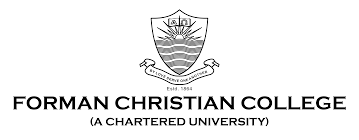
******

***AI Project Deliverable 4***

***By:***

***Iqra Amin (221441560)***

***Idwah Sajjad (221440716)***

***Sumera Shafi (231452028)***

***Submitted to: Dr Aasia Khanum***

***Course: COMP 360 – B***

***Date: July 17, 2021***

***Evaluation Results***

In our project we are predicting the spread of covid-19(cases) using regression algorithms by integrating socio-economic and demographic factors.

***Preprocessing data***

We are using two data sets files. One of the files is of cases that contains data related to covid-19 cases of different countries with some other details. The other file is of demographic data which has specifics related of the socio-economic factors. Both files contain important variables which are being used in our project. These files are being input in our code before the actual processing of the software. We read data from these two files.

***Code Input***

We selected a specific date i.e.- “2020-03-23” for the data input from cases. The reason behind this is that we are facing some limitations which are explained later. We are only taking one date from the cases file, and we are taking other factors as input from demographic file for the year 2018. We are taking these two combinations for the process. We store these all particulars as merger to see at this specific date and location how many cases and other details were there. We save this in covid’s data frame. We are making a column of 2018’s factors from demographic file as covid data’s frame list. We are doing this in a loop. This as whole is going as input in our project.

***Datasets and limitations of Datasets***

***Cases Data:***

The data we are going to use through this link displays the total number of cases, deaths due to COVID, tests for corona diagnosis, and vaccinations for corona being done all over the world till to date. The data is collected from various authentic organizations including “World Bank World Development Indicators, sourced from Food and Agriculture Organization and World Bank estimates”, “National government reports”, “European CDC for European countries / UK Government / HHS for the United States”, “COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University”, “Our World in Data”, “UN Population Division, World Population Prospects, 2017 Revision” and many more. These data will help us analyzing different numbers of cases and how they are a component of spread of COVID. We will be using these to make our ML algorithm more precise and accurate.

***Demographic data:***

It is one of the most important facts and figures to our project. It has the specifics of socio-economic Indicators. The data contains the economic statistics of different countries around the world from 2014 to 2019. It will assist us in our project when we will be comparing each countries’ economic particulars and COVID cases. There are different factors for which the data is available for different countries such as total population of each country for every year (population growth annual). The more helpful is the fact that it has explicit data for different age groups as it can help in finding out which age group is being affected the most and many more contributing factors will be used through this dataset. The input of data is genuine and taken from “United Nations, Department of Economic and Social Affairs, Population Division (2019). World Population Prospects: 2019 Revision”, for some special cases, population data are derived from “Eurostat (Demographic Statistics) or National Statistical Offices”, “The World Bank. World Development Indicators: December 2020”,” International Monetary Fund”, “World Economic Outlook database: October 2020, and United Nations Statistics Division, National Accounts Main Aggregates Database (December 2020).

***Limitations:***

Since Covid-19 is still a very new phenomenon, the data of cases available per country is still a few months old. We barely have data per day for a year. Ideally for a machine learning algorithm to give accurate results we need data of worth at least some years. But since we cannot take input of data per year, we will be taking data per day.

***Testing and Training Split Ratio***

Like in any machine learning algorithm we have testing and training data. So, we split

80% of our data into training data and 20% of our data into testing data. Training set is the one on which we train and fit our model basically to fit the parameters whereas test data is used only to assess performance of model. Training data's output is available to model whereas testing data is the unseen data for which predictions have to be made, it helps us to validate the progress of the algorithm's training and adjust or optimize it for improved results.

***Machine Learning Algorithm for the Project***

We used Regression algorithms for our project, since Regression analysis consists of a set of machine learning methods that allow us to predict a continuous outcome variable (y) based on the value of one or multiple predictor variables (x), and for our project we needed output as numeric and not as classes so, we deemed it best for the outcome we were trying to achieve. Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum.

***Data Models***

We used 3 models for predicting covid cases:

***Linear Model***

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is used for finding out the relationship between variables and forecasting. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). The regression line is the best fit line for our model.

***Decision Tree***

Decision Trees are a type of Supervised Machine Learning where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

While making decision tree, at each node of tree we ask different type of questions. Based on the asked question we will calculate the information gain corresponding to it.

Information gain is used to decide which feature to split on at each step in building the tree. Simplicity is best, so we want to keep our tree small. To do so, at each step we should choose the split that results in the purest daughter nodes. A commonly used measure of purity is called information. For each node of the tree, the information value measures how much information a feature gives us about the class. The split with the highest information gain will be taken as the first split and the process will continue until all children nodes are pure, or until the information gain is 0.

***Neural Networks***

Machine learning algorithms can improve without being explicitly programmed. In other words, they are able to find patterns in the data and apply those patterns to new challenges in the future. [Neural networks](https://www.ibm.com/cloud/learn/neural-networks)—and more specifically, artificial neural networks (ANNs)—mimic the human brain through a set of algorithms. At a basic level, a neural network is comprised of four main components: inputs, weights, a bias or threshold, and an output. Neural networks learn (or are trained) by processing examples, each of which contains a known "input" and "result," forming probability-weighted associations between the two, which are stored within the data structure of the net itself. The training of a neural network from a given example is usually conducted by determining the difference between the processed output of the network (often a prediction) and a target output. This difference is the error. The network then adjusts its weighted associations according to a learning rule and using this error value. Successive adjustments will cause the neural network to produce output which is increasingly similar to the target output. After a sufficient number of these adjustments the training can be terminated based upon certain criteria. This is known as [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning).

Such systems "learn" to perform tasks by considering examples, generally without being programmed with task-specific rules.

***Errors***

There are different error detecting techniques that can be used in machine learning for the precision checking of the algorithm. We are using different techniques for this purpose. These are listed below:

***Root Mean Square Error (RMSE)***

RMSE is calculated as the square root of the mean of the squared differences between actual outcomes and predictions. Squaring each error forces, the values to be positive, and the square root of the mean squared error returns the error metric back to the original units for comparison

***Mean Absolute error***

Mean absolute error refers to the magnitude of difference between the prediction of an observation and the true value of that observation. MAE takes the average of absolute errors for a group of predictions and observations as a measurement of the magnitude of errors for the entire group.

***Mean Absolute percentage error***

Mean Absolute Percentage Error (MAPE) is a statistical measure to define the accuracy of a machine learning algorithm on a particular dataset. MAPE can be considered as a loss function to define the error termed by the model evaluation.