**Report on covid impact prediction milestones**

**Introduction**

The Covid-19 pandemic had the potential to devastate the world’s economies and populations. If left unchecked, it could have led to widespread shortages of food, water, and medical supplies; large-scale loss of life; and massive economic disruptions. The World Health Organization (WHO) currently estimates that between 50% and 90% of the population will be affected by this pandemic at some point in time. The impact varies significantly from country to country, with some likely to experience more severe consequences than others. Countries in Africa, Southeast Asia, and the Caribbean are particularly at risk for severe infection and poverty outcomes due to weak health systems. In developed countries, such as the United States and Europe, relatively few people are expected to become infected with Covid-19 – but even these populations may experience significant setbacks if access to healthcare is limited or disrupted. In light of these risks, governments around the world have responded swiftly by mobilizing emergency response teams, coordinating domestic vaccination programs, beefing up security measures, and providing humanitarian assistance.

**Project Background**

The dataset that we are analysing has data regarding Covid cases in different countries of the world. It also has data regarding the impact on GDP, unemployment of 170 countries in the world. We will use this data to understand which are the countries that have been most badly effected due to the corona virus. We will plot visualisations that will shed light on different aspects of how countries have been affected by this virus. We will also build different machine learning models to predict the total cases of corona virus in different countries based on the features provided. We will select the best model based on performance metrices.

**Milestone 1:**

**Data**: The World Bank provides a dataset on COVID-19-related economic indicators, such as GDP growth, unemployment rates, and poverty rates.

Datasets:

1. Training Dataset

— 40334 samples of data (80%)

— 8 unique features

1. Test Dataset

— 10083 samples (20%)

— 8 unique features

Text

Description automatically generated

Preprocessing the data: your COVID-19 impact prediction project. The preprocessing steps can vary depending on the specific dataset and the objectives of your project, but here are some general steps that you can follow:

1. Data Cleaning: Remove any duplicate or irrelevant data, and handle missing or incorrect values. You can use techniques such as imputation or deletion to handle missing values, and use domain knowledge to determine which data points are outliers and should be removed.

• As per the analysis the total missing values or the NA values are in total cases and in total deaths columns

• After the time series analysis we found that the deaths are almost similar so we have

**Text

Description automatically generated**

**Data Modelling:**

* A good prediction is one that is accurate in the sense that it predicts the right thing for most or all instances.
* One common goal of machine learning is to automate some aspects of data analysis.

it is the process of creating a visual representation of either a whole information system or parts of it to communicate connections between data points and structures.

**Graphical user interface, text, application, email

Description automatically generated**

1. **Feature Selection**: Select the most relevant features that will be used in the machine learning models. You can use techniques such as correlation analysis, feature importance analysis, or dimensionality reduction to identify the most relevant features.

2. **Feature Scaling**: Scale the selected features to ensure that they are on a similar scale, which can improve the performance of the machine learning models. Common scaling techniques include normalization and standardization.

3. Data Encoding: Convert categorical variables into numerical variables, which can be used in machine learning models. Common encoding techniques include one-hot encoding and label encoding.

4. Data Splitting: Split the dataset into training and testing sets to evaluate the performance of the machine learning models. You can use techniques such as cross-validation to improve the reliability of the model performance evaluation.

5. Feature Engineering: Create new features that can be used in machine learning models based on domain knowledge. This step can involve transforming existing features, creating interaction terms between features, or creating new features based on external data sources.

By following these steps, you can prepare the dataset for use in the machine learning models in your COVID-19 impact prediction project.

**Training data model**: training a machine learning model for COVID-19 impact prediction involves selecting the appropriate algorithm, splitting the data, training the model, evaluating its performance, fine-tuning it, and deploying it for use. Choose a machine learning algorithm that is suitable for your specific problem and dataset. Depending on the problem, you may choose from classification, regression, or clustering algorithms, among others.

1. **Split the data into training and testing sets**: Split the preprocessed dataset into two parts: a training set and a testing set. The training set is used to train the machine learning algorithm, while the testing set is used to evaluate the performance of the model.

2. **Train the machine learning algorithm**: Use the training set to train the machine learning algorithm using the selected algorithm and appropriate parameters. The algorithm will learn the patterns in the training data, which will be used to make predictions on new, unseen data.

3. **Evaluate the performance of the model**: Use the testing set to evaluate the performance of the trained model. This involves predicting the target variable using the testing set and comparing the predictions with the actual values. Common performance metrics include accuracy, precision, recall, and F1-score.

4. **Fine-tune the model**: If the performance of the model is not satisfactory, you can fine-tune the algorithm by adjusting the hyperparameters, modifying the feature selection or engineering process, or using a different algorithm altogether.

5. **Deploy the model**: Once you are satisfied with the performance of the model, you can deploy it to make predictions on new, unseen data. You can use the model to make real-time predictions or integrate it into a larger system for decision-making.

**Test the data**: Use the training set to train the machine learning algorithm using the selected algorithm and appropriate parameters. If the performance of the model is not satisfactory, you can fine-tune the algorithm by adjusting the hyperparameters, modifying the feature selection or engineering process, or using a different algorithm altogether.

**Milestone 2:**

**ML algorithms**: There are several machine learning algorithms that can be used for COVID-19 impact prediction. Some common algorithms are:

Linear regression: This algorithm is used for predicting a continuous target variable. It assumes that the relationship between the features and the target variable is linear.

Logistic regression: This algorithm is used for predicting binary outcomes (e.g., positive or negative diagnosis for COVID-19). It assumes a linear relationship between the features and the log odds of the outcome.

Random forest: This algorithm is used for both classification and regression tasks. It works by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

Support vector machines (SVMs): This algorithm is used for classification tasks. It works by finding the hyperplane that maximally separates the classes in the feature space.

1. Gradient boosting: This algorithm is used for both classification and regression tasks. It works by iteratively adding weak learners (e.g., decision trees) to the model, with each new learner correcting the errors of the previous ones.

2. Neural networks: These algorithms are used for both classification and regression tasks. They work by simulating the structure and function of the human brain, with multiple layers of interconnected neurons that can learn complex patterns in the data.

**Offline results**: Offline results refer to the performance of a machine learning model on a dataset that is separate from the dataset used for training and testing. This is sometimes referred to as a validation set or holdout set.

Offline results are important because they help to assess the generalization performance of the machine learning model. In other words, they help to determine how well the model will perform on new, unseen data. The idea is that if the model performs well on the validation set, it is likely to perform well on new data.

To obtain offline results, a portion of the available data is set aside and not used during the training phase. This data is used to evaluate the model after training is complete. The performance of the model on the validation set can then be compared to its performance on the training set. If the performance is significantly worse on the validation set than on the training set, this is a sign of overfitting, which means the model has learned to fit the training data too closely and is not generalizing well to new data.

It is important to note that offline results are not always a perfect indicator of real-world performance. In some cases, the offline results may be overly optimistic, and the model may not perform as well when deployed in the real world. Therefore, it is important to also test the model in real-world scenarios and continually monitor its performance to ensure it is still accurate over time.

**Milestone 3:**

Linear regression is a simple algorithm that can be used to model the relationship between a dependent variable (e.g., the number of COVID-19 cases) and one or more independent variables (e.g., time, temperature, population density

Text

Description automatically generated

Text

Description automatically generated

Decision trees are a type of algorithm that can be used for classification or regression tasks. They have been used to predict the severity of COVID-19 cases, as well as to identify risk factors for infection and mortality.

Text

Description automatically generated

Text

Description automatically generated

SVMs are a type of algorithm that can be used for classification or regression tasks. They have been used to predict the spread of COVID-19, as well as to predict the severity of cases and patient outcomes.

Text

Description automatically generated

Text

Description automatically generated

XGBoost has been used to predict the number of COVID-19 cases and deaths in different regions, to identify high-risk groups and locations, and to analyze the effectiveness of various interventions. XGBoost has also been used in conjunction with other machine learning algorithms to develop models that can predict the spread of COVID-19 and help public health officials make informed decisions about how to manage the pandemic.

Overall, XGBoost has proven to be a powerful tool for analyzing the impact of COVID-19 and developing effective strategies to mitigate its spread. However, it is important to note that the effectiveness of XGBoost and other machine learning algorithms depends on the quality of the data used to train and test the models, as well as the assumptions and limitations of the models themselves. Therefore, caution should be taken when interpreting the results of these models and applying them to real-world situations.

Text

Description automatically generated

Chart, scatter chart

Description automatically generated

We have built models using linear regressor, Decision tree regressor and SVM regressor to predict the total covid cases. The performance metric we are considering is R2 square. Since the Decision tree model has very high value of R2 score it is the best model for this prediction task.

**Comparison of Training and Testing Accuracy and Run Time**

Chart, bar chart

Description automatically generated