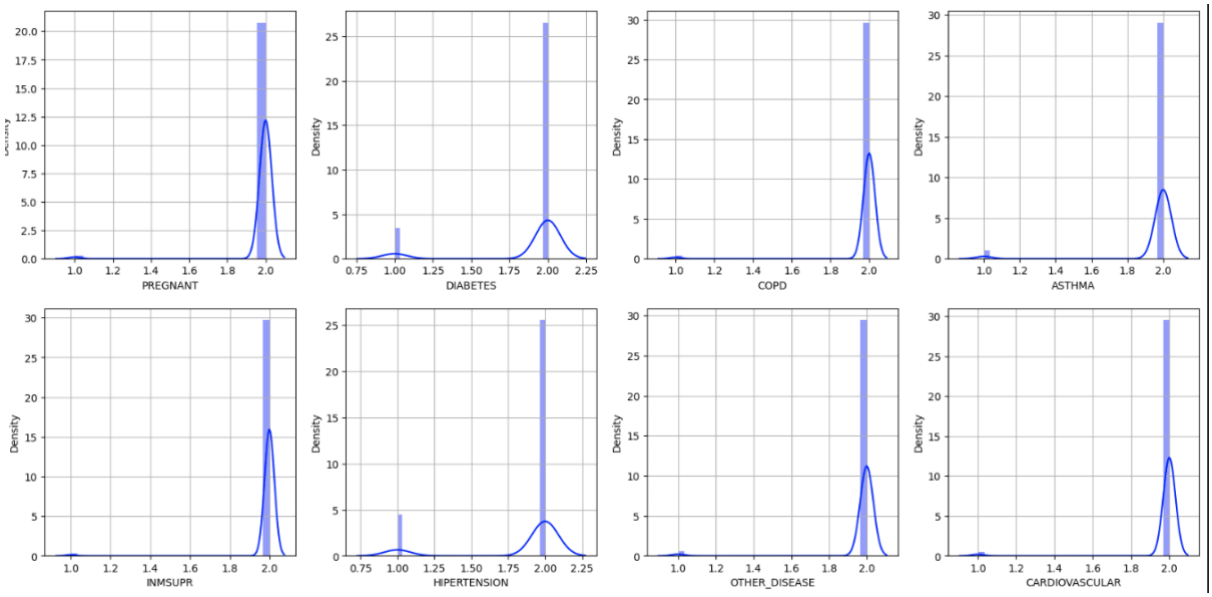
  In this step, we analyze the dataset before applying any preprocessing.

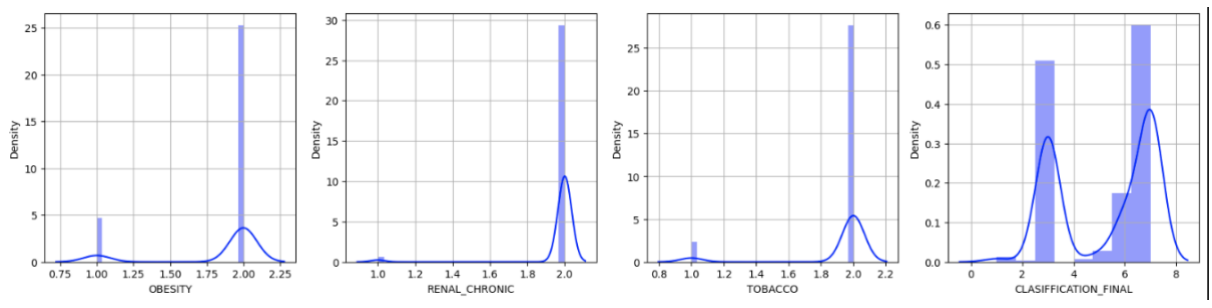
  - Check the percentage of the null value in each column

```
ICU                   81.336161
INTUBED               81.336161
PREGNANT              48.992577
PNEUMONIA              1.378579
OTHER_DISEASE          0.424178
TOBACCO                0.318134
OBESITY                0.318134
INMSUPR                0.212089
RENAL_CHRONIC          0.212089
CARDIOVASCULAR         0.212089
DIABETES               0.212089
COPD                   0.106045
ASTHMA                 0.106045
HIPERTENSION           0.000000
MEDICAL_UNIT           0.000000
AGE                    0.000000
DIED                   0.000000
PATIENT_TYPE           0.000000
SEX                    0.000000
CLASIFFICATION_FINAL   0.000000
USMER                  0.000000
Name: missing_percentage, dtype: float64
```

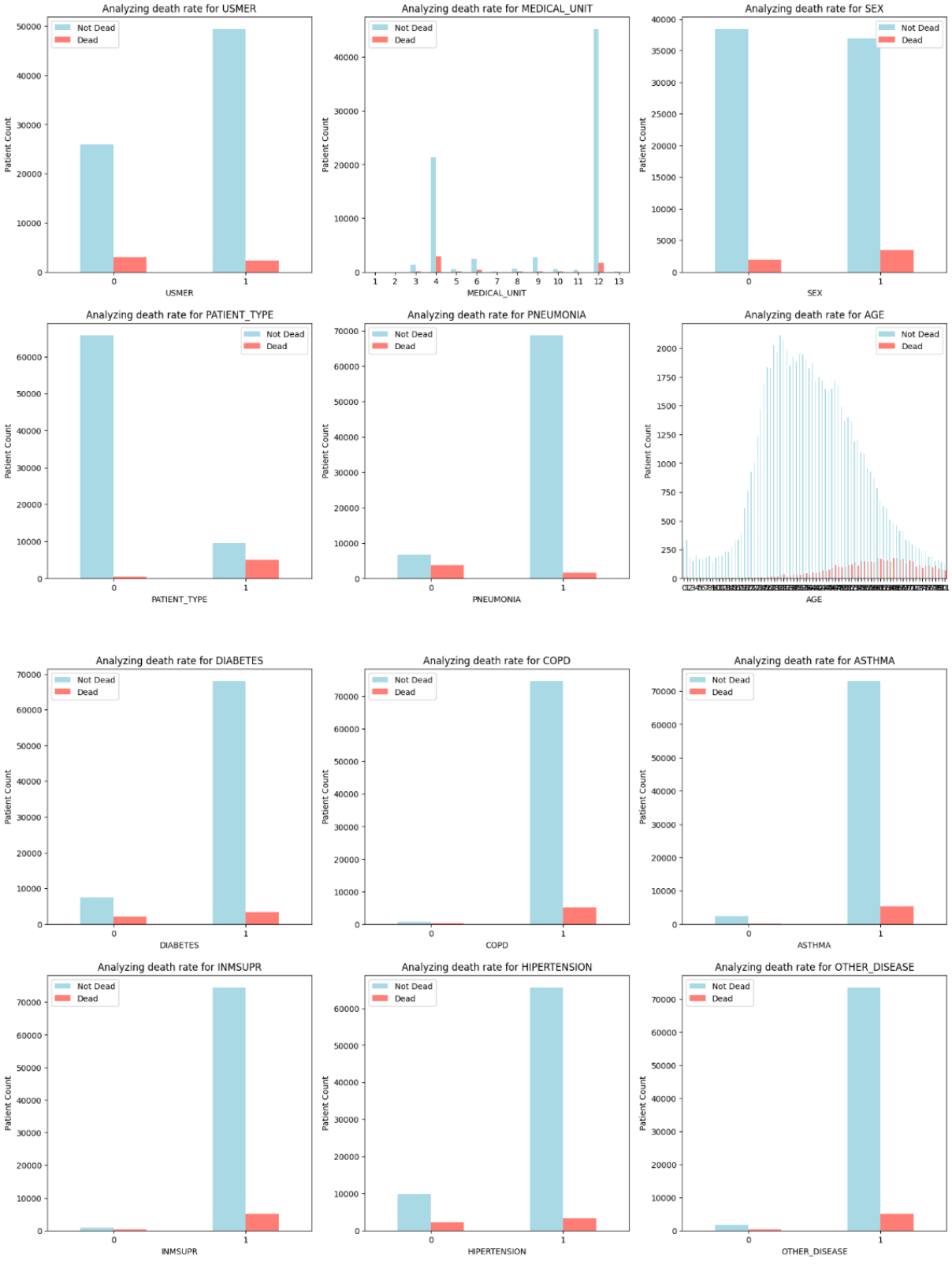
- Get unique values for each column
- Draw histograms to see the distribution of each feature.

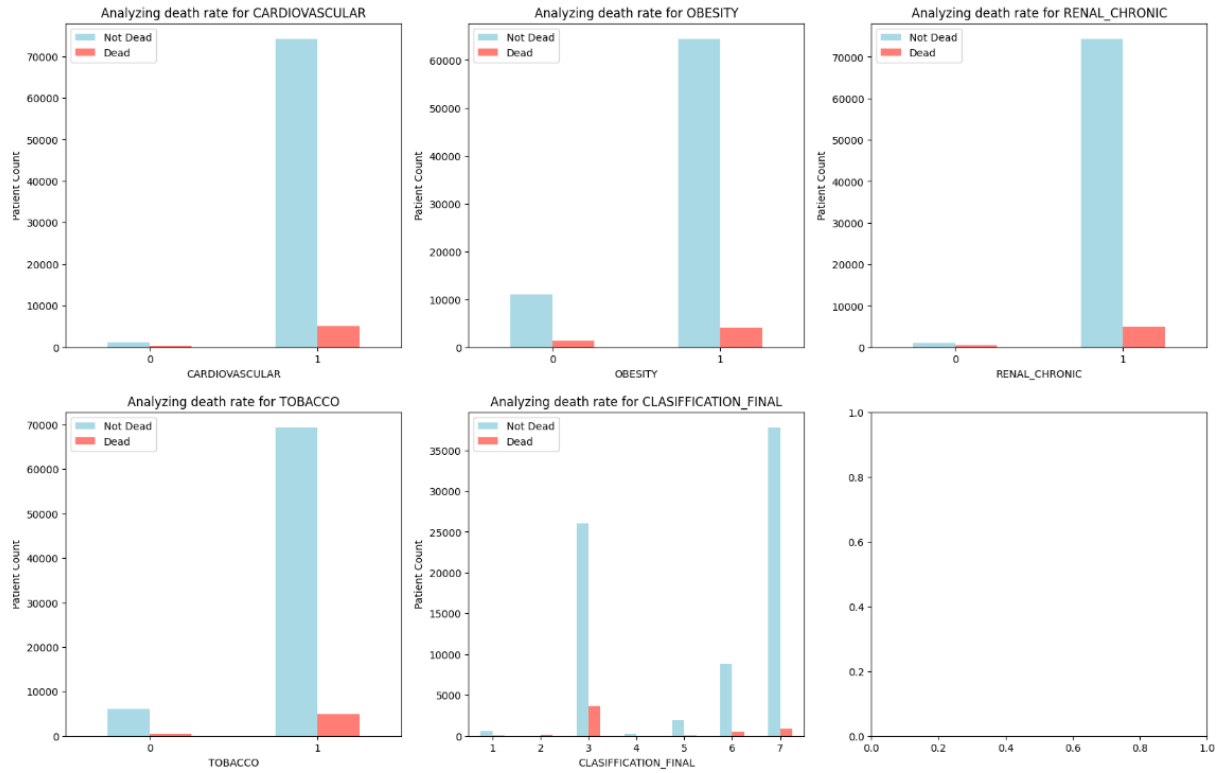




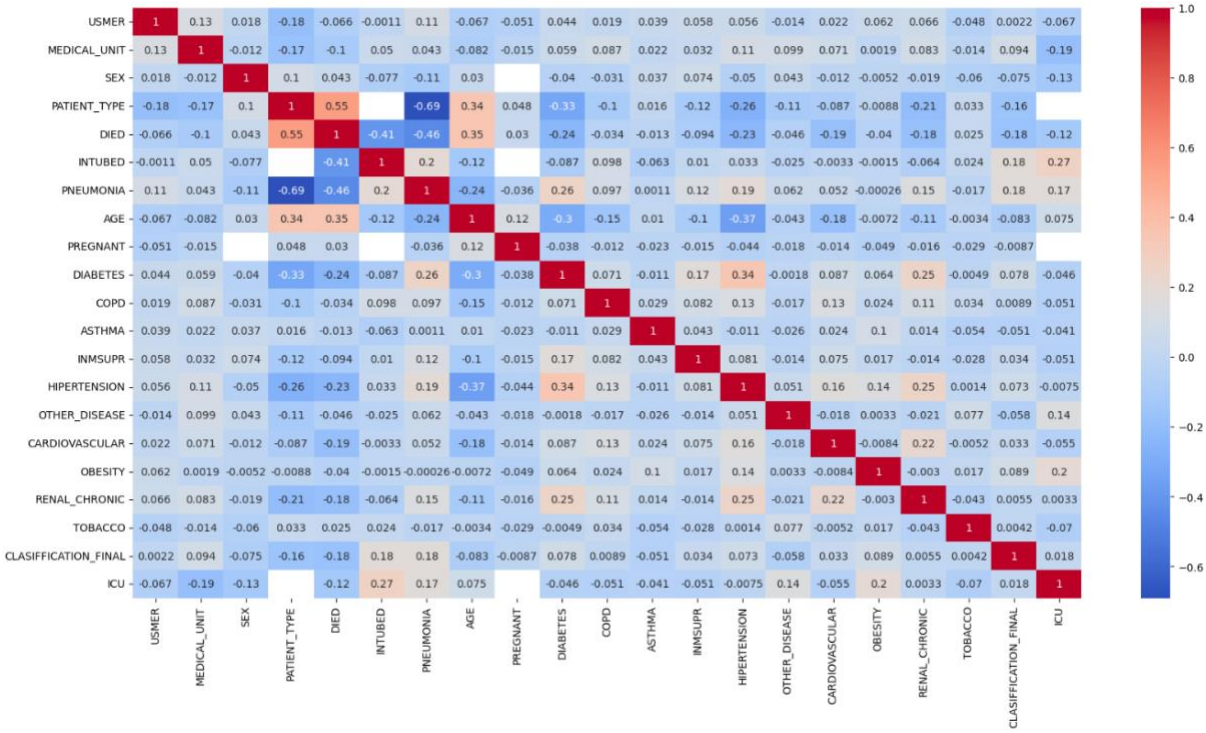


- Draw the death value with each feature.

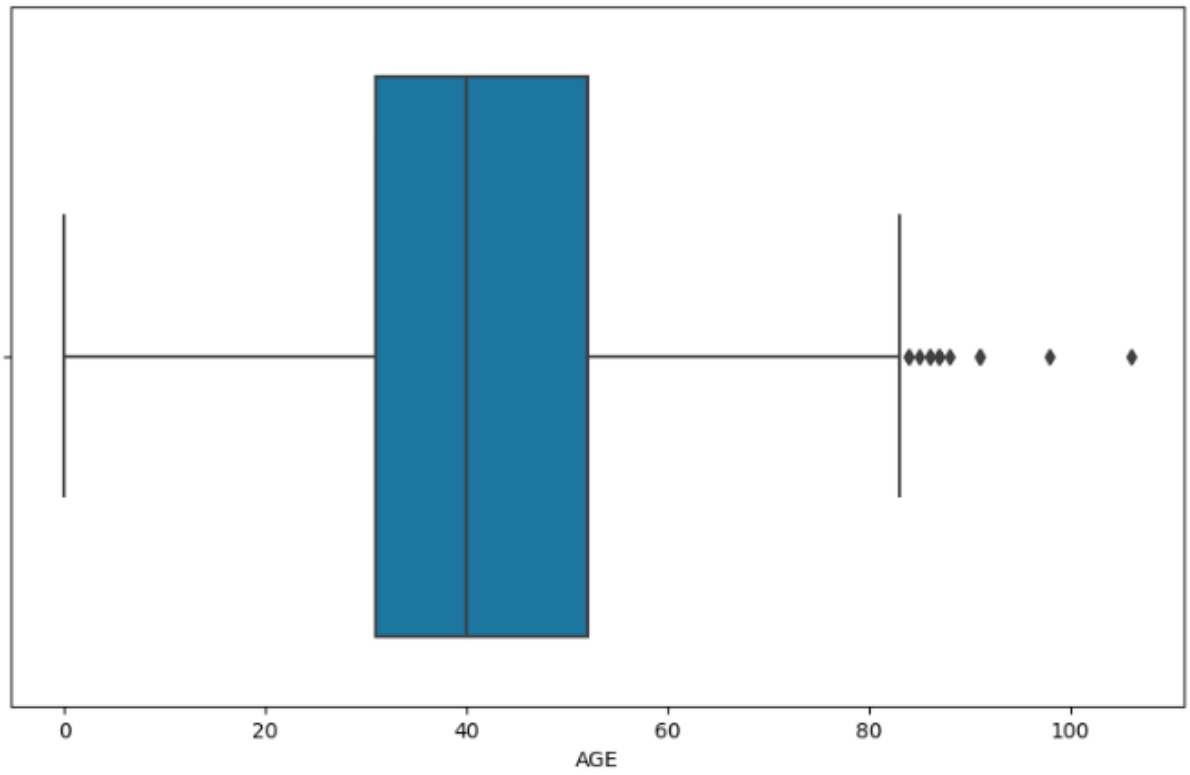




- Get the correlation between each pair of features.

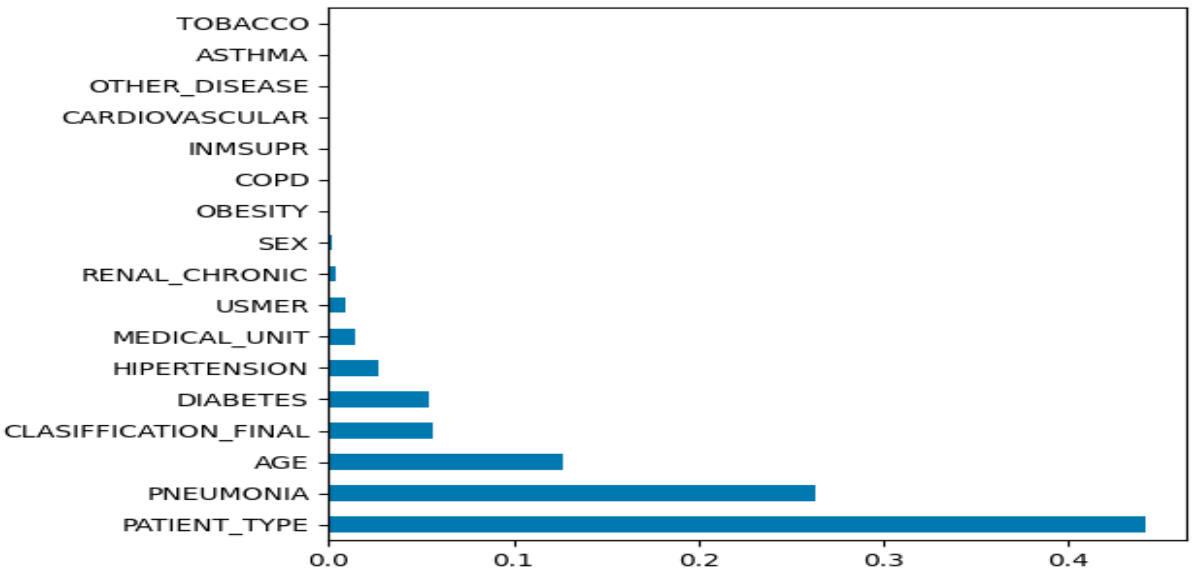


- Detect outliers in numerical columns (Age)

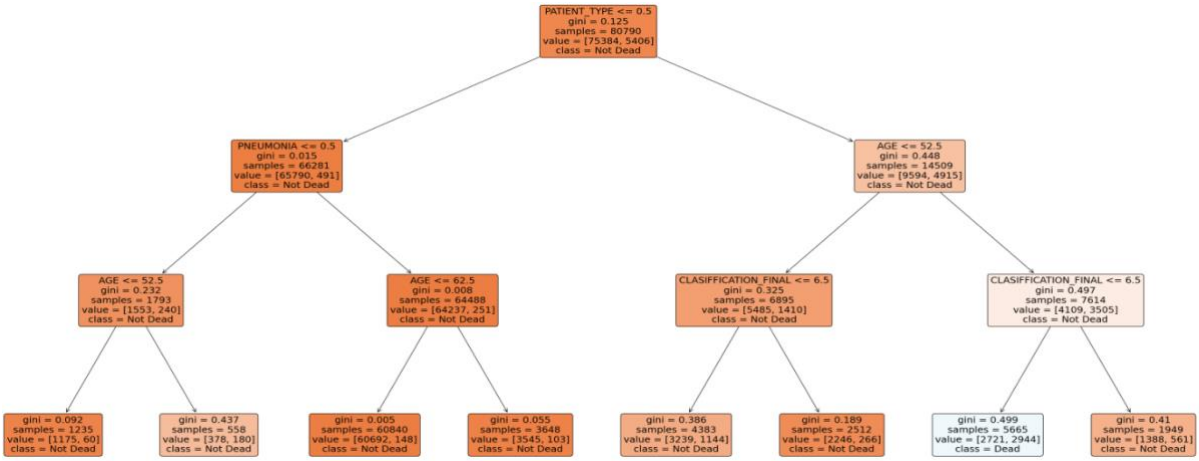


- Preprocessing
    - Drop columns that have high missing values (ICU/ INTUBED / Pregnant)
    - Drop rows with null values
    - Remove age outliers
    - Convert Boolean columns to 0,1 instead of 1,2
- EDA-2

  The same steps in EDA-1 were applied and the following are the new steps
    - Feature importance
        - Random Forest



- Decision trees



Observations and Insights

- From EDA
    - There is a class imbalance in the data.
    - The ratio of patients having diabetes, asthma, and other diseases is very high compared to the normal averages.
    - People who have an infection with a classification value of 3 are more likely to die.
    - The death rate in males is higher than in females.
    - People who got admitted to intensive care have a lower death rate.
    - The death rate increases with the increase in age.
    - Death is inversely correlated with diabetes and pneumonia.
    - Death is highly correlated with patient type which indicated that hospitalized patients are more likely to do. Being in the hospital not the home means that the case is at risk which gives a higher probability of death.
    - There is a good observation that indicates that there is an inverse correlation between age and hypertension.

- From Association Rules
    - Items with the highest support

	support	itemsets
2	0.988132	(COPD)
5	0.986913	(INMSUPR)
9	0.982614	(RENAL_CHRONIC)
1	0.982207	(CARDIOVASCULAR)
36	0.975818	(INMSUPR, COPD)
7	0.974579	(OTHER_DISEASE)
24	0.972063	(COPD, CARDIOVASCULAR)
40	0.971835	(COPD, RENAL_CHRONIC)
61	0.971468	(INMSUPR, RENAL_CHRONIC)
27	0.970408	(INMSUPR, CARDIOVASCULAR)

    - Items with the highest confidence

antecedents	consequents	support	confidence	lift	leverage
(ASTHMA, CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, INMSUPR, HIPERTENSION, USMER, DIABETES, OTHER_DISEASE, PNEUMONIA, TOBACCO)	(COPD)	0.378053	0.997204	1.009181	0.003439
(ASTHMA, CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, INMSUPR, HIPERTENSION, USMER, DIABETES, PNEUMONIA, TOBACCO)	(COPD)	0.384016	0.997196	1.009173	0.003491
(ASTHMA, CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, HIPERTENSION, USMER, DIABETES, OTHER_DISEASE, PNEUMONIA, TOBACCO)	(COPD)	0.379826	0.997191	1.009168	0.003451
(ASTHMA, CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, HIPERTENSION, USMER, DIABETES, PNEUMONIA, TOBACCO)	(COPD)	0.386245	0.997187	1.009164	0.003507
(ASTHMA, RENAL_CHRONIC, OBESITY, INMSUPR, HIPERTENSION, USMER, DIABETES, OTHER_DISEASE, PNEUMONIA, TOBACCO)	(COPD)	0.380064	0.997141	1.009117	0.003434
(ASTHMA, CARDIOVASCULAR, OBESITY, INMSUPR, HIPERTENSION, USMER, DIABETES, OTHER_DISEASE, PNEUMONIA, TOBACCO)	(COPD)	0.379618	0.997138	1.009114	0.003429
(ASTHMA, RENAL_CHRONIC, OBESITY, INMSUPR, HIPERTENSION, USMER, DIABETES, PNEUMONIA, TOBACCO)	(COPD)	0.386186	0.997135	1.009112	0.003487
(ASTHMA, CARDIOVASCULAR, OBESITY, INMSUPR, HIPERTENSION, USMER, DIABETES, PNEUMONIA, TOBACCO)	(COPD)	0.385631	0.997131	1.009107	0.003480
(ASTHMA, RENAL_CHRONIC, OBESITY, HIPERTENSION, USMER, DIABETES, OTHER_DISEASE, PNEUMONIA, TOBACCO)	(COPD)	0.381857	0.997129	1.009105	0.003445
(ASTHMA, RENAL_CHRONIC, OBESITY, HIPERTENSION, USMER, DIABETES, PNEUMONIA, TOBACCO)	(COPD)	0.388455	0.997126	1.009103	0.003504

Because the COPO is the most frequent item with support of 0.988, so this justifies why it appears as the consequence in the items with higher confidence.

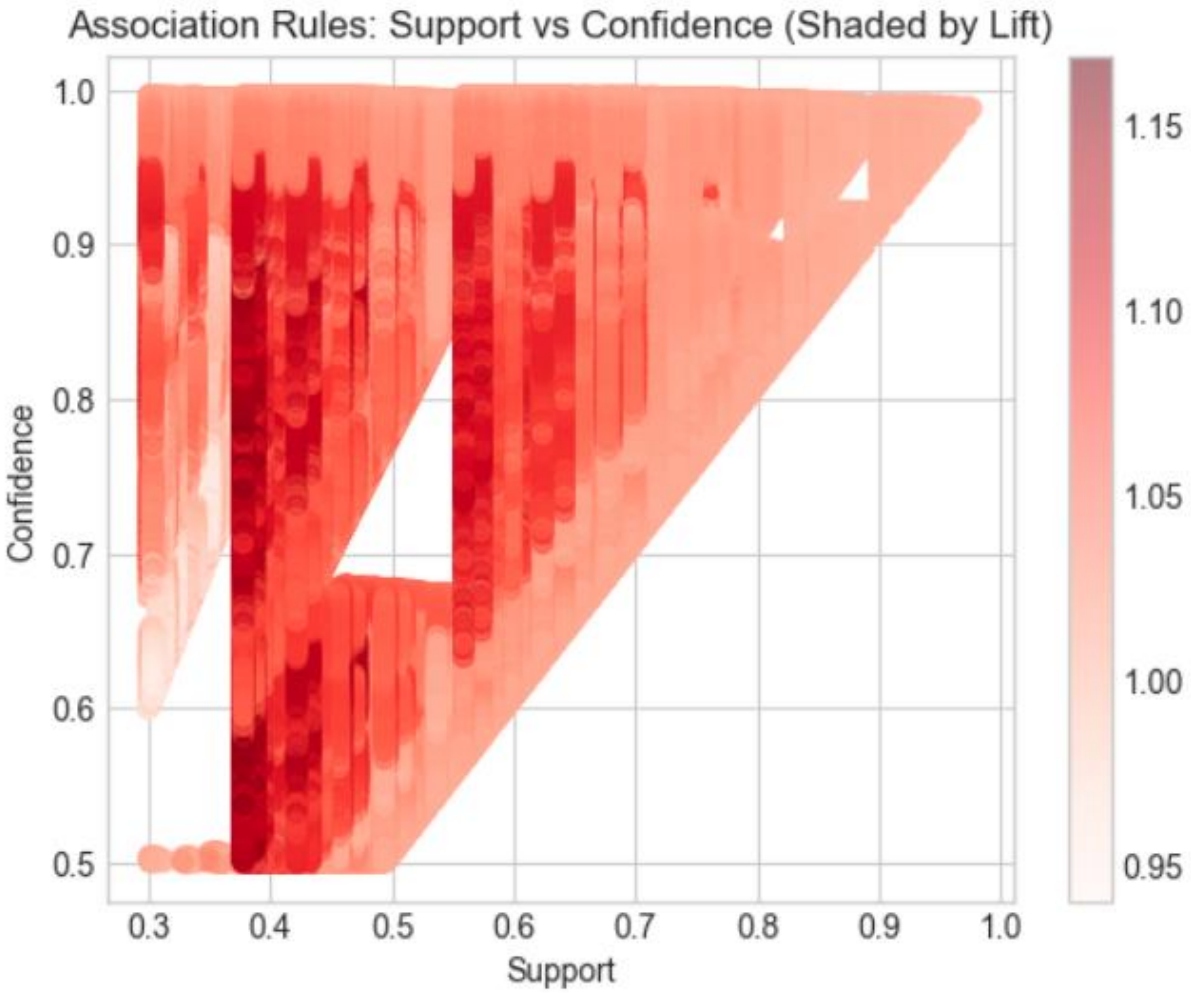
    - Items with the highest lift

antecedents	consequents	support	confidence	lift	leverage
(ASTHMA, HIPERTENSION, OTHER_DISEASE, PNEUMONIA, TOBACCO)	(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, INMSUPR, COPD, USMER, DIABETES)	0.378053	0.564579	1.168576	0.054537
(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, INMSUPR, COPD, USMER, DIABETES)	(ASTHMA, HIPERTENSION, OTHER_DISEASE, PNEUMONIA, TOBACCO)	0.378053	0.782501	1.168576	0.054537
(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, COPD, USMER, DIABETES, OTHER_DISEASE)	(ASTHMA, INMSUPR, HIPERTENSION, PNEUMONIA, TOBACCO)	0.378053	0.791541	1.167892	0.054347
(ASTHMA, INMSUPR, HIPERTENSION, PNEUMONIA, TOBACCO)	(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, COPD, USMER, DIABETES, OTHER_DISEASE)	0.378053	0.557803	1.167892	0.054347
(CARDIOVASCULAR, OBESITY, INMSUPR, COPD, USMER, DIABETES)	(ASTHMA, RENAL_CHRONIC, HIPERTENSION, OTHER_DISEASE, PNEUMONIA, TOBACCO)	0.378053	0.777241	1.167146	0.054140
(ASTHMA, RENAL_CHRONIC, HIPERTENSION, OTHER_DISEASE, PNEUMONIA, TOBACCO)	(CARDIOVASCULAR, OBESITY, INMSUPR, COPD, USMER, DIABETES)	0.378053	0.567703	1.167146	0.054140
(ASTHMA, COPD, HIPERTENSION, OTHER_DISEASE, PNEUMONIA, TOBACCO)	(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, INMSUPR, USMER, DIABETES)	0.378053	0.566868	1.166712	0.054020
(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, INMSUPR, USMER, DIABETES)	(ASTHMA, COPD, HIPERTENSION, OTHER_DISEASE, PNEUMONIA, TOBACCO)	0.378053	0.778097	1.166712	0.054020
(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, COPD, USMER, DIABETES)	(ASTHMA, INMSUPR, HIPERTENSION, OTHER_DISEASE, PNEUMONIA, TOBACCO)	0.378053	0.776293	1.166710	0.054020
(ASTHMA, INMSUPR, HIPERTENSION, OTHER_DISEASE, PNEUMONIA, TOBACCO)	(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, COPD, USMER, DIABETES)	0.378053	0.568185	1.166710	0.054020

From the pervious results it is observed that relations with higher lift have low support and confidence.

    - Items with the highest leverage

antecedents	consequents	support	confidence	lift	leverage
(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, COPD, DIABETES, OTHER_DISEASE)	(INMSUPR, PNEUMONIA, HIPERTENSION, TOBACCO)	0.573394	0.789381	1.131092	0.066456
(INMSUPR, PNEUMONIA, HIPERTENSION, TOBACCO)	(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, COPD, DIABETES, OTHER_DISEASE)	0.573394	0.821608	1.131092	0.066456
(ASTHMA, INMSUPR, HIPERTENSION, PNEUMONIA, TOBACCO)	(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, COPD, DIABETES, OTHER_DISEASE)	0.558732	0.824390	1.134922	0.066423
(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, COPD, DIABETES, OTHER_DISEASE)	(ASTHMA, INMSUPR, HIPERTENSION, PNEUMONIA, TOBACCO)	0.558732	0.769196	1.134922	0.066423
(HIPERTENSION, PNEUMONIA, ASTHMA, TOBACCO)	(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, INMSUPR, COPD, DIABETES, OTHER_DISEASE)	0.558732	0.817946	1.134280	0.066145
(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, INMSUPR, COPD, DIABETES, OTHER_DISEASE)	(HIPERTENSION, PNEUMONIA, ASTHMA, TOBACCO)	0.558732	0.774818	1.134280	0.066145
(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, COPD, DIABETES, OTHER_DISEASE)	(INMSUPR, HIPERTENSION, PNEUMONIA, ASTHMA)	0.600123	0.826178	1.123819	0.066120
(INMSUPR, HIPERTENSION, PNEUMONIA, ASTHMA)	(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, COPD, DIABETES, OTHER_DISEASE)	0.600123	0.816325	1.123819	0.066120
(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, INMSUPR, COPD, DIABETES, OTHER_DISEASE)	(PNEUMONIA, HIPERTENSION, TOBACCO)	0.573394	0.795150	1.130294	0.066098
(PNEUMONIA, HIPERTENSION, TOBACCO)	(CARDIOVASCULAR, RENAL_CHRONIC, OBESITY, INMSUPR, COPD, DIABETES, OTHER_DISEASE)	0.573394	0.815071	1.130294	0.066098



The best are to have a good association rules in the top right corner quarter with high lift value. This is not available, instead there is few good rules.

Model Training and Evaluation

In the training and evaluation process, we used the most important 10 features and yielded the best results.

	Dataset	Accuracy	Weighted Precision	Weighted Recall	Weighted F1
Logistic Regression	Train	0.941	0.933	0.941	0.935
	Test	0.941	0.932	0.941	0.935

Random Forest	Train	0.941	0.928	0.941	0.929
	Train	0.939	0.926	0.939	0.927
Decision Trees	Train	0.94	0.927	0.94	0.927
	Test	0.939	0.926	0.939	0.925
Naïve Bayes (map-reduce)	Train	0.921	0.945	0.921	0.93
	Test	0.922	0.944	0.922	0.93

Unsuccessful Trials

1. Impute the values of the three columns that have missing values of more than 40%. When using these columns, the model results were not good enough as the final models.
2. After Dropping these columns using the imputation for the missing values for other columns is not good as dropping null columns. We were able to drop null rows as the size of the data was large enough.
3. Before using feature, importance, and dropping unimportant columns the performance was lower than the final performance and this is justified by that there are so columns that are not correlated to the feature we are predicting.
4. Making the classification final feature a Boolean was no good idea, it is better to have a measure of how the patient is infected.

Future Work

1. Using other models
2. Try a neural network on this large number of rows
3. Do more feature engineering and try to define a feature that mixes two or more features to predict the results correctly.
4. Doing more data mining on the data.