**Group Project Proposal**

Group 66

Tingzhou Wan 22329065

Jiaming Deng 22302794

Yifan Zhu 22301710

**Motivation:**

A group of viruses known as the coronaviruses are called for their spiky crowns. A infectious respiratory virus with the moniker SARS-CoV-2, the novel coronavirus was first discovered in Wuhan, China. The sickness brought on by the novel coronavirus was given the designation COVID-19 by the World Health Organization. Through data analysis and forecasts using machine learning techniques, COVID-19 will be investigated in this research.

**Dataset:**

We will use a dataset which is the daily number of people infected with Covid-19 and our mission is to predict the number of covid-19 infections per day in Irish cities in the next year. The raw data of Covid-19 is collected from the health protection surveillance center website. We use the crawler technology to store the data of the number of Covid-19 infected people on the web page in a file in the csv format.

**Method:**

Model: kNN, logistic regression, SVM, decision tree, naive bayes, Lasso Regression,

train and evaluate: cross-validation, confusion matrix, standard deviation, f1 score

To forecast the COVID-19 pandemic in the next year, we want to use the techniques of KNN, logistic regression, SVM, decision tree, naive bayes, and lasso regression. To validate the results of the models, cross-validation, confusion matrix, standard deviation and f1 score are selected as the tools.

KNN: The k-nearest neighbors’ algorithm, sometimes referred to as KNN or k-NN, is a supervised learning classifier that employs proximity to produce classifications or predictions about the grouping of a single data point.

Logistic Regression: Since the results of COVID-19 prediction are categorical and Logistic Regression is used when the dependent variable(target) is categorical, the model is selected.

SVM: SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable.

Decision tree: Decision tree algorithm is a method of approximating discrete function values, which can effectively prevent the occurrence of overfitting.

Naive Bayes: Naive Bayes is a kind of classifier which uses the Bayes Theorem. It predicts membership probabilities for each class such as the probability that given record or data point belongs to a particular class.

Lasso Regression: Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean.

**Intended Experiment:**

In this project, we use the above six models to predict the future development of Covid-19. To evaluate the accuracy of our machine learning algorithms, we would like to use the methods of cross-validation, confusion matrix, standard deviation and f1 score and keep comparing this model, and finally selecting the most suitable model to predict the target value. Besides, for each model, we need to adjust the parameter of the model method, and the ratio of the test set and training set, making the model more accurate for data training and prediction.