**Documentation for COVID-19 Analysis**

**CRISP – Deep Learning Framework:**

1. collecting data and business understanding.
2. Exploratory data analysis
3. Data preprocessing
4. Building a Deep Learning model
5. Evaluation of the Deep Learning model
6. Deployment of Deep Learning model

**About Dataset :**

**Context :**

* Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus.
* Most people infected with the COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment.
* Older people, and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness.
* During the entire course of the pandemic, one of the main problems that healthcare providers have faced is the shortage of medical resources and a proper plan to efficiently distribute them. I
* n these tough times, being able to predict what kind of resource an individual might require at the time of being tested positive or even before that will be of immense help to the authorities as they would be able to procure and arrange for the resources necessary to save the life of that patient.

**The main goal of this project is to build a Deep learning model that, given a COVID-19 patient's current symptom, status, and medical history, will predict whether the patient is at high risk or not.**

**Content :**

* The dataset was provided by the Mexican government [(link)](https://datos.gob.mx/busca/dataset/informacion-referente-a-casos-covid-19-en-mexico). This dataset contains an enormous number of anonymized patient-related information including pre-conditions.
* The raw dataset consists of 21 unique features and 1,048,576 unique patients. **In the Boolean features, 1 means "yes" and 2 means "no". values as 97 and 99 are missing data**.
* **sex**: 1 for female and 2 for male.
* **age**: of the patient.
* **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-19 or that the test is inconclusive.
* **patient type:** type of care the patient received in the unit. 1 for returning home and 2 for hospitalization.
* **pneumonia:** whether the patient already has air sac inflammation or not.
* **pregnancy:** whether the patient is pregnant or not.
* **diabetes:** whether the patient has diabetes or not.
* **COPD**: Indicates whether the patient has Chronic obstructive pulmonary disease or not.
* **asthma**: whether the patient has asthma or not.
* **Insurer**: whether the patient is immunosuppressed or not.
* **hypertension**: whether the patient has hypertension or not.
* **cardiovascular**: whether the patient has heart or blood vessel-related disease.
* **renal** **chronic**: whether the patient has chronic renal disease or not.
* **other** **disease**: whether the patient has other disease or not.
* **obesity**: whether the patient is obese or not.
* **tobacco**: whether the patient is a tobacco user.
* **Usmr** : Indicates whether the patient treated medical units of the first, second or third level.
* **medical** **unit**: type of institution of the National Health System that provides the care.
* **intubated**: whether the patient was connected to the ventilator.
* **icu**: Indicates whether the patient had been admitted to an Intensive Care Unit.
* **date** **died**: If the patient died indicate the date of death, and 9999-99-99 otherwise.

**Exploratory Data Analysis Report :**

**Problem statement: We need to predict whether the patient is affected by COVID-19 OR NOT.**

* **In EDA**

1. **Imputing missing values**

* Categorical = Mode
* Real Values = Mean, Interpolation, BFill, FFill etc.

1. **Detecting outliers**

* Z-Score method
* Boxplot
* Scatter plot

1. **Treatment for outliers**

* IQR Method
* Z-Score method
* K-NN Method

1. **Visualization**
2. **Univariate Analysis**

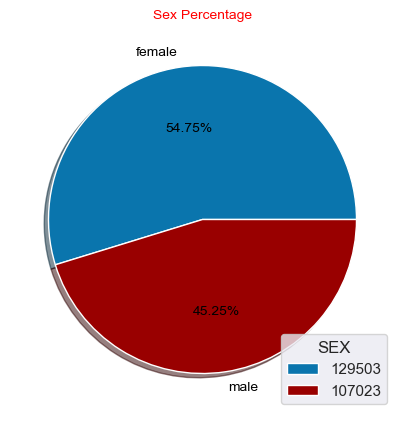
* **Numerical**
* Histogram plot
* Density plot
* **Categorical**
* Count plot

1. **Bivariate Analysis**

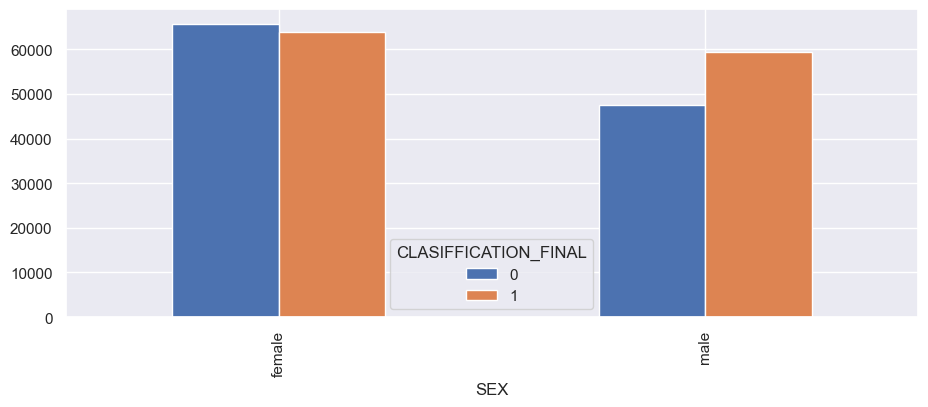
* **Numerical vs numerical**
* Scatter plot
* Histogram plot
* **Numerical vs Categorical**
* Boxplot
* **Categorical vs Categorical**
* Stacked bar

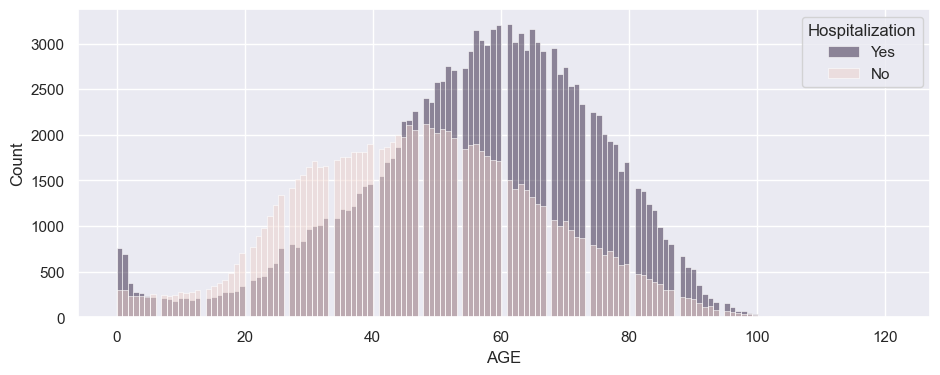
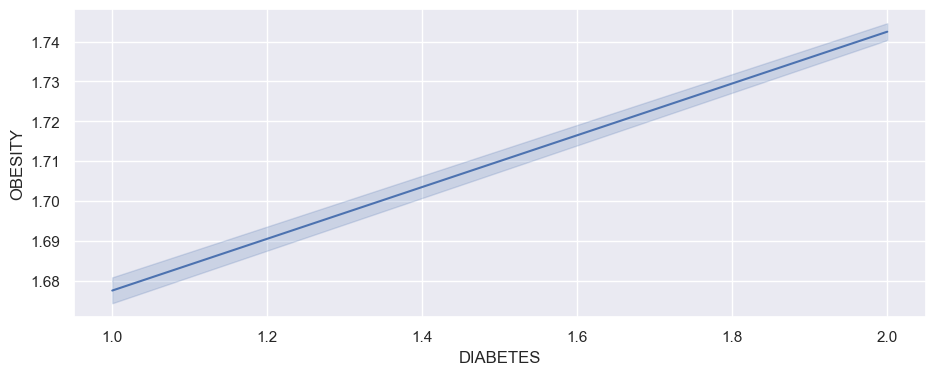
1. **Multi-Variate Analysis**

* **Heat Map**
* **Pair Plot**



* By seeing this pie chart we can observe the Gender ratio of Males and Females, so the female count is higher than the male.



* From the pivot table, all the people were affected by COVID19.
* From the histogram plot, we can see a very clear distribution of AGE and HOSPITALIZED.
* The age between from 30 to 90 have been affected with COVID19 and Hospitalized.
* This is a line plot, where diabetes increases obesity also increases.
* Means they are highly correlated to each other.

**Deep Learning:**

My data is **Imbalanced**, so I need to balance the data by using **SMOTE Techniques.**

In **Oversampling** Technique we have **Majority sampling and minority sampling**. So, we need to do **Oversampling** upon **minority sampling**, so now the number of samples is increased and now our data is balanced.

**Data Cleaning Report:**

After the completion of data cleaning, Now we are ready to send our data to Neural Network.

But, before sending our data to the neural network, we must observe our data, whether it is properly scaled or not, because our numerical data should be in the same scale then only our neural network properly works and training could be speed.

So, we must Rescale our data by using **Normalisation techniques or Standardisation techniques**, so the values will lie between 0 to 1.

This technique helps us to train our model very fast and our optimizers will run very fast.

**Model Report:**

To create the ANN Model we used the Keras Tuner. Where keras tuner can help us to return the best model corresponding to No.of.hidden layers and no.of.Neurons and which activation function to be used, which optimizer and weight initializers etc.

Our target variable is binary classification, so we used the loss function as “binary\_crossentropy”.

**Model Accuracy:**

Finally, my model accuracy is 63%.

The total hidden layers are 11, activation is tanh, weight\_init is glorot\_normal, optimizer is adadelta.