**Name: Samar Wagih Elshedody**

**ID: 211002021**

**Covid -19 Report**

[**GitHub-Link**](https://github.com/samarwagih/Covid-19-ML-project)

[**Streamlit-WebApp**](https://covid-19-ml-project-samar.streamlit.app/)

[**Presentation-Link**](https://nileuniversity-my.sharepoint.com/:v:/g/personal/s_wagih2121_nu_edu_eg/Ea-964Zd8VVOshdlY2-oBfwB_TwhW9zm05gaKGr-HOAMdg?nav=eyJyZWZlcnJhbEluZm8iOnsicmVmZXJyYWxBcHAiOiJTdHJlYW1XZWJBcHAiLCJyZWZlcnJhbFZpZXciOiJTaGFyZURpYWxvZy1MaW5rIiwicmVmZXJyYWxBcHBQbGF0Zm9ybSI6IldlYiIsInJlZmVycmFsTW9kZSI6InZpZXcifX0%3D&email=MElSayeh%40nu.edu.eg&e=LVehv1)

**Introduction**

**1.1 | About dataset**

An infectious condition known as coronavirus disease (COVID-19) is brought on by a recently identified coronavirus. Most COVID-19 virus infections result in mild to moderate respiratory disease and recovery without the need for special care. Serious sickness is more likely to strike the elderly and those with underlying medical conditions such diabetes, cancer, cardiovascular disease, and chronic respiratory diseases. One of the biggest issues facing healthcare professionals during the epidemic has been the lack of medical resources and an effective plan for allocating them. Knowing what sort of resource, a person would need in these difficult times, either before or at the time of a positive test, can be extremely helpful as they will be able to get and decide on the resources required to save that patient's life.

**1.2 | Problem statement**

The objective of this dataset is to build a predictive model for diagnosing. the model should predict whether a patient has Covid or does not have Covid based on several diagnostic measurements, including pregnancy, ICU, hypertension, renalchronic, obesity, copd, age and other diseases.

**1.2 | Description of the dataset**

This data set contains an enormous amount 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**.  
  
The features can be explained as follows:

* **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 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 sacs 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.
* **inmsupr**: whether the patient is immunosuppressed or not.
* **hypertension**: whether the patient has hypertension or not.
* **cardiovascular**: whether the patient has heart or blood vessels 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 care.
* **intubed**: 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.
  1. **| Output**
* provided table displays summarizing the performance of different machine learning models used f for classification. The table includes the following columns:
* **Model**: This column lists the names of the machine learning models evaluated.
* **Score**: This column provides the performance score for each model, presumably in terms of accuracy or another evaluation metric, on a scale from 0 to 100.
* The models and their respective scores are:

1. **Gradient Boosting Classifier**: Achieved the highest score of 91.37.
2. **XgBoost**: Scored 91.20.
3. **Random Forest Classifier**: Scored 91.15.
4. **Decision Tree Classifier**: Scored 90.85.
5. **Logistic Regression**: Scored 90.57.
6. **KNN (K-Nearest Neighbors)**: Scored the lowest at 89.72.

A screenshot of a computer

Description automatically generated

* The table is sorted in descending order based on the scores, highlighting that the Gradient Boosting Classifier performed the best among the evaluated models, closely followed by XgBoost and the Random Forest Classifier. This ranking suggests that ensemble methods (like Gradient Boosting, XgBoost, and Random Forest) tend to outperform simpler models (like Logistic Regression and KNN) for this classification task.

**1.4 | model performance**

* The provided image displays a bar chart titled "Performance Evaluation - Covid-19," comparing the performance of various machine learning models in terms of Accuracy (%) and ROC AUC (%) for classifying Covid-19 cases. The chart is organized as follows:
* **X-Axis**: Lists the different machine learning models evaluated:
  + Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbors (KNN), XgBoost, Random Forest (RF), Gradient Boosting Decision Trees (GBDT)
* **Y-Axis**: Shows the performance scores, ranging from 60% to 95%.
* **Bars**: Each model has two bars:
  + **Accuracy (%)**: Represented by light purple bars.
  + **ROC AUC (%)**: Represented by dark purple bars.

**A graph of performance evaluation

Description automatically generated**

**Analysis:**

* The Gradient Boosting Decision Trees (GBDT) model exhibits the highest performance in both accuracy and ROC AUC, indicating it is the best model for this classification task.
* XgBoost and Random Forest models also show strong performance, closely following GBDT.
* Logistic Regression performs well, with both accuracy and ROC AUC scores around 90-91%, making it a reliable model.
* Decision Tree and K-Nearest Neighbors (KNN) have slightly lower performance compared to the ensemble methods but still show decent accuracy and ROC AUC scores around 89-90%.

**1.5 | model Deployment**

The screenshot shows an application interface titled Covid-19 App, which is likely used for entering and tracking patient data related to Covid-19. The fields include:

* user
* Medical unit
* sex
* patient type
* Age
* Pneumonia
* Death
* Covid\_or\_Not

**A screenshot of a computer

Description automatically generated**

* I developed a model deployment system for a COVID-19 dataset to determine whether an individual has COVID-19 based on a variety of diagnostic measurements. These measurements include factors such as pregnancy, ICU, hypertension, chronic renal disease, obesity, chronic obstructive pulmonary disease (COPD), age, and other diseases. Utilizing Gradient Boosting Classifier algorithms, which are renowned for their high accuracy, I was able to create a robust predictive model. This model effectively analyzes the input data and provides reliable predictions on a person’s COVID-19 status, aiding in timely and accurate diagnosis.

**2 | Conclusion**

I have done the "Covid 19" project. Originally, the dataset had 1048575 records. Besides, creating 6 different models and tuning their parameters were very useful for evaluating their performance to decide which one is the most effective for predicting Covid 19 From the above results we can see that although we got the highest accuracy (about 91.37%) from Gradient Boosting Classifier algorithms, Comes next the "XgBoost" with an accuracy of 91.20% and with a recall of 91% (both classes).As mentioned above, because of the unique characteristics of the medical industry, correctly predicting Covid such as HIPERTENSION is crucial so choosing which model provided the highest value of Recall should be on the top of priority, also in the feature engineering I create new columns: enhance the predictive power of the dataset.

**2.1| Future Possible Work:**

- Betterment of results using different hyperparameters for tuning

- Implementing more models to gain better results