**Predicting High-Risk COVID-19 Patients Using Ensembling of Machine Learning classifiers**

**Chaitanya Datta Maddukuri A20568393**

**Vamshi Krishna Cheeti A20582646**

**Mark Aiken A20607497**

**Rounak Laddha A20555466**

**Project Github -** [**https://github.com/Rounakladdha832001/CSP-571-DPA-.git**](https://github.com/Rounakladdha832001/CSP-571-DPA-.git)

**Abstract**  
COVID-19, a global pandemic caused by the SARS-CoV-2 virus, has presented unprecedented challenges to healthcare systems worldwide. Effective analysis and prediction of COVID-19 outcomes, such as mortality, are essential for strategic healthcare planning and resource allocation. In this study, we leverage a large-scale dataset provided by the Mexican government, sourced from Kaggle, which includes 1,048,576 patient records with 21 features encompassing demographics, symptoms, pre-existing conditions, and outcomes. The primary objective is to predict the mortality indicator (Death\_Indicator) and gain insights into the factors contributing to patient outcomes.

The proposed framework employs robust data pre-processing, outlier analysis and feature engineering techniques. Dimensionality reduction methods such as Principal Component Analysis (PCA) are employed to visualise the multi-dimensional data. Predictive models including Decision trees, Random forest, Extreme gradient boosting and Logistic regression are utilized for classification and analysis. To enhance model performance, hyperparameter tuning is applied, ensuring accuracy and generalization across diverse subsets of data.

The proposed approach demonstrates improved performance metrics compared to baseline models, highlighting the efficiency of ensemble learning and machine learning models in handling high-dimensional datasets. The findings from this study provide actionable insights into COVID-19 outcomes and underscore the potential of machine learning in public health decision-making during pandemics.

**Introduction**

The primary objective of this study is to leverage machine learning techniques to analyze and predict COVID-19 mortality using a large-scale dataset provided by the Mexican government and hosted on Kaggle. The dataset contains 1,048,576 patient records with 21 features, including demographics, symptoms, pre-existing conditions, and outcomes, with the target variable being the Death\_Indicator..

Ensemble learning, a technique that integrates multiple classifiers, is incorporated to enhance the model's prediction capability. Ensemble methods such as Random Forest are utilized for their ability to combine the strengths of individual classifiers, improving overall model performance and reliability. Additionally, hyperparameter optimization ensures that each model operates at its peak performance for the given dataset. In this literature, we propose a comprehensive framework that employs various machine learning classifiers and dimensionality reduction techniques to achieve accurate and interpretable predictions.

**Experimental Setup**

**Experiment 1**

To predict COVID-19 deaths, we employed machine learning classifiers such as Decision tree, logistic regression, random forest and XGBoost. The predictive methodology is illustrated in Figure 1. The process began with data acquisition and preprocessing, where we collected and cleaned the COVID-19 dataset. Subsequently, we split the dataset into training and testing sets.

A diverse set of machine learning algorithms were utilized as base classifiers, including logistic regression, decision trees, random forest, and XGBoost. These models were trained on the training set and their predictions were combined using a voting ensemble technique.

To enhance model performance, we employed hyperparameter tuning techniques like Halving Grid Search CV. Additionally, we utilized Stratified K-fold cross-validation to assess model robustness. Finally, the ensemble model was evaluated on the testing set using various performance metrics, including accuracy, precision, recall, and F1-score.



Figure 1. A figure illustrating the predictive framework for predicting COVID-19 deaths.

**Experiment 2 – Proposed Methodology**  
  
To predict COVID-19 deaths, we utilized an ensemble learning approach. We employed the Covid-19 dataset, which comprised [1048576] data samples with [21] features. The target variable indicated whether a patient would succumb to COVID-19. We conducted exploratory data analysis (EDA) to gain insights into the data distribution and relationships between features.

A diverse set of machine learning algorithms were employed as base classifiers, including logistic regression, decision trees, random forest, and XGBoost. These models were trained on the dataset and their predictions were combined using a voting ensemble technique. The architecture of the proposed ensemble model is illustrated in Figure 2.

We evaluated the performance of our ensemble model using several evaluation metrics such as Accuracy, Precision, recall and F1 score.

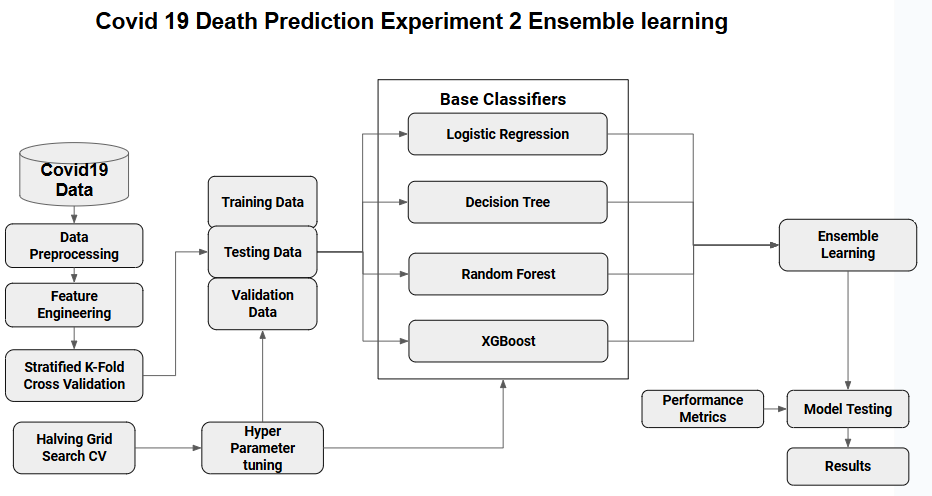


Figure 2. A figure illustrating the architectural framework for predicting COVID-19 deaths.

**Dataset Description**

The dataset was provided by the Mexican government and is publicly available on Kaggle (<https://www.kaggle.com/datasets/meirnizri/covid19-dataset/data>).It contains anonymized patient-related information, including pre-existing conditions, symptoms, and outcomes, for a total of 1,048,576 unique patients. The dataset comprises 21 features and a target variable (Death\_Indicator). Boolean features are represented as 1 for "yes" and 2 for "no," while values such as 97 and 99 indicate missing data. Table 1 summarizes the attributes and their descriptions.

The predictive features include 1 continuous variable the patients’ ages, 3 categorical variables ‘classification’ indicating covid test findings, ‘usmr’ the level of medical treatment for the patient, and ‘medical unit’ the type of unit for treating the patient. A single binary indicator variable for sex included having a 1 for female 2 for male. The rest of the predictor features (N:12) were binary indicator variables of health conditions with a 1 indicating the presence of a health condition and 2 for the absence. Health condition indicators made up the vast majority of predictors including conditions like, diabetes, tobacco usage, hypertension, and COPD. Finally there were 3 indicator variables for care received including patient type (1= sent home 2, hospitalized) admitted to the ICU , and if the patient had to be intubated. Table 1 summarises the description of the dataset.

The pre-processed response variable was the age of the patient when they died.

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Description** | **Type** | **Values** |
| Sex | 1 for female and 2 for male. | Categorical | 1,2 |
| Age | age of paitent | Continuous | 0-121 |
| Classifcation | Covid test findings. Values 1-3 mean that the patient was diagnosed with covid in different | Categorical | 1,2,3 |
| Patient Type | type of care the patient received in the unit. 1 for returned home and 2 for hospitalization | Categorical | 1,2 |
| Pneumonia | Whether the patient already have air sacs inflammation or not. | Indicator | 1=Yes, 2=no |
| Pregnancy | Whether the patient is pregnant or not. | Indicator | 1=Yes, 2=no |
| Diabetes | Whether the patient has diabetes or not. | Indicator | 1=Yes, 2=no |
| COPD | Whether the patient has Chronic obstructive pulmonary disease or not | Indicator | 1=Yes, 2=no |
| Asthma | Whether the patient has asthma or not. | Indicator | 1=Yes, 2=no |
| Imsupr | Whether the patient is immunosuppressed or not. | Indicator | 1=Yes, 2=no |
| Hypertendion | Whether the patient has hypertension or not. | Indicator | 1=Yes, 2=no |
| Cardiovascular | Whether the patient has heart or blood vessels related disease | Indicator | 1=Yes, 2=no |
| Renal Chronic | Whether the patient has chronic renal disease or not. | Indicator | 1=Yes, 2=no |
| Other Disease | Whether the patient has other disease or not. | Indicator | 1=Yes, 2=no |
| obesity | Whether the patient is obese or not. | Indicator | 1=Yes, 2=no |
| Tobacco | Whether the patient is a tobacco user. | Indicator | 1=Yes, 2=no |
| Usmr | Indicates Whether the patient treated medical units of the first, second or third level. | Categorical | 1,2,3 |
| Medical Unit | Type of institution of the National Health System that provided the care. | Categorical | 1-13 |
| Intubed | Whether the patient was connected to the ventilator. | Indicator | 1=Yes, 2=no |
| Icu | If the patient died indicate the date of death, and 9999-99-99 otherwise | Date | DD-MM-YYYY |
| Date Died | Indicates Whether the patient had been admitted to an Intensive Care Unit. | Indicator | 1=Yes, 2=no |

**TABLE 1**. Attribute information and corresponding description for COVID 19 dataset

**Data Preprocessing**

The COVID-19 dataset provided by the Mexican government underwent systematic preprocessing to ensure it was clean, accurate, and ready for analysis. Missing values, represented by codes 97, 98, and 99, were identified and replaced with NaN in relevant columns, excluding the AGE column to preserve potentially meaningful values.

A new binary target variable, DEATH\_IND, was created from the DATE\_DIED column to indicate patient mortality. This simplified modelling by replacing complex date strings with binary indicators. Variables were categorized into binary (e.g., PNEUMONIA, INTUBATED) and categorical types (e.g., CLASSIFICATION) based on their unique values.

Data distribution was visualized to identify trends, such as the prevalence of conditions and the balance in the target variable, using bar charts and other visual tools. The preprocessing pipeline ensured the dataset was complete, consistent, and structured, laying a solid foundation for predictive modelling. The correlation among the features are present in Fig. 3.

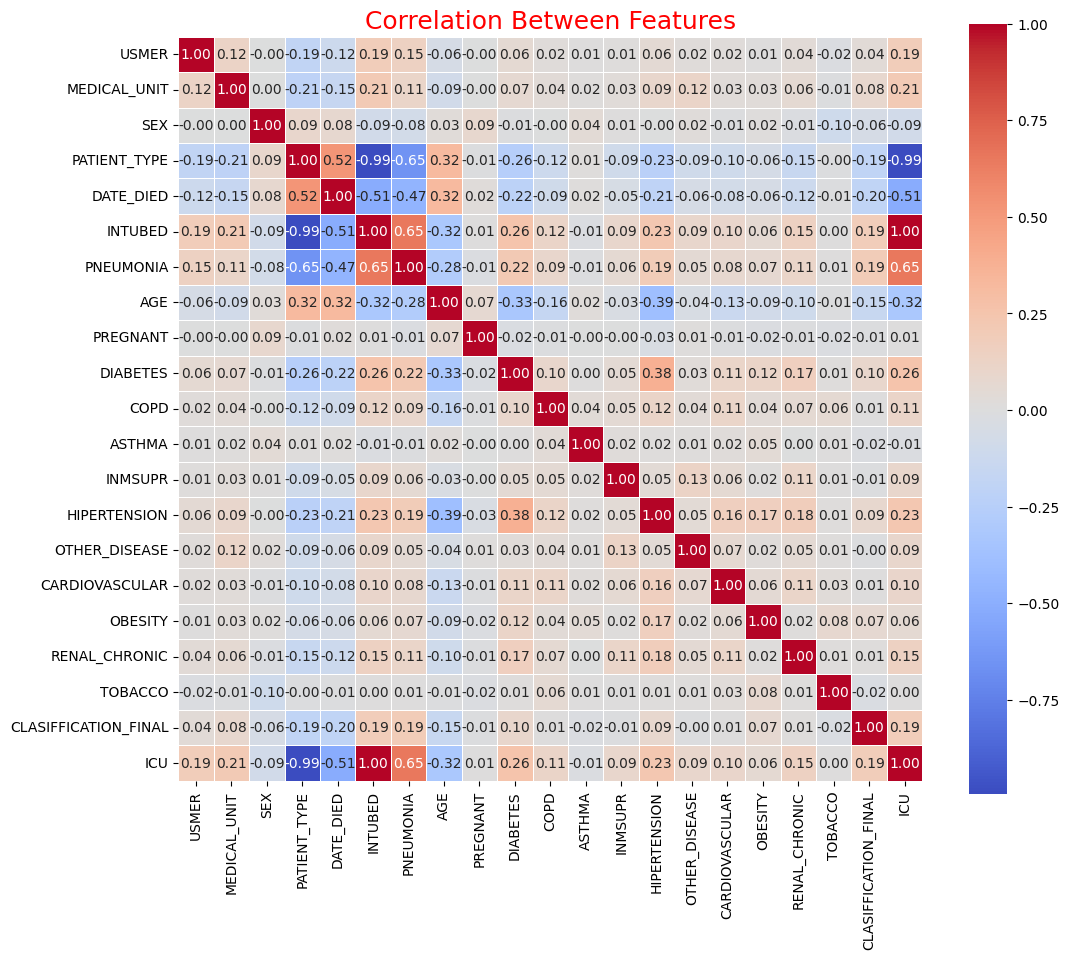


Figure 3. The correlation among the attributes.

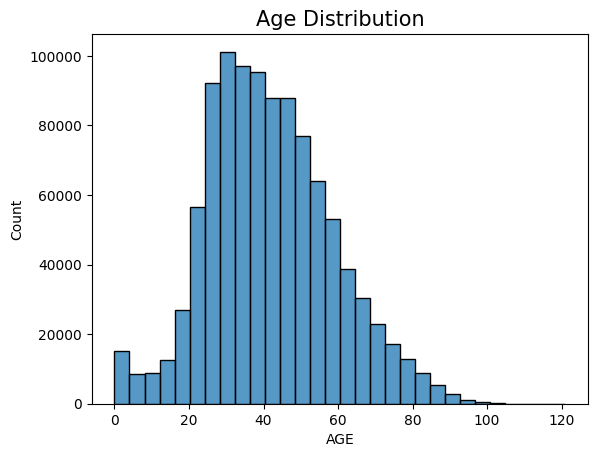


Figure 4. The frequency distribution of the AGE variable, showing its central tendency and spread

## Outlier detection & Analysis

Outliers are often considered noise among the hidden data and need to be removed or identified using an analysis of the data distributions among attributes. This may cause the model to see a sudden change in the data and may lead to inaccurate models, which can lead to inappropriate values being distributed between them. In our proposed approach, outliers are identified among the data points using boxplots, pair plots, K Means Clustering algorithm and PCA which are visualised in Fig. 7.

Outliers can disrupt predictive modelling by introducing noise, leading to inaccurate results. In this analysis, dimensionality reduction techniques like **Principal Component Analysis (PCA)** and were used to detect patterns, clusters, and potential outliers in the dataset.

First, the data was normalized using **StandardScaler** to standardize feature values with zero mean and unit variance, a crucial step since PCA is sensitive to scaling. Missing values (NaN) were removed, and the cleaned dataset was used for further processing. PCA was then applied to reduce the dimensionality of the dataset to two components (PC1 and PC2). These components were analyzed for variance explained, but results showed a low variance contribution, indicating that PCA alone might not fully capture the dataset's complexity.

This step is crucial for refining the dataset, ensuring that the machine learning model is trained on reliable and representative data for better performance. The outliers identified are removed from the dataset to achieve better performance.

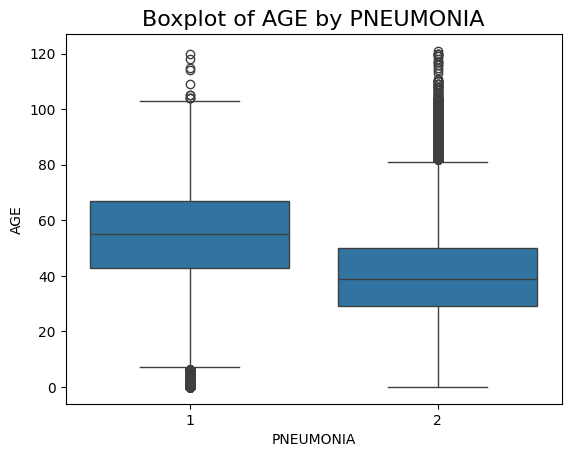


Figure 4. Boxplot of Age Grouped by Pneumonia Status*.*

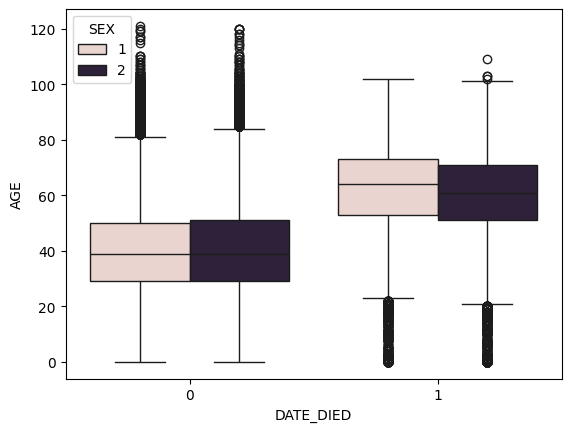


Figure 5. Box Plot Age vs Target feature

**K-Means Clustering Algorithm**

In the outlier analysis, K-Means clustering is utilized for detecting multivariate outliers among the data points. In this phase, the training data is divided into a predefined number of clusters by minimizing the variance within each cluster. The central idea of this approach is to partition the data into distinct clusters, where each cluster is represented by its centroid (mean point). Data points are assigned to the cluster with the nearest centroid, ensuring that each cluster contains points that are closer to its centroid than any other. Outliers are identified as data points that are positioned far from their assigned cluster centroid, as they deviate significantly from the typical patterns within the cluster. This method is particularly effective in identifying anomalies in datasets where the data naturally forms distinct groups.

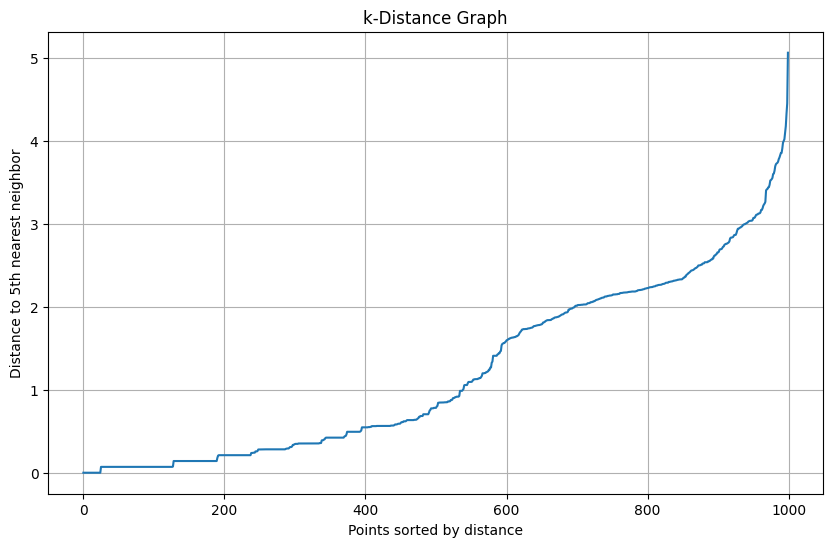


Figure 6. K-distance graph with Points sorted by distance

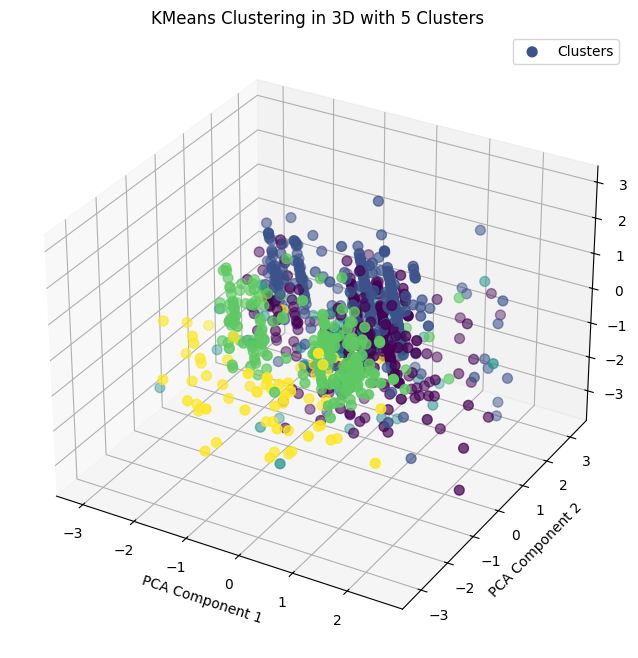


Figure 7.K-Means Clustering for PCA Component 1 and PCA Component 2

**Feature Selection**

After fitting our models using the full feature set we wished to see if we could reduce it to a subset of more relevant predictive features. For our binary indicator variables (N:15) we used a chi-squared test to see if any of the features were independent of our death outcome response variable and therefore poor choice to include in our subsequent model fits. After testing all but 2 predictors were found to be dependent with our death outcome and associated with a p-value < 0.05. The two predictors that did not meet the threshold of significance were ‘tobacco use’ and ‘pregnancy’.

For our single continuous variable, an A.N.O.V.A F-test was used to determine if the age of a patient was significantly related to our response variable death. At a 95% confidence level the test indicated that Age showed extreme significance and related to the response variable with a F-score of 18,926.43 and as such should be contained within the model. Going forward we hope that this new reduced set of predictors which no longer include indicator variables for pregnancy or tobacco use.

**Chi-Squared Test**

This method refers to a univariate filter-based feature selection method based on the common χ² statistical test, that calculates the divergence from the expected distribution only possible if we assume that the feature occurrence is independently based on the class value. In the other univariate methods, we determine the chi-square between each feature and the target variable. As a result, we observe the existence of a relationship between the feature and the target. If the target variable is independent of the feature, then the value will be lower and if they are dependent, if we get a higher value, and obviously if the score is less often, the feature is considered as important. The lower the value of chi-square results in the feature is more dependent on the target, highly correlated and important concerning the class and helps in building an effective model. This test can be conducted by the mathematical expression,

=

|  |  |  |
| --- | --- | --- |
| **Chi-Squared Test Results** | | |
| **Feature** | **Chi-SQ Score** | **P-Value** |
| **INTUBED** | 4,627,070.86 | *0.00000* |
| **ICU** | 4,594,485.78 | *0.00000* |
| **PNEUMONIA** | 14,152.52 | *0.00000* |
| **HYPERTENSION** | 3,074.85 | *0.00000* |
| **DIABETES** | 2,702.05 | *0.00000* |
| **SEX** | 1,118.20 | *0.00000* |
| **OBESITY** | 229.16 | *0.00000* |
| **RENAL\_CHRONIC** | 129.63 | *0.00000* |
| **CARDIOVASCULAR** | 58.43 | *0.00000* |
| **COPD** | 57.61 | *0.00000* |
| **OTHER\_DISEASE** | 43.86 | *0.00000* |
| **INMSUPR** | 16.41 | *0.00005* |
| **ASTHMA** | 4.45 | *0.03482* |
| **PREGNANT\*\*** | 1.72 | *0.18947* |
| **TOBACCO\*\*** | 1.09 | *0.29661* |
| **\*\* *Indicates predictor variables which were tested to be independent of response variable death.*** | | |

## Decision Tree

A Decision Tree (DT) is a supervised machine learning algorithm for solving classification tasks. The trees are constructed using a top-down recursive approach using divide and conquer (DAC). The presence or absence of diabetes is predicted by developing a tree structure in which intermediate nodes indicate a feature, each branch defines an outcome of the feature, and each terminating node impacts an existence of diabetes. The outcome label helps in identifying the diabetes. Consider a dataset with N classes. The formula can be used to calculate entropy.

## Where  is the probability of randomly selecting an example in class ‘’.

## Logistic Regression

Logistic Regression (LR) is a supervised machine learning classification approach which is found evident in solving several chronic diseases. This approach estimates the relation between the independent variables = {, ,…,} and a binary dependent variable (i.e., outcome variable) by estimating the probabilities using a sigmoid/activation function. This approach is applied only to a binary dependent variable with classes and an independent variable that can be ordinal, ratio-level and binomial. The logistic/sigmoid function is mathematically given as

where and are the weights or coefficients, is the bias or the intercept.

**Random Forest**

A Random Forest (RF) is a supervised machine learning technique that uses an ensemble learning method to include numerous decision trees. The tree learning is carried out by applying a bootstrap bagging and aggregating process to the given data, X = {, ,…,}with known labels Y = {, ,,..., } and iterated repeatedly. The process of predicting covid 19 death involves creating a variety of decision trees that take into account the class label, and each intermediary node is filled with a attribute and the output of each tree only represents one particular class label. A class label is selected from all sets of uncorrelated decision trees based on a voting criteria. The class label denotes the existence or absence of Covid 19.

## XGBoost

Extreme gradient boosting (XGBoost) is a supervised machine learning classification algorithm. The implementation of this technique is carried using gradient boosted trees with some tuning and enhancements using regularisation and loss function, column sampling. The weights are assigned to the independent variables that are to be passed to sequentially generated decision trees to produce predictive results. The weights are adjusted accordingly in a sequence and ensembled to minimise the loss using an objective function L. Mathematically, the XGBoost classifier can be represented as follows.

Where f is the functional sample space in F, F defines the sample space consisting of all possible decision trees, and K is the total number of trees. The phrase defines the differentiable convex loss between prediction and the target . The regularisation parameter , fines the model complexity with number of leaves in a tree T. For each , an independent tree structure q and leaf weight w are assigned. The phrase represents the threshold and pre-pruning is performed during optimisation to cut the tree growth and is often used for smoothening the final weights after learning to avoid overfitting.

## Ensembling classifiers using voting rule

The outcome of an ensemble model is dependent on the outcome of the base classifiers used in developing the ensemble model. This technique binds the probabilities of each prediction for each model and considers the prediction with the highest total probability. The integration of several distinct classifiers can aid in achieving better performance with the help of the soft voting rule. In this study, the soft voting rule applied, where be the classifier of an ensemble model having N distinct classifiers which are set to predict the target class from a set of possible classes , ,…,} (in our study it’s binary). For a provided input x, the classifier will result in a dimensional vector , where is an outcome of the classifier for the class label. All *N* distinct classifiers are considered with same priority unless weights are assigned for each classifier. The final classification is resulted by averaging the individual outcomes with different class label and then choosing the label with respect to the highest majority vote (hard voting). The proposed ensemble model considers the class from a total of classes that results in a maximum probability and combines them by adding all the probabilities of each class predicted by classifier.

– (1)

In this soft voting rule, is an estimate of the posterior probability obtained by the base classifier.

# **Result Discussion**

As a part of the evaluation of the model, several measures must be employed to verify the model’s performance and to find the necessary improvements that can occur to improve the overall performance. The evaluation is carried out with the confusion matrix, which has mainly four outcomes known as True Positive (TP) denotes that the actual and predicted values are true, True Negative (TN) denotes that the actual and predicted values are false, False Positive (FP) denotes that the actual and predicted values are true, and False Negative (FN) denotes that the actual and predicted values are false. The actual value refers to the existing value, and the predicted value refers to the value that is predicted by the proposed model. The measures employed for evaluation are as follows:

### **Accuracy**

The most common measure to evaluate any model that finds how accurate the model is built. It is defined as the ratio of the sum of correctly classified values to the sum of correctly classified values and misclassified values. Accuracy (ACC) is determined as ACC = (TP + TN) / (TP + FN + FP + TN).

### **Precision**

### Precision is defined as the exactness or preciseness of the classification for evaluating the positively classified data values and is considered the ratio of the positively classified values to the sum of correctly classified data.

PRE = TP/ (TP + FP) is used to calculate precision (PRE).

### **Recall**

Recall is defined as the remembering/recollection/memorizing as a measure of quality and it is said to be the ratio of the positively classified data to the sum of correctly positive and negative classified data. Recall (or) Sensitivity (or) True Positive Rate (REC, SEN, TPR) is determined as

REC = TP/ (TP + FN)

### **F1-Score**

F1-Score is defined as the harmonic mean of the precision and recall to determine the average rate. The purpose of this metric is mainly used to evaluate imbalanced datasets, which do not have a balance among those classes between features. It works well on categorical attributes. As accuracy is not a good metric since it's more biased towards the positives rather than the negatives. F1-Score can compensate for them and hence gives an unbiased result while evaluating the model. F1-Score (F1) is determined as F1 = 2 (PRE \* REC) / (PRE + REC); where PRE, REC indicates precision and recall respectively.

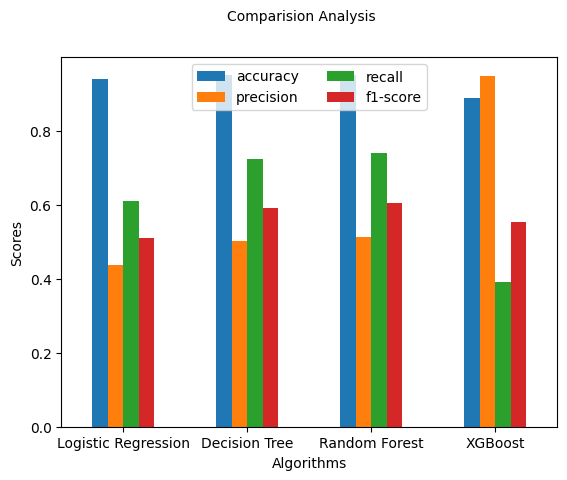


Figure 8. Comparison analysis for the models on Experiment 1

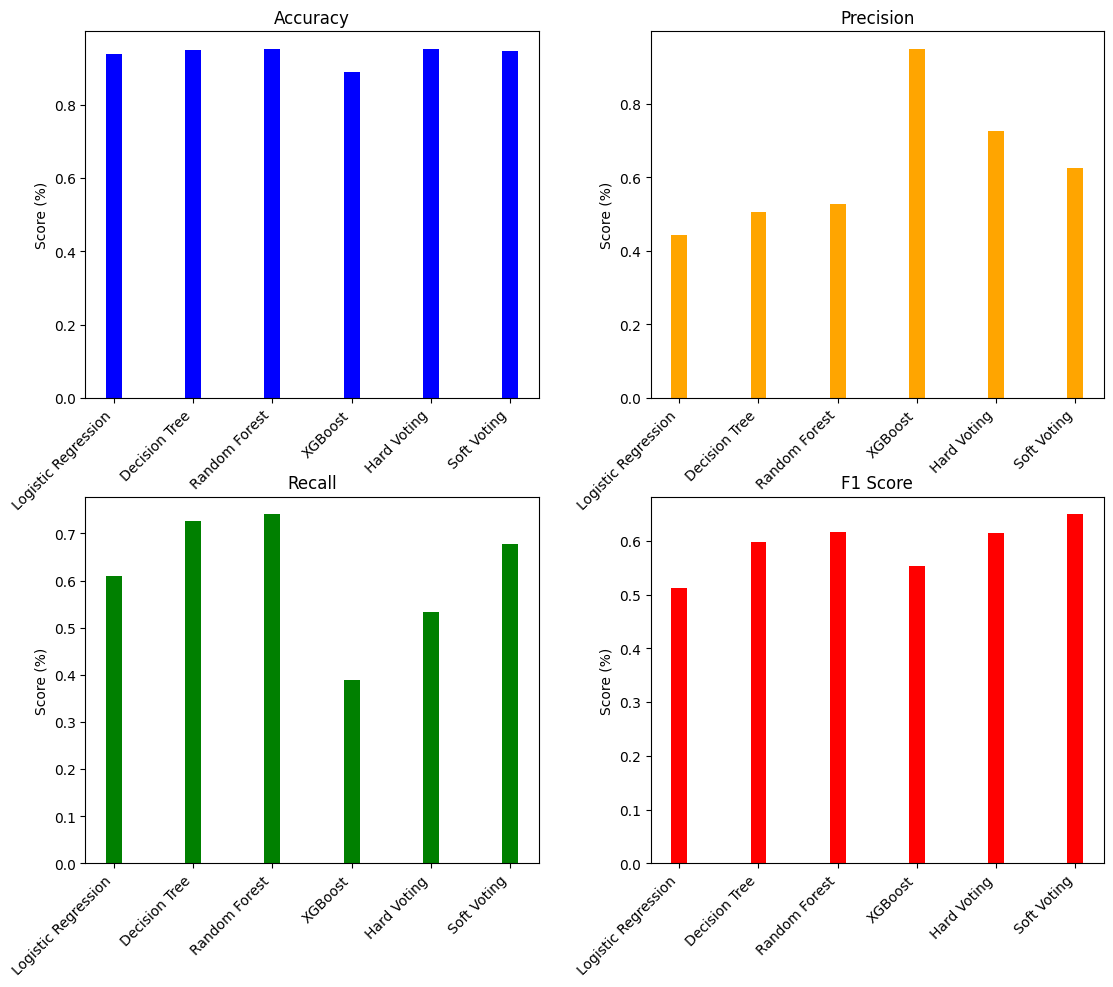


Figure 9. Comparison analysis for the models on Proposed approach