**INTRODUCTION**

This is a dummy dataset that imitates the characteristics of data collected from hospital patients during covid-19 pandemic.

*Discalimer: The dataset used for this analysis is a dummy dataset.*

**PROBLEM STATEMENT**

This dataset comprises twenty-one columns, encompassing the input column that encapsulates information regarding the severity of COVID for specific individuals. The aim of this project is to build a ML model and train the model with the available features to predict the severity of Covid in individuals. The scale of severity in individuals ranges from

Level 1: Lowest level of covid detected (asymptotic)

Level 2: Mild symptoms\*

Level 3: Moderate symptoms\*

Level 4: Moderate to severe symptoms\*

Level 5: Severe symptoms\*

Level 6: Critical symptoms

Level 7: Life threatening conditions

Now let’s dive into it

All important and useful libraries were imported, it is important to note that it is conventional to import all libraries at the start of the script in a data science project as this allows for a structured and logical workflow. This process is iterative as you would usually realize that a library that has not been imported is needed in your workflow. This is not due to poor planning, it is just natural as you come across challenges that are not envisaged in your project planning phase. Flexibility is among the core skills of a data scientist. This is then followed by a routine exploration of the dataset.

Luckily, all columns in the dataset seems to be suited for a ML model building and there are no missing values in the dataset. An observed challenge is that the responses for this dataset was created using numbers to elicit response, although a data dictionary that explains all variables exist. Due to the quantitative nature of numbers, there is a possibility that an arbitrary ordering may have been introduced in the dataset. This issue is analogous to the drawback of label encoding (Hancock & Khoshgoftaar, 2020).

**EXPLORATORY DATA ANALYSIS**

Firstly, a histogram of the patients age distribution was plotted. Majority of the patients fall between the age range of 20-70 years with a similar distribution between genders. However, the histogram is skewed to the left. Skewness is a measurement of the distortion of symmetrical [distribution](https://www.investopedia.com/terms/d/distribution.asp) or asymmetry in a data set. Skewness is demonstrated on a bell curve when data points are not distributed symmetrically to the left and right sides of the median on a bell curve (Sofaer & Strech, 2012). There are several techniques that can be used to address the issue of skewness. However, none of the technique will be applied at this instance because we intend to use non-parametric ML models. According to (Sedgwick, 2015), A statistical test that makes no assumptions about the data's underlying distribution is known as a non-parametric test. It is used when the data does not meet the assumptions of parametric tests. In addition, other advantages of such models are **Flexibility:** non-parametric models are not limited by a particular functional form, they can handle a broad range of data types and patterns and **POWER:** They can result in higher performance models for prediction (Jason, 2015). In contrast, non-parametric models can be computationally expensive, especially when dealing with large datasets. If not carefully adjusted, these models may be prone to overfitting the data, capturing noise rather than the underlying relationship (Jason, 2015).

A box plot was generated to visualize the distribution and identify outliers in the Age feature. Evidently, some outliers were observed in the Age feature, prompting the need for outlier treatment. (Beckman & Cook, 1983) defines outlier as data point that substantially deviates from other observations. An outlier may be due to a variability in the measurement, an indication of novel data, or it may be the result of experimental error. Although, this is a definition from an article written in 1983, the definition of outlier has not changed. According to (Mobarak, 2022), In machine learning, outliers cause anomaly in the models. It leads to a slight modification of the model's typical thinking from the typical pattern, which in machine learning is known as overfitting.  Overfitting occurs when the model cannot generalize and fits too closely to the training dataset instead. Overfitting can happen due to several reasons (Mobarak, 2022). In light of this, an outlier treatment was done on a copy of the dataset. However, after the outlier treatment, it was revealed that patients over the age of 70 are regarded as outliers. Considering the domain of analysis, it is not uncommon to have patients above the age of 70, also the maximum number in the age feature is also 107. These figures are not uncommon amongst human population. This is corroborated by an extract from (Robine, 2021) , A plethora of statistics suggests that adult life expectancy in the industrialised countries has increased since the mid-1900s. For instance, there has never been a higher percentage of individuals turning 100 years old than there is now. In addition, there is a total of 18085 patients who are above the age of 70. This represents about 9% of the entire dataset. If the results of the outlier is accepted, it will reduce the number of features by 9%. This is corroborated by (Jim, 2023), Given the problems that outlier can cause, one might think that it’s best to remove them from a dataset. But, that’s not always the case. **Removing outliers is legitimate only for specific reasons**. If an outlier results from a measurement error or data entry error and the observation is not representative of the population you are researching, it can be corrected; if not, it can be eliminated. In contrast, an outlier should never be removed if it is a natural part of the population under study. excluding extreme values solely due to their extremeness can distort the results by removing information about the variability inherent in the study area. You’re forcing the subject area to appear less variable than it is in reality (Jim, 2023).

As seen in the above graph, all patients whose covid condition was life threatening had almost a fifty percent chance of survival. In addition, an issue of imbalanced dataset was observed. Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e one class label has a very high number of observations and the other has a very low number of observations. When the records of a certain class are much more than the other class, our classifier may get biased towards the prediction treating other classes as noise or completely disregarding them is a drawback of imbalanced datasets (Haixiang *et al.*, 2017). The issue of imbalanced datasets will be addressed before model deployment. A further exploration of all the features against the target variable was done to better understand the relatiomship that exist between them.

**IMBALANCED DATASET TREATMENT**

A new variable, X, representing the input, and y, representing the output, were created. The dataset was then split into training and testing sets using the train-test split method, with a test size of 20% of the dataset. Since an imbalanced dataset has been previously identified, it is imperative to address the imbalanced dataset to avoid building a ML model that overfits the data. Given the identification of an imbalanced dataset, it is crucial to address this issue to prevent the development of a machine learning model that might overfit the data.

Smote sampling technique was used to address the issue of imbalanced dataset in this project. It is an oversampling technique that modifies the training data by adding artificial minority samples. Among the benefits of SMOTE are its capacity to produce additional data from the minority class, overcoming data scarcity, and enhancing classification performance (Elreedy & Atiya, 2019). However, the Smote is not a technique without its own issues. The technique may create instances in noisy and overlapping areas, far from safe regions . This can lead to the generation of instances that are not representative of the minority class and may negatively impact classification performance (Vuttipittayamongkol, Elyan & Petrovski, 2021) Keeping this in mind, the SMOTE (Synthetic Minority Over-sampling Technique) oversampling technique was employed to tackle the imbalanced dataset issue.Top of Form

**MODEL BUILDING**

Four machine learning models were utilized to construct a predictive model for assessing the severity of COVID-19 cases. These models comprise the Decision Tree Classifier, K-Nearest Neighbors Classifier, Bagging Classifier, and Random Forest Classifier. Following the utilization of the training set to train the model, it was subsequently employed to predict outcomes on the testing set. Various metrics, including accuracy, precision, and others, were then employed to evaluate the model's performance.

**HYPER PARAMETER TUNING**

Hyperparameter tuning is an essential part of machine learning and it allows data scientists to tweak model performance for optimal results (Yang & Shami, 2020).  Hyperparameter tuning is an important part of any machine learing project although it can be a computationally intensive process. Due to the dataset's large size and the computational intensity associated with hyperparameter tuning, a tuning fraction of 60% of the dataset was allocated for this purpose.

Hyper parameter tuning did not really improve the models performance except for the Decision Tree classifier which model performance was increased from 0.6616 to 0.6953.

**VOTING CLASSIFIER**

After hyper parameter tuning, the best parameters of each model was fiited on a voting classifier which trains on all the numerous models and predicts an output based on their highest probability of chosen class as the output.