

# Economic Crisis Prediction due to Pandemic Outbreak using Machine Learning

by

Tanvir Ahmed

17301054

MD. Nahid Hasan

17301096

Md.Ashik

17301131

Md. Jahid Hasan

17301227

A thesis submitted to the Department of Computer Science and Engineering  
in partial fulfillment of the requirements for the degree of  
B.Sc. in Computer Science

Department of Computer Science and Engineering  
Brac University  
January 2021

© 2021. Brac University  
All rights reserved.

# Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

## Student's Full Name & Signature:



---

Tanvir Ahmed  
17301054



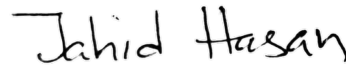
---

MD. Nahid Hasan  
17301096



---

Md. Ashik  
17301131



---

Md. Jahid Hasan  
17301227

# Approval

The thesis/project titled “Economic Crisis Prediction due to Pandemic Outbreak using Machine Learning” submitted by

1. Tanvir Ahmed (17301054)
2. MD. Nahid Hasan (17301096)
3. Md.Ashik (17301131)
4. Md. Jahid Hasan (17301227)

Of Fall, 2020 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 15, 2021.

## Examining Committee:

Supervisor:  
(Member)

---

Mohammad Zavid Parvez, PhD  
Assistant Professor  
School of Data and Sciences  
Department of Computer Science and Engineering  
Brac University

Thesis Coordinator:  
(Member)

---

Md. Golam Rabiul Alam, PhD  
Associate Professor  
School of Data and Sciences  
Department of Computer Science and Engineering  
Brac University

Head of Department:  
(Chair)

---

Mahbubul Alam Majumdar, PhD  
Professor and Dean  
School of Data and Sciences  
Department of Computer Science and Engineering  
Brac University

## Abstract

Since pandemic disease outbreaks are causing a major financial crisis by affecting the worldwide economy of a nation, machine learning techniques are urgently required to forecast and analyze the economy for early economic planning and growth and to resettle it. A large number of studies have shown that the spread of the disease has experienced a significant change in the economy. As a consequence, we will use machine learning to construct early warning models for economic crisis prediction. This paper used a publicly available data-set containing information about National Revenue, Employment Rate, and Workers Earnings of USA over 239 days (1 January 2020 to 12 May 2020). Then, we applied Multilayer Perceptron (MLP) Neural Network and Random Forest classifier to identify recession in revenue, employment rate, and workers earnings, therefore, got accuracy 95%, 81%, 89% and 85%, 81%, 89% respectively. To analyze how much the economy affected by this pandemic, we drive revenue, employment rate, and workers earnings data-set into Long short-term memory (LSTM) and Random Forest Regressor hence got accuracy 92%, 90%, 90%, and 95%, 93%, 93% respectively. Before a country faces an economic recession, it is important to identify which sector to emphasize to minimize this unexpected scenario. Using machine learning, we analyzed the data and predicted the economy so we could help save a significant amount of capital for a country.

**Keywords:** Multilayer perceptron, Long short-term memory, Random forest, Economic recession, Machine learning, Pandemic, Covid-19

## **Acknowledgement**

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.

Secondly, to our supervisor Dr. Mohammad Zavid Parvez sir for his kind support and advice in our work. He helped us whenever we needed help.

And finally to our parents without their throughout support it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

# Table of Contents

Declaration	i
Approval	ii
Abstract	iii
Acknowledgment	iv
Table of Contents	v
List of Figures	vii
List of Tables	viii
Nomenclature	ix
<b>1 Introduction</b>	<b>1</b>
1.1 Problem Statement . . . . .	2
1.2 Aims and Objective . . . . .	2
1.3 Overview of Contents . . . . .	2
<b>2 Background Study</b>	<b>3</b>
2.1 History of Pandemic Records . . . . .	3
2.2 Economic impact . . . . .	4
2.3 Algorithm Description . . . . .	6
2.3.1 Neural Network . . . . .	6
2.3.2 Random Forest . . . . .	10
<b>3 Literature Review</b>	<b>12</b>
<b>4 Proposed Material and Methods</b>	<b>15</b>
4.1 Data Description . . . . .	15
4.2 Data Preprocessing . . . . .	16
4.3 Implementing Classifier in Economic Recession Prediction Model . . .	16
4.3.1 Multi-layer Perceptron Classifier . . . . .	17
4.3.2 Random Forest Classifier . . . . .	18
4.4 Implementing Regression in Economic Recession Prediction Model . .	20
4.4.1 Recurrent Neural Network . . . . .	20
4.4.2 Random Forest Regressor . . . . .	21

<b>5</b>	<b>Result and Discussion</b>	<b>22</b>
5.1	Result of Classifiers . . . . .	22
5.1.1	MLP . . . . .	22
5.1.2	RFC . . . . .	25
5.2	Result of Regressors . . . . .	27
5.2.1	RNN . . . . .	28
5.2.2	RFR . . . . .	29
5.3	Discussion . . . . .	31
<b>6</b>	<b>Conclusion and Future research</b>	<b>33</b>
6.1	Conclusion . . . . .	33
6.2	Future Work . . . . .	33
	<b>Bibliography</b>	<b>37</b>

# List of Figures

2.1	Wages received by workers and new confirmed test cases . . . . .	5
2.2	Employment level and new confirmed test cases . . . . .	5
2.3	National employment rate, workers' wages and net revenue . . . . .	6
2.4	Multi-layer perception neural network . . . . .	7
2.5	Long Short Term Memory . . . . .	9
2.6	Random forest classifier . . . . .	10
2.7	Random Forest Regressor . . . . .	11
4.1	architecture of proposed model . . . . .	15
4.2	Random forest model . . . . .	18
5.1	Results from MLP classifier before and after cross validation (CV) . .	23
5.2	ROC of national revenue using MLP . . . . .	23
5.3	ROC of employment rate using MLP . . . . .	24
5.4	ROC of employer's earnings rate using MLP . . . . .	24
5.5	Results from Random Forest classifier before and after cross validation (CV) . . . . .	25
5.6	ROC of national revenue using random forest . . . . .	26
5.7	ROC of employment rate using random forest . . . . .	26
5.8	ROC of employer's earnings rate using random forest . . . . .	27
5.9	Comparison of real and modeled (LSTM) data of worker earnings of all sectors. . . . .	28
5.10	Comparison of real and modeled (LSTM) data of employment rate of all sectors. . . . .	28
5.11	Comparison of real and modeled (LSTM) data of national revenue all sectors. . . . .	29
5.12	Comparison of real and modeled (Random Forest Regression) data of worker earnings of all sectors. . . . .	30
5.13	Comparison of real and modeled (Random Forest Regression) data of employment rate of all sectors. . . . .	30
5.14	Comparison of real and modeled (Random Forest Regression) data of national revenue all sectors. . . . .	30



# List of Tables

5.1	Results from MLP classifier (CV=cross validation) . . . . .	22
5.2	Results from Random forest classifier . . . . .	25
5.3	Results from LSTM regression . . . . .	29
5.4	Results from Random forest regression . . . . .	29

# Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

$\epsilon$  Epsilon

$v$  Upsilon

*CV* Cross-validation

*GDP* Gross domestic product

*LSTM* Long Short-Term Memory

*MLP* Multilayer perceptron

*RFC* Random Forest Classifier

*RFR* Random Forest Regressor

*RNN* Recurrent Neural Network

*SARS* Severe acute respiratory syndrome

# Chapter 1

## Introduction

During a pandemic outbreak caused by contagious organisms, an economic forecast has a direct effect on the decisions taken by financial institutions and a wrong decision may have significant adverse effects. As a result, in finance, economic prediction plays a significant role, contributing to the creation of various models of the economic forecast using machine learning. Over the last few decades, multiple pandemic outbreaks have taken place including Spanish Flu (1918-1919) [44] which is estimated to have infected 500 million people and killed around 20 million people causing a severe impact in economic sectors, Asian Flu (1957-1958) which was first identified during 1957 and eventually spread throughout the world killing about two million people worldwide causing a depletion in economy throughout the world, SARS coronavirus (2003) which caused respiratory illness eventually reported to have killed 774 people creating a huge negative impact on human health as well as the economy, Swine Flu (2009) which was caused by a strain of influenza virus resulting the death of 150,000 to 575,000 people causing a depletion of stock markets, tourism, food as well as transportation industries incurring a huge loss and lastly the novel Coronavirus (COVID-19) (2019-present) [42] causing the death of 1.6 million people till now and completely destroying the economy throughout the world. Significant numbers of countries have already gone through a financial crisis after a pandemic outbreak. For instance, the SARS outbreak hit many Asian countries' economies [9] and took millions of lives. The economic costs from a global disease go beyond the direct damages incurred in disease-inflicted countries' affected sectors. It is not only because the disease spreads rapidly through countries across networks linked to global travel, but also because, through expanded trade and financial ties linked to globalization, any economic shock to one country spreads quickly to other countries. Infectious diseases likely to increase global cost. As a result, it is essential to take the early initiative to revive the economy. This paper provides a method for the prediction of the economy during a pandemic outbreak. In particular, we examine the techniques that have been used, examine the experiments that have been conducted, and consider directions of future work from the perspective of machine learning [19]. Our model will capture many crucial linkages across different economic sectors, thereby offering a more comprehensive measure of disease-related costs.

## 1.1 Problem Statement

Millions of livelihoods around the globe are affected by the economic instability created by the pandemic. Everyone fights a longstanding battle against the microorganism, which cause the pandemic and affects people's livelihood throughout the world. However, such pandemics are a problem of an entirely different nature and one that would demand an unparalleled scale of response. To respond to the imminent challenges to the long-term impact on the country's economy due to a pandemic, machine learning technique to forecast economic crisis will help the public and private sector leaders as well as the policymakers plan a better solution to combat the economic crisis caused by the pandemic.

## 1.2 Aims and Objective

The paper aims at defining robust financial crisis predictors. In our thesis, we found the link between a covid-19 pandemic and its impact on people's revenue, earnings, and employment. Our dataset consisting of different sectors revenue, earnings, and employment helped us find how a pandemic affects the economy. In comparison, this paper's uniqueness lies with the variables that are commonly co-related with the novel covid-19 pandemic. Overall this paper contributes to decision-makers and leaders to predict the economic crisis during a pandemic and give time to introduce measures which can eliminate or dampen a crisis entirely.

## 1.3 Overview of Contents

The remainder of this paper is organized as follows,

- \* Chapter 2 : Background study
- \* Chapter 3 : Literature Review
- \* Chapter 4 : Explanation of our proposed methods.
- \* Chapter 5 : Experimental results and discussions based on a comparative study.
- \* Chapter 6 : Conclusion and Future Works.

# Chapter 2

## Background Study

In this chapter, we will discuss how the pandemic outbreak influences a country's economy and what kind of algorithm we will use to detect this issue. We will first give an overview of previous pandemic records that occurred in a different time then how much losses a country faced due to those pandemics. However, we will also include a generalized description of a few algorithms we will use in our model.

### 2.1 History of Pandemic Records

The most common outbreak of the virus is a pandemic, which is the result of human-to-human transmission. Many diseases have been reported and recorded in the past, including SARS, Hong Kong flu, Spanish flu, Ebola, H7N9, Zika, virus, and currently a new coronavirus pandemic. Pandemic associated outbreaks have had tremendous negative impacts on national and global populations' health, environment, culture, and stability. They also caused massive political and social unrest. The pandemic's globally recognized concept is simple and known as an outbreak arising worldwide or over a diverse range crossing international frontiers, typically affecting a significant number of people. Since the 1500s or about every 10 to 50 years, influenza pandemics have occurred about three times in a century. Every pandemic threatens human survival and economic growth. The outbreak of infectious diseases will quickly cross the border, threatening financial and regional stability. The pandemic has a wide variety of negative social and financial and political ramifications, and not only has detrimental and sometimes tragic effects on those directly involved. Pandemic poses a huge danger not only to the world's people but also to the economy. The effect of financial losses will add to economic uncertainty. The impact is due to direct costs, long-term burdens and indirect costs. The direct costs of coping with the spread of the disease can be very high. Furthermore, Ebola epidemic [28] has severely disrupted economics across West Africa, and Sierra Leone in 2015 cost USD 6 billion in direct costs (hospitals, personnel, medicines) as well as economic losses of USD 1.6 billion have been estimated. Losing those who have died is one of the most critical pressures. Prager, Wei et al. (2016)[33] estimated that the economic loss from the US influenza pandemic would be USD90-220 billion and that 80% of the expected future lifetime earnings of those who would die would benefit. The World Bank calculation predicts large-scale outbreaks such as hantavirus, severe acute respiratory syndrome, H5N1 Flu, H1N1 Flu, respiratory syndrome of the Middle East, and Ebola virus, which cost more than two billion

dollars to the world. Indirect expenses are very high as well. They are all that lead to a decrease in GDP. In New Zealand, the Treasury looked at a pandemic with a % attack rate and a 2% case-fatality rate, concluding that GDP will be cut by 5-10% in the following year of the event. Some areas of the economy could be more highly affected than others. Many critical pandemics have been reported in modern history, and the resulting pandemic crisis has created substantial adverse effects on global health as well as in the economy. The pandemic has a significant negative effect. Millions of people have suffered from pandemics that have resulted in massive, severe disease and thousands of deaths in vast communities. Not just the world's inhabitants, but its economy are badly damaged. The effect of economic losses will add to economic fluctuations, due to direct costs, long-term burdens and indirect costs. An effective emergency response through machine learning methods will reduce the economic crisis. .

## 2.2 Economic impact

Previously Keogh-Brown et al. [14] described that due to SARS virus pandemic outbreak losses of various sectors losses are 3.7% in GDP(US\$ billion), 4.75% in Growth(US\$ billion), 23.1% in export and trade(US\$ billion), and 0.86% in tourism(US\$ billion) sector which was in Hong-Kong in 2003. Again, in the same year Canada's several sectors losses are 3.2-6.4% in GDP(US\$ billion), 1% in Growth(US\$ billion), 5.2% in export and trade(US\$ billion), 0.03% in tourism(US\$ billion) and 6.25% in Airline(US\$ billion).

In 2020 the USA, due to Covid-19 outbreak different economic sectors affected. Those sectors affected by increasing covid-19 positive test cases daily, which causes a national economic recession. Earnings received by workers get decreased by increasing positive test cases comparing with January 2020. Falling wages of workers suggests recession which causes unemployment and downward pressure on wages. It is visible that (from figure-2.1) Manufacturing sector workers earnings decreased 41%, Retail Trade workers earnings decreased 36%, Transportation and Warehousing decreased 30% and Health Care and Social Assistance decreased 31%.

Again, Employment level is also affected by this pandemic situation. Employment level or employment rate is defined as the number of people engaged in productive economic activities. Here,(from figure-2.2) Manufacturing employee's level decreased 41%, Retail Trade employee's level decreased 38%, Transportation and Warehousing decreased 32% and Health Care, and Social Assistance decreased 31% comparing with January 2020.

So, these sector-wise recessions of employment rate and worker wages turn out the USA's national economic recession. Moreover, net Revenue for all businesses is also decreasing at the same time. From the above figure-2.3, the maximum employment rate was 1% at the end of January 2020. Still, later with the increase of positive test cases, it gradually decreases and within the first week of May 2020 rate decreased 39% comparing with January 2020. Also, workers' wages decreased 38% within the first week of May 2020 again. However, Revenue for all small businesses starts increasing from the last week of January 2020, and the maximum value was 14%.

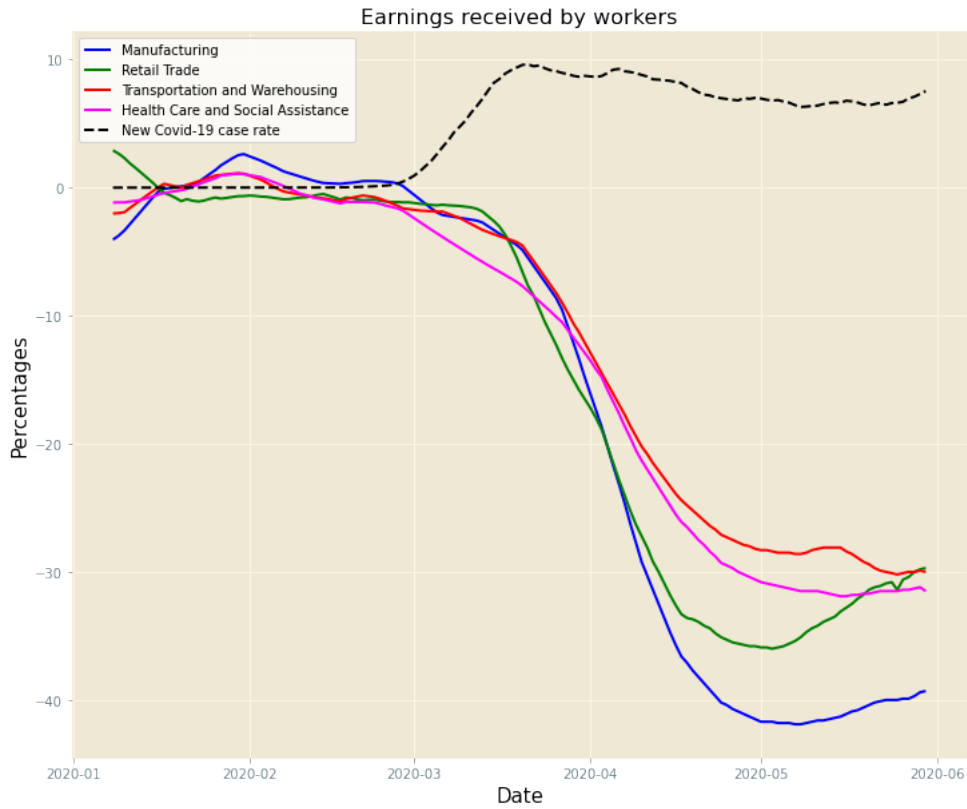


Figure 2.1: Wages received by workers and new confirmed test cases

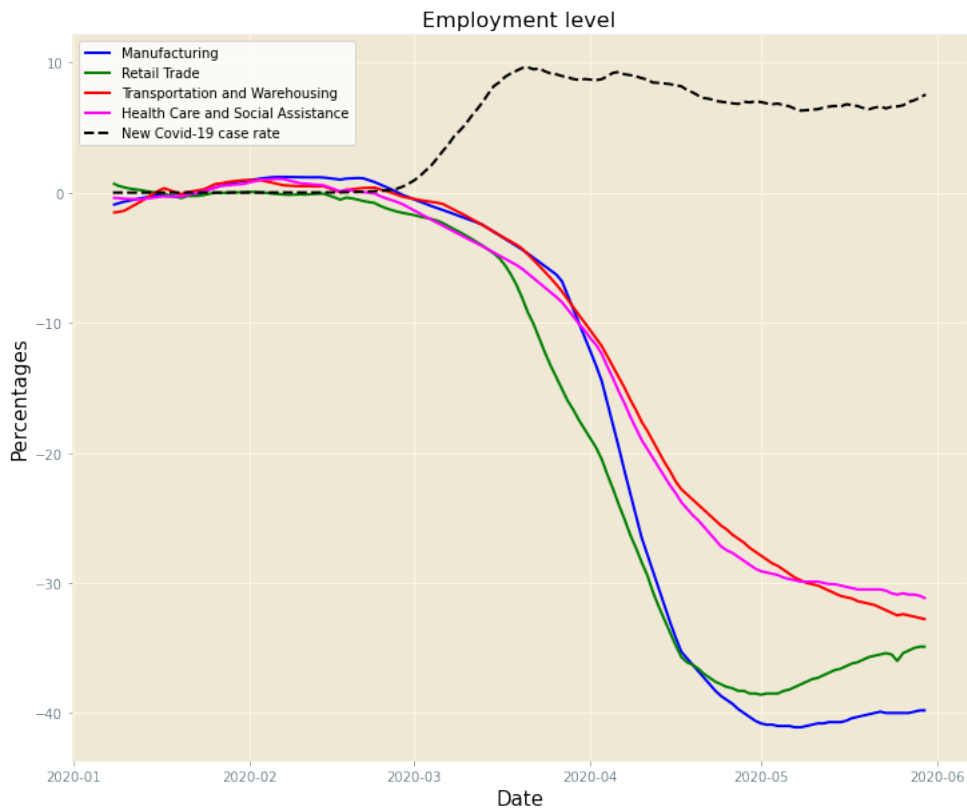


Figure 2.2: Employment level and new confirmed test cases

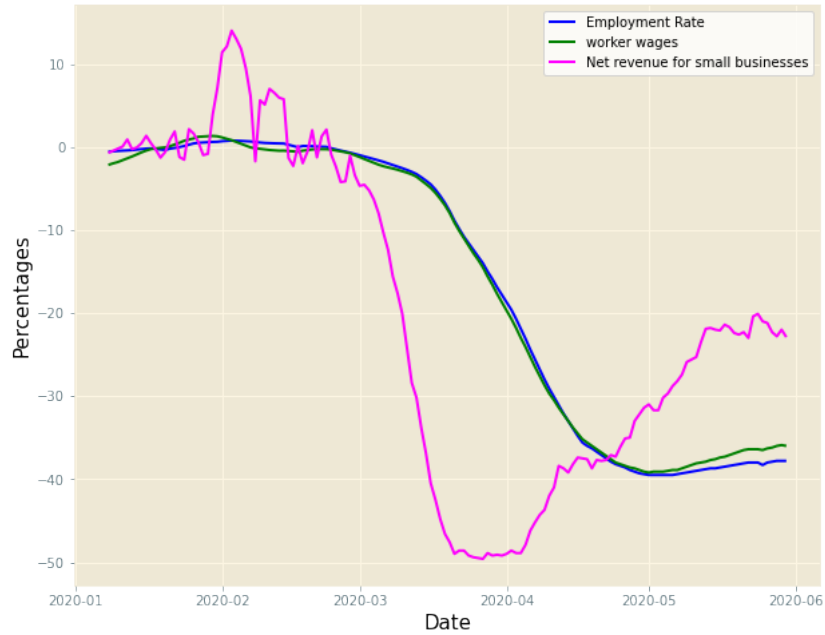


Figure 2.3: National employment rate, workers' wages and net revenue

Suddenly, it starts decreasing and at the end of March 2020, Revenue decreased 49% comparing with January 2020. After that, Revenue starts to rise again while the pandemic effect was increasing. After May 2020, both the employment rate and workers' wages start increasing.

## 2.3 Algorithm Description

It is difficult to predict the economy of a country where uncertainties involved such as pandemic outbreak, natural disaster etc within a country. So, for predicting future economy at present we have selected Neural Network and Random Forest machine learning algorithm. We will discuss about them in detail in this section.

### 2.3.1 Neural Network

Neural network is a machine learning algorithm for interpreting patterns and modeling complex data. Although it has many variant and nowadays it became more popular. Among all those variants to keep the model simple with low training time we have selected *Multi-layer Perceptron (MLP)* and *Long short-term memory (LSTM)*.

#### Multi-layer Perceptron

In Artificial Neural Network, Multi-layer Perceptron (MLP) is a fully connected feed-forward neural network that mimics the human brain to build up a machine learning model. Multi-layer Perception (MLP) is a class of Feed-forward Artificial Neural Network (ANN). MLP has three kinds of a layer, such as input layer ( $i$ ), which has an equal number of neurons, same as data-set features, hidden layer ( $j_0$  to  $j_n$ ) and output layer ( $k$ ) consist of a single neuron. Signal flows from left to right



layer by layer for computing output of each neuron, which is named as Forward Propagation of Function Signals [17]. Again, to minimize errors of the network originated error signal in the output neuron propagates backward layer by layer referred to as Backward Propagation of Error Signal.

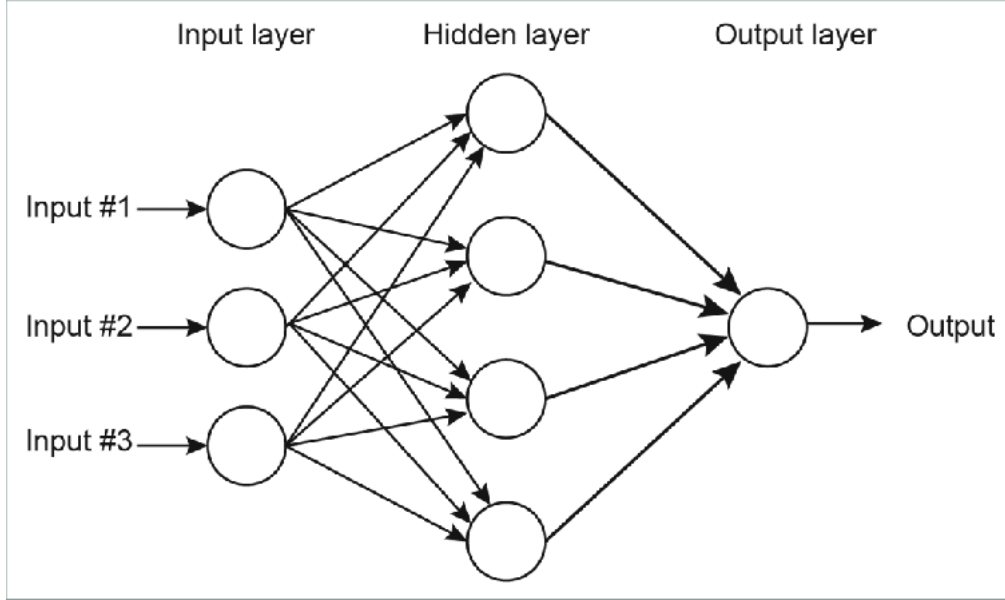


Figure 2.4: Multi-layer perception neural network

To construct a simple neural network model that will give higher accuracy with lower training time MLP is an excellent choice compared to other complex models [39]. For developing an ideal MLP architecture, choosing hyper-parameters are essential because they directly rule over the behaviour of the model also have a more significant impact on performance. These hyper-parameters can be set manually. But to find best-fitted hyper-parameters few algorithms can be used such as Grid Search Algorithm. By using *Grid Search Algorithm* hyper-parameters such as weights, learning rate, the number of hidden layers, activation function etc. can be found those will give a higher accurate result.

Each neuron in the network will produce output following some steps of calculation. Firstly, each input will multiply by the assigned weight of that link, merely a summation of inputs and weights. Then, the summation value will pass through an activation function. An activation function is added into a neural network to learn intricate patterns in the data. Hence, the total equation of output from the  $i$ -th neuron is,

$$O(j) = f\left(\sum_i^n w_{ji} y_i + b_i\right) \quad (2.1)$$

Here,  $f$  is the activation function,  $W_{ji}$  is the weights of connected links,  $y_i$  are inputs, and  $b_i$  is the bias. This process of computing output is the same for all hidden layer and output layer neurons.

Using the above equation output from  $k$ -th neuron may not be accurate, so to

minimize errors, MLP uses backward propagation of error signal technique, also known as Back Propagation Algorithm. It adjusts the weights in such a manner that produces results exactly as the desired output. The formula for update weights is *Delta rule*,

$$w_{new} = w_{old} + \eta \delta_k y_j \quad (2.2)$$

Where  $\eta$  is the learning rate, which decides how quickly the network will converge. Again,  $\delta_k$  is Local Gradient, and  $y_j$  is the input of  $k$ -th neuron or output from  $j$ -th neuron. After that, equations for calculating  $\delta$  is different for hidden and output layer. If  $k$ -th is a neuron of the output layer,

$$\delta_k = \phi' \left( \sum_k w_{kj} y_j \right) (d_k - y_k) \quad (2.3)$$

Here,  $\delta_k$  is the desired output from the  $k$ -th neuron, and  $\phi'$  is the first derivative of the activation function. Again  $\delta_j$  for  $j$ -th neuron of a hidden layer,

$$\delta_j = \phi' \left( \sum_j w_{ji} y_i \right) \left( \sum_{c \in k} \delta_c w_{cj} \right) \quad (2.4)$$

Where  $c$  is the set of next hidden/output layer neurons and  $y_i$  is a set of input layer neurons or hidden layer neurons, but it is the previous layer of  $j$ -th neuron's layer. In this way, MLP model training will continue until it meets stopping criteria. When the Average square error (loss function) change is sufficiently small for per epoch, then the training process will stop. Another measure is, after each period, the MLP model will be tested for Generalization, and if this generalization performance is suitable only then, training will stop.

Multi-layer Perceptron (MLP) is a popular choice in the research area for its ability to solve a complex problem with lower training time and most accurate results. Previously, by using MLP many complex problems such as Financial Prediction, Pattern Recognition/Computer Vision, Speech Recognition, ECG Noise Filtering and so on solved. Problematic issues with the simpler network are a right combination by which we will get an accurate result also the machine will also train itself efficiently.

### LSTM regression

Long short-term memory (LSTM) [37] is a recurrent neural network that is used in the field of deep learning area. LSTM has overcome some major limitations of recurrent neural networks. To be specific, LSTM has solved the vanishing gradient problem, which is encountered when training artificial neural networks with gradient-based learning methods and back-propagation. A common LSTM cell is formed of three gates - *input gate*, *output gate*, and *forget gate* [35].

In LSTM structure the first layer is called the forget gate layer which selects the information to forget. It can produce any value between 0 (completely forget) to 1 (use fully)[43].

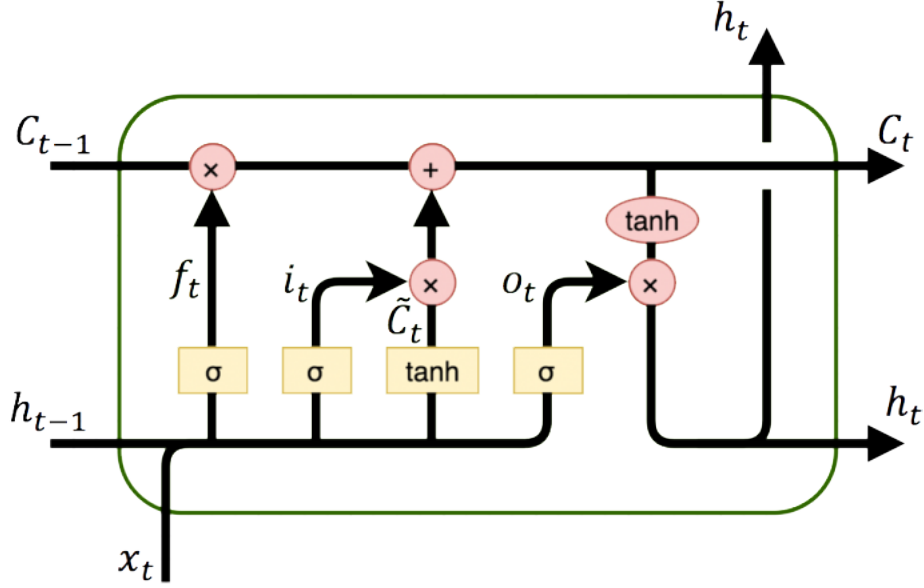


Figure 2.5: Long Short Term Memory

$$f_t = \sigma (W_f \cdot (h_{t-1}, x_t) + b_f) \quad (2.5)$$

Here,

$\sigma$  = sigmoid function,

$W_f$  = weight vector of inputs,

$h_{t-1}$  = forecast vector from previous periods,

$x_t$  = new input vector and

$b_f$  = bias of function

Next step of the algorithm is to determine the new input which needs to be added.

$$i_t = \sigma (W_i \cdot (h_{t-1}, x_t) + b_i) \quad (2.6)$$

After that, the algorithm determines the new candidate value of neural cells.

$$\tilde{C} = \tanh (W_c \cdot (h_{t-1}, x_t) + b_c) \quad (2.7)$$

After computing above equations, it states the new cell states by computing  $c_t$ .

$$c_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (2.8)$$

$$C_t = \sigma (W_f \cdot (h_{t-1}, x_t) + b_f) \times C_{t-1} + \sigma (W_i \cdot (h_{t-1}, x_t) + b_i) \times \tanh (W_c \cdot (h_{t-1}, x_t) + b_c) \quad (2.9)$$

Lastly, the output function will predict the value. This layer is called prediction layer and the prediction will then be  $h_t$ .

$$o_t = \sigma (W_o \cdot (h_{t-1}, x_t) + b_o) \quad (2.10)$$

$$h_t = \tanh o_t \tanh (C_t) \quad (2.11)$$

### 2.3.2 Random Forest

Random forest learning model used to train a machine for classification and regression both. To complete our task we will use Random Forest classifier and regressor.

#### Random Forest Classifier

Random forest classifier [29] uses an ensemble classification learning system that incorporates several decision trees to predict individual trees during the training and performance average. This classifier creates forests with a random number of trees. Typical algorithms for the decision tree are rule-based and based solely on data-set prediction rules. Instead of using the Gini index [22] or data benefit for root node calculation, random forest classifiers find the root node and randomly divide the functions.  $X$  reflects the classifier's input. Random decision trees, respectively  $tree_1, tree_2$ , and  $tree_B$ , with corresponding outputs of  $k_1, k_2$ , and  $tree_B$  are generated. The plurality vote is taken, and class  $k$  is chosen out of  $k_1, k_2, \dots, k_B$ . With the bulk of the ballots, the classifier's performance is class  $K$ .

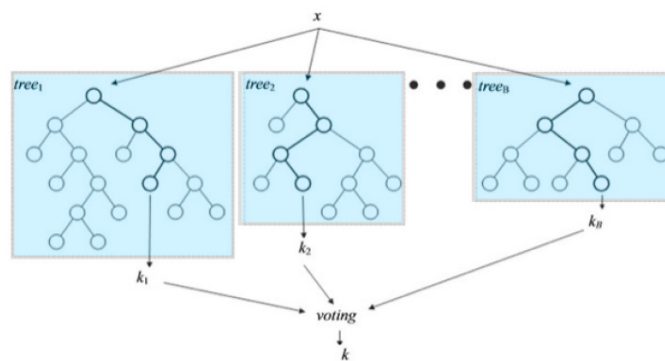


Figure 2.6: Random forest classifier

#### Random Forest Regressor

One of the most common and most crucial supervised machine learning algorithms is a random forest regression (RFR). It generates multiple decision trees. Generally, the more trees in the forest, the more robust the forecast and give higher accuracy. With the help of gain or Gini index, multiple decision trees are created in the forest. When a large proportion of data is missing, the RFR can tackle the missing values and retain accuracy, and it can handle a range of data sets of greater dimensionality. RFR is an ensemble machine learning technique, which takes predictions from the same algorithms multiple times to make more detailed predictions. Predictions from a single model cannot provide a very detailed outcome. Therefore, we use RFR to obtain higher accuracy. During the training time, RFR creates several decision trees and integrates the outcome of prediction from each decision trees to decide the final output. We first choose a random  $N$  number of data points from the training set to create a random forest. Using the data-set, we select the subsets from it. Then, we build the decision trees connected to those  $N$  data points; then we are constructing

decision trees for those data point subsets here. Then, we pick how many trees we want to create, which is the  $K$  tree, and repeat the steps above, thereby generating lots of decision trees for regression. Lastly, we make each of our trees forecast the ideal data point's value and allocate the average for a new data point to the same data point over all the expected values. So, if the value of several trees is set to 600, then we will get individual projections from all 600 trees for the value of the desired data, and then we will take an average of the expected value of all 600 trees from our desired data.

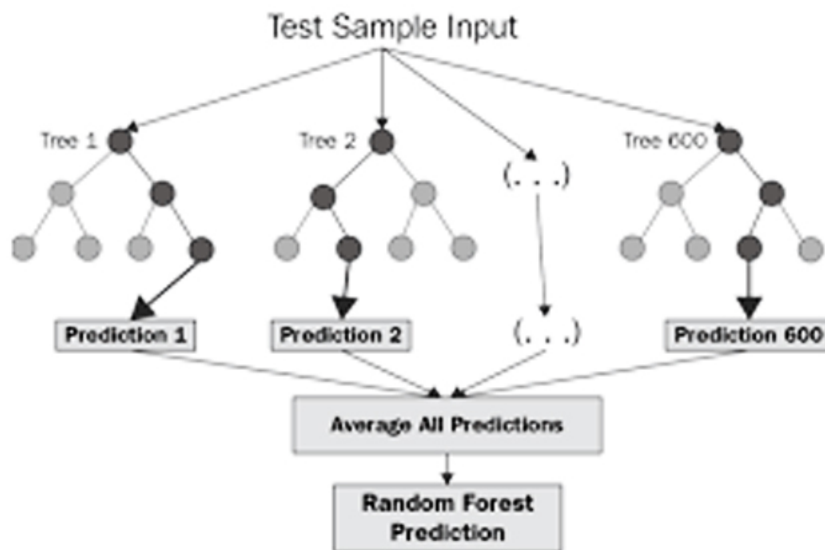


Figure 2.7: Random Forest Regressor

# Chapter 3

## Literature Review

There is enough research conducted regarding economic rise and fall due to the pandemic outbreak. These studies are related to ours and shows how a pandemic can influence a country's economy, culture, peoples and others.

Smith et al. [21] show the UK economy's impact based on the Computable General Equilibrium (CGE) model. This research estimates that only pandemic influenza can minimize GDP by 0.3%, 0.4%, and 0.6% respectively for mild, moderate, and severe. Additionally losses of other sectors are 1.5% in domestic output, 2% to household consumption, 3% to exports, 2.5% to imports, 2% to government consumption calculates from data-set 2003 supply chain in the UK and other researches using benchmark equilibrium. This paper presented that a large portion of the economy was damaged due to the pandemic.

Keogh-Brown et al. [14] suggest macro-economic outcome after Severe Acute Respiratory Syndrome (SARS, 2003) outbreak through affected countries and their economic sectors evaluated by economic indicators. Researchers calculated losses of various sectors such as GDP, Growth, Exports and trade, and Tourism, food and travel; for instance, Hong-Kong losses are 3.7% in GDP, 4.75% in Growth, 23.1% in export and trade, and 0.86% in tourism using national statistics databases. This research gives a clear picture of affected sectors of individual countries and why stakeholders could develop solutions to be aware of the next pandemic outbreak.

Fernandez-Delgado et al. [26], compared 179 classification algorithms in a journal paper, such as Bayesian, neural network, logic, polynomial regression, etc. to predict the economic crisis. Their output and implementation were tested on 121 data sets in the learning database of the University of California, Irvine. The results showed that multiple models showed different prediction results. Hence, deciding which one to be selected was a challenge. Lastly the random forest family algorithm showed the best results in the early warning of the economic crisis.

In recent years, bankruptcy forecasts have been made by machine-learning [12] using standard statistic approaches. To obtain improved failure – prediction solution of problems, analysis has been done using the mathematical and machine learning techniques[13]. In particular, the data sets, financial partnerships, country of origin and the timeline of the analysis where used and the results and implementation were

compared with several different prediction accuracy techniques [19] [13].

This paper, [19] aims to predict or forecast the financial status of institutions or individuals by using machine learning algorithms to take practical steps to reduce the sufferings and take the best decision to overcome the financial crisis. Here, tools such as neural networks, decision trees, etc. have been extensively studied to predict financial crises. Use MLP, Random Forest (RF) to apply machine learning methods, such as pattern classification techniques, single classification techniques and soft classification techniques. In the observation, we have seen some problems that are not widely discussed in the literature. When different data sets are used for different training and evaluation sets and cross-validation, more accurate results can be obtained.

It is difficult to make crucial decisions due to the vulnerability of the emerging coronavirus outbreak. Deep learning and fuzzy prediction approach are proposed by Fong et al. [41] on different outcomes of Coronavirus and its effect. The current events and their possible actions were described using the Composite Monte-Carlo simulation system. Fong et al.[41] address the daunting challenges of predicting the fate of an outbreak correctly effectively using the availability of dataset, the layout for picking the best predictive model and finely tuning each models parameter.

The statistical data were benchmarked in a paper by Bluwstein et al. (2020) [38], used in many machine learning models, such as decision trees, random forests, large-scale random trees, SVM and artificial neural networks. The author [38] and found that, except for human decision-making agencies, all machine learning models performed better than logistic regression. The best performing machine learning model (a tree that is too random) can correctly predict the global financial crisis of 2007-2008 and provide outstanding signals in countries/regions with different economic realities and achievements [38].

In a paper [32], Nyman gives an actual forecasting scenario where they use a small number of financial variables for the prediction of the economic recession. The author also improves the prediction of the actual GDP growth by using a certain machine learning approach. The author uses two Model estimation techniques, including ordinary least squares regression and random forest machine learning. Random forests [7],[16] noisy, non-linear, high-dimensional prediction can be tackled using machine-learning models. The author obtains qualitatively similar results for the UK and USA through the random forest algorithm's predictive power is more efficient for the USA. Finally, the author says the machine learning approach is very efficient for forecasting horizons and providing better information.

A nation and a financial institution need to model an accurate and earlier forecasting model to estimate the possible probability of a company's financial loss status in advance. The prediction of the financial crisis produces a binary classification model that has been rationally resolved by Martin et al. 2012[1]. The classifier model's findings can be divided into two types: the first one representing a company's failure status and the second one representing its non-failure status. The economic crisis prediction has faced significant problems with rapid growth in the quantity

of financial data. The difficulty of the classification algorithm continues to increase with the vast number of the dataset. Thus, Clustering strategies are used before the classification task (Evans et al.) [18] to enhance the classification efficiency in both a broad and complex dataset. The clustering strategies aim at splitting the observation into single clusters, and the findings in similar collections are closely related to those assigned to distinct sets (Guojun et al. 2007) [36].

To date, many models have been developed with different domain knowledge for the prediction of economic crisis. Logistic regression (Ohlson 1980) [1] and factor analysis are some of the conventional statistical approaches (West 1985) [2]. Some of Artificial Intelligence methods such as artificial neural network ANN (Atiya 2001) [6], Support vector machine (SVM) (Min and Lee 2005) [10] (Shin et al. 2005) [11], Bayesian networks (Sarkar and Sriram 2001) [8] (Sun and Shenoy 2007) [8] and many integrated machine learning techniques (Fedorova) et al. 2013) [24] (Abellán and Mantas, 2014) [25], several hybrid methods are widely used to predict economic crises.

Because it can handle the nonlinear interaction between different features in the actual data set, ANN (Tsai and Wu 2008) [15] (Zhang et al. 1999) [5] is considered to be an essential tool for financial crisis prediction. However, it should also be noted that ANN learning methods (such as backpropagation) rely on gradient descent strategies that lead to the best local results. In addition, a sufficient number of network parameters should be adjusted frequently.

Car, Z. et al. [39] built a model that can detect the spread of covid-19 infection to predict its impact. They used a 51-day data set to train a multi-layer perceptron (MLP) neural network, and chose to use a grid search algorithm for hyperparameter optimization. After performing k-fold cross-validation, they found that the accuracy of positive confirmed cases was 94%, the accuracy of recovered patients was 78%, and the accuracy of deceased patients was 98%. This is an excellent model for the detection of covid-19 infection.



# Chapter 4

## Proposed Material and Methods

In this chapter, we will discuss about implementing this model to find out the recession in a country due to the pandemic outbreak. Firstly, we will start with data description that will describe from where we collected data-sets, structure etc. Then, how we process data-set in order to fit to our model. Finally implementation of algorithms for detecting recession in a country due to this covid-19 pandemic. The process we followed is given in figure-4.1,

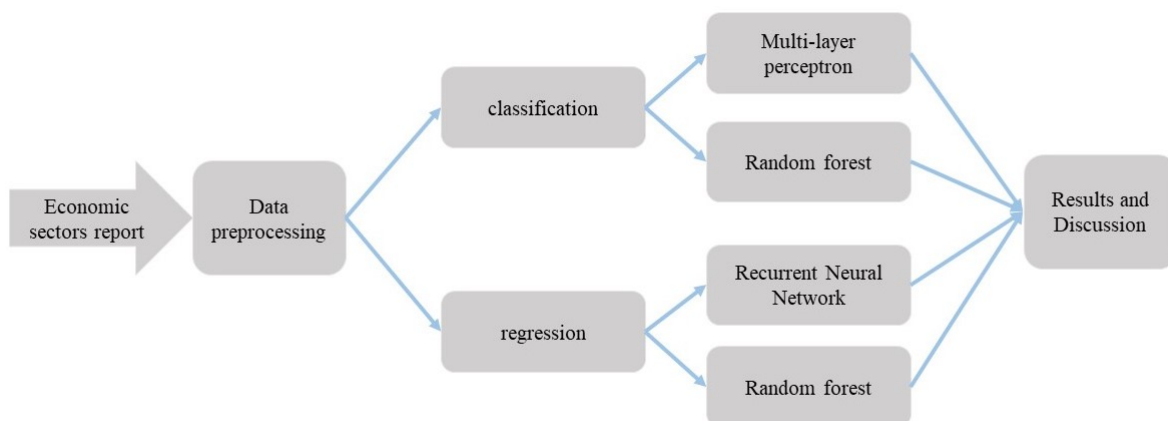


Figure 4.1: architecture of proposed model

### 4.1 Data Description

The datasets used in this research are obtained from a publicly available archive operated by the Opportunity Insights Economic Tracker, a non-partisan, not-for-profit organization located at Harvard University [40]. The archive contains data from leading private companies – from credit card providers to payroll firms - to offer a real-time image of indicators such as employment rates, consumer spending, and job postings across counties, industries, and income groups. Our dataset is split into three groups - employment rate, worker’s earnings, and national revenue

of various economic sectors. During this research, the dataset contained the data of 239 days for COVID-19 positive cases, small business revenue, and employment rates among low-income workers. Moreover, the dataset, as published, is organized as time-series data. Besides, our dataset contains data points of 239 days, which is split into five folds.

## 4.2 Data Preprocessing

The data-set has three different kinds of information: employment rate, worker’s earnings, and national revenue of various economic sectors. Firstly, the data-set was time-series data, so we convert it in a way that fits the model. As we are classifying, is there any recession or not? If any value from the employment rate of all sectors is negative, then there exists a recession; otherwise, not. To full-filling this purpose, we are adding one extra feature depending on the overall employment rate, which will be our attribute column for classifying recession in employment rate and the same for others.

In the real world data-set, missing data points are not unexpected. In this paper, used data-sets also have missing data points. To reduce biases, we used statistical imputation also to fill up the gaps of missing data points. Although there are varieties of techniques for imputation, we have used mean value for each feature. Replace missing data points with mean is simple; besides does not introduce many biases in the data.

$$Mean = \frac{x1 + x2 + x3 + \dots + xn}{n} \quad (4.1)$$

Each data set is divided into five random folds, which are used to train and validate the model. Here, in order to train the model (k-1), folding is used, and the rest will be used to verify the model, where k = 5. This process continued unless each fold was used as a test set for the model, which gave us the advantage of having a generalized model. In addition, because our data points are limited; therefore, this K-fold cross-validation technique has helped us use each data point as a training set and a test set [20]. To determine the cross-validation result, we calculate the average of the R2 scores.

$$R^2 = \frac{1}{5} \sum_{n=1}^5 R_k^2 \quad (4.2)$$

## 4.3 Implementing Classifier in Economic Recession Prediction Model

In this section , we are going discuss about how we detect recession due to pandemic outbreak. For detecting this issue first we will use MLP then Random forest. Both the algorithm will help us to find out recession a country might face due to this covid-19 affect.

### 4.3.1 Multi-layer Perceptron Classifier

Multi-layer perceptron (MLP) is a fully connected feed-forward network often applied to supervised learning problems that mimic the human brain to build up a machine learning model. MLP has three kinds of a layer, input layer ( $i$ ) which has an equal number of neurons same as data-set features, hidden layer ( $j_0$  to  $j_n$ ) and output layer ( $k$ ) consist of a single neuron. There is a Forward Propagation of Function Signals mechanism for calculating each neuron's outputs except input layer neurons and a Backward Propagation of Error Signal mechanism for minimizing the errors generated in the network [17].

To generate a network that can produce a higher accurate result, choosing hyper-parameters is a crucial factor. For the different value of hyper-parameters network will show different performance. For instance, if hyper-parameter like the number of iterations is excessive than network might face an over-fitting problem. Additionally, depending on the learning rate network will converge slowly or quickly. We applied the Grid Search Algorithm [30] to find best-fitted hyper-parameters such as weights, learning rate, number of hidden layers, etc., for network convergence. Besides, we set a few hyper-parameters constant such as stochastic gradient descent, which is a solver for weight optimization, a maximum number of iterations is 100, activation function (*relu*), an initial learning rate 0.1. However, Summation of inputs and connected links weights will pass through *relu* activation function. *Rectified Linear Unit (relu)* function shows better convergence performance and computationally efficient since it maps only the value  $\max(0, z)$  where  $z$  is  $z > 0$ . This function is also differentiable. Hence, the total equation of output from the  $i$ -th neuron is,

$$O(j) = \text{relu}\left(\sum_i^n w_{ji}y_i + b_i\right) \quad (4.3)$$

Here, *relu* is the activation function,  $W_{ji}$  is the weights of connected links,  $y_i$  are inputs, and  $b_i$  is the bias in our case it is 1. This process of computing output is the same for all hidden layer and output layer neurons.

An activation function decides whatever it will fire the output of that particular neuron or not. Now,  $\eta$  is the learning rate, which decides how quickly the network will converge and we put  $\eta = 0.1$  so that machine can learn slowly and perform with higher accuracy in the long run. Hence updated *Delta rule*,

$$w_{new} = w_{old} + (0.1 * \delta_k * y_j) \quad (4.4)$$

Where  $\delta_k$  is Local Gradient, and  $y_j$  is the input of  $k$ -th neuron or output from  $j$ -th neuron,  $w_{old}$  is the previous iteration's weight.

After putting the values mentioned above, we have got a satisfactory result from the network. However, to check this network's robustness, we have applied the k-fold (k=5) cross-validation technique. After applying this technique, some of the results dropped, and some showed more accuracy. Nevertheless, this difference is shallow, so this network fulfilled our exception for detecting the issue that we wanted to solve. Finally, this is a simplex network having the ability to solve a complex problem like detecting financial crisis with lower training time.

### 4.3.2 Random Forest Classifier

The random forest classifier is a collection of projected trees, where each tree is subordinate to independently evaluated random vectors, with comparable transport within the random forest with one another tree. It, too accomplishes the proper speed required and productive parameterization within the process. The random forest classifier bootstraps random tests where the expectation with the most elevated vote from all trees is chosen. The distinction of each tree is ensured due to the taking after qualities. To begin with, each tree training within the test employs random subsets from the beginning training tests. Besides, the ideal part is chosen from the unpruned tree nodes' arbitrarily chosen features. Thirdly, each tree develops without limits and ought not to be pruned at all. In our model, we have implemented a random forest classifier for classification that uses an ensemble learning approach to prediction [22], which uses several decision trees during the training process and average individual tree prediction outputs. Random Forest efficiently runs on massive datasets, can handle thousands of input variables without variable deletion, produces significant variable for forecasts, [23] creates an internal unbiased measure of generalization error as forest growth increases, has an adequate method to estimate lost data [7], and maintains accuracy where a large proportion of data is lacking. The chosen RFC's key objective is the power of the individual decision tree and the relation between base trees [7]. Random forest classifier comprises various individual classification trees, where each tree may be a classifier given diverse weighted classification. The output of the classification determined the overall classification. It builds each tree by part number of features for each part without pruning [27].

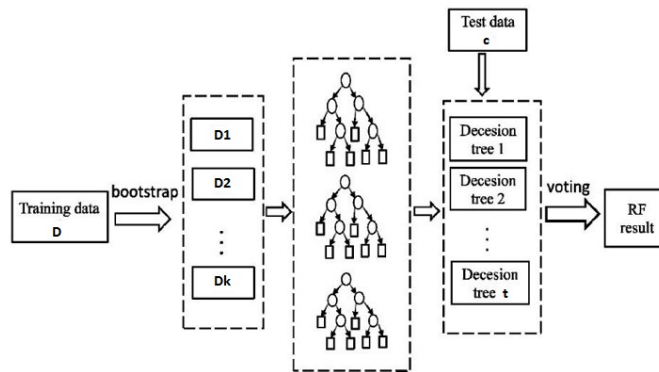


Figure 4.2: Random forest model

The classifier of Random Forest functions as consequently [34]: We Choose  $C$  subsets from which to construct the training set  $T$  and initial training data  $D$  using bootstrap sampling

$$T = D1, D2, \dots, DC \quad (4.5)$$

The algorithm automatically generates the decision tree for  $C$  models with a random vector  $c$  for each subset:

$$M = M_1, M_2, \dots, M_c \quad (4.6)$$

The random vectors  $\theta_c, c = 1, 2, \dots, C$  are both distinct and distributed identically. Each decision tree evolves freely without pruning so that all the trees are combined to get a forest.

$$r(T, c), c = 1, 2, \dots, C \quad (4.7)$$

For determining the classification of a new input variable  $t$ , the final vote of each established decision tree model is accurate. The result for classification is:

$$R(t) = \sum_{i=1}^c I(r_i(t) = Z) \quad (4.8)$$

Where  $R(t)$  denotes the product of the classification is;  $r_i(t)$  denotes the outcome of the decision tree classification;  $Z$  is the target group, and  $I(r_i(t)=Z)$  stands for the characteristic function.

n-estimators is several trees in the forest. Since having a few distinctive forms of trees with different depths and sizes boosts the generalization the n-estimators trees that the algorithm needs to construct. If the estimator is  $n$ , it gives  $n$  other decision trees. By using n-estimators, we build several trees within a Random Forest before aggregating the predictions. Though the higher the number is better, we want to make sure the computational expense while taking the trees as it delays our code to run.

Using the criterion Gini-impurity, we determine the feature of a tree has to split the parameters. It measures the quality of each split. We use a max-depth to determine how much further the tree has to be expanded to each node until we get to the leaf node. We set the max-depth because if we let it run down to the purist option, it risk over-fitting our model. By using the max-features, we consider the maximum number of features for splitting a node. We are going to observe the impact of the max-features hyper-parameter. This takes into account the number of top features given to each tree in a random forest. We realize that the random forest selects a few random tests to explore the main break from the functions. We can see that the execution of the first increments demonstrate that the number of max features is increasing. Nonetheless, the training score keeps expanding after a certain point. However, the test-score saturates and begins diminishing the conclusion, which implies that the show starts to over-fit. Ideally, the general execution of the demonstration near six of the highest highlights is the most notable. In general, the ideal number of total features tend to be similar to this value. To decide whether the algorithm will avoid further splitting, we use min samples split to define the minimum number of records present in each node. If the split number is less than  $n$ , there will be no further split. We use max-leaf nodes to expand the tree in a best-first manner resulting in a relative reduction in impurity. We use max-leaf nodes to grow the tree in a best-first way resulting in a relative decrease in impurity. The random state makes it simple for others to imitate our results if given the same training data and parameters. For sampling data points, we use the bootstrap process. The algorithm randomly selects many rows with replacement to construct

the trees using bootstraps once we provide the Random Forest Classifier model’s training data. If the bootstrap option is set to False, there will be no random sampling, and the entire data-set will be used to build the trees. We use oob-score as it is much quicker because it gathers all of the trees’ observations and finds the highest score on each observation base’s trees that did not use that observation to train. Oob-score, however, is a cross-validation technique that is somewhat similar to a leave-one-out validation technique in which a model’s generalized approximate output is trained on n-1 data samples. We set n-jobs to -1 will often lead to faster processing. If we use -1, there is no limit on how much computing resources the code can use n-jobs helps the program know how many processors it can use. The default value of 1 means that only one processor can be used. We set the logging output to be verbose, which gives us continuous feedback on what the model is doing as it is processed. This parameter defines the verbosity of the construction method of the tree. We use false Warm Start for recursive feature collection, and false Warm Start suggests that other features will gain in value as we drop such features, and it will be repeated used. It is often used in regression models with backwards exclusion and is not often used in classification models.

## 4.4 Implementing Regression in Economic Recession Prediction Model

In this section , we are going to cover how much economic recession a country might face due pandemic outbreak. First we will discuss about how RNN implementation helped in this regard secondly Random forest.

### 4.4.1 Recurrent Neural Network

In our model, we used the long-term short-term memory (LSTM) method [4] to estimate the economic impact of the COVID-19 outbreak on different regions of the United States. When dealing with time series, we used LSTM modelling, which is an in-depth learning method that is useful when trying to model time series. In the LSTM unit, there are four functions in our model, which are 3 *Sigmoids* (f, i and o below) and *Tanh* (c function below). It is mentioned that the coefficient of deviation may be a general feature of all functions in the learning model, which can be set or measured in advance during the training process. In order to help it adjust to the various situations of each case, the bias can be used to model calibration.

$$f_t = \sigma (W_f \cdot (h_{t-1}, x_t) + b_f) \quad (4.9)$$

$$i_t = \sigma (W_i \cdot (h_{t-1}, x_t) + b_i) \quad (4.10)$$

$$o_t = \sigma (W_o \cdot (h_{t-1}, x_t) + b_o) \quad (4.11)$$

$$\tilde{C}_t = \tanh (W_c \cdot (h_{t-1}, x_t) + b_c) \quad (4.12)$$

$$c_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (4.13)$$

$$h_t = \tanh o_t \tanh (C_t) \quad (4.14)$$

We add the LSTM layer and later integrate a few Dropout layers to prevent over fitting. We integrate the LSTM layer with 50 units which is the dimensionality

of the output space. We define the dropout layers 0.2. This indicates that it will reduce the number of layers by 20%. After that, we merge the dense layer, which determines the output of 1 unit. Next, we used a very common optimizer, the Adam optimizer. Then, we fit the LSTM model to run at 100 epochs and 32 batch sizes.

We will conduct k-fold on the data in order to improve the complexity of our results. K -fold is a cross-validation technique that requires the creation of various models on subsets of the data set. This approach would be very beneficial in achieving the desired predictive precision standard.

#### 4.4.2 Random Forest Regressor

The random forest regression (RFR) incorporates a wide collection of decision trees trained together to produce a more precise and reliable final forecast [7]. Based on some parameters, a regression tree is ordered from the leaf to the root node. RFR consists of a supervised learning algorithm for predicting output target feature average by bootstrap aggregation or bagging [3] of independently built decision trees. Bootstrap aggregation or bagging is used for lowering variance error source of independently built decision trees. We first randomly select k points from our training data set of employment rate, worker income and national income of different sectors. Then we created a bootstrap sample of the random data with replacement and created the root node, eventually forming a decision tree. After that, We have taken N number of trees and created multiple decision trees from it. To estimate an economic recession, each of our N numbers of trees forecasts a data point value and assigns the average new data point over all the expected values. Our forest-random regression trees consist of a greedy top-down method to define ideal recursive divisions in binary nodes at each node by minimizing variance locally in every terminal node by inserting the squared error function at every point.

$$\text{Min(SSE)} = \left( \sum_{i=1}^n (y_i - y_s)^2 \right) \quad (4.15)$$

Here SSE=Sum Squared error,  $y_t$  =output target feature data, and  $y_s$  =terminal node output target feature mean And  $y_s$  = is calculated by,

$$y_s = \frac{1}{m} \sum_{i=1}^m (y_t) \quad (4.16)$$

Here  $y_s$  is the mean terminal node, m= number of observations in the terminal node, and  $y_t$  =output target feature data.

In a Random forest regression algorithm, tree bagging consists of predicting the output feature of an independently built decision tree by calculating the arithmetic mean,

$$y_p = \frac{1}{k} \sum_{i=1}^m (y_s) \quad (4.17)$$

Here  $y_p$  =mean output target feature prediction, k=number of independently built decision trees, and  $y_s$ =independently built decision trees output feature prediction.

# Chapter 5

## Result and Discussion

In this chapter, by following the described methodology, we will discuss the achieved results in detail. We will also analyze essential factors for achieving these results.

### 5.1 Result of Classifiers

As we choose to use two different classifiers for detecting recessing in a country which might introduce from a pandemic effect. Now we are going to discuss results found from multi-layer perceptron and random forest algorithm described in chapter 4. The formulas for calculating the above values for both MLP and Random forest classifier are,

$$\text{Accuracy} = \frac{\text{Correctly predicted class} \times 100\%}{\text{Total testing class}} \quad (5.1)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5.2)$$

where, TP=true positive and FN= false negative

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (5.3)$$

where, TN=true negative and FP=false positive

#### 5.1.1 MLP

Firstly, from the MLP classifier, we got an accuracy of 81% for the employment level, 89% for worker earnings, and 95% for national revenue. To achieve these results, we have used the initial learning rate (alpha) 0.1, *relu* activation function, three hidden layers with different units of neurons, and others.

Section	Accuracy	Sensitivity	Specificity	Accuracy(cv)
Employment level	81%	100%	0%	82%
Worker Earnings	89%	100%	0%	88%
National Revenue	95%	100%	67%	87%

Table 5.1: Results from MLP classifier (CV=cross validation)



After applying k-fold cross validation where k=5, suddenly accuracy level showed different results. Taking the mean value of 5 folds we get, 82% for employment level, 88% for worker earnings, and 87% for national revenue.

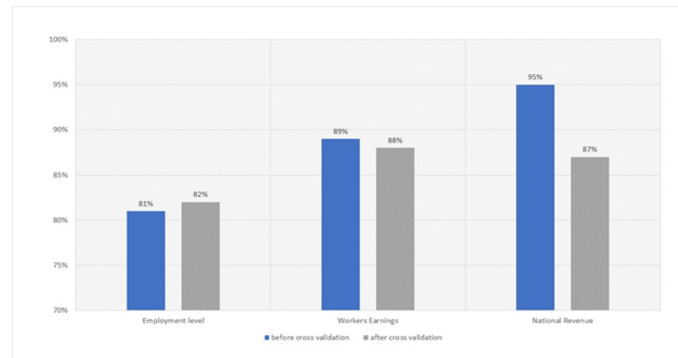


Figure 5.1: Results from MLP classifier before and after cross validation (CV)

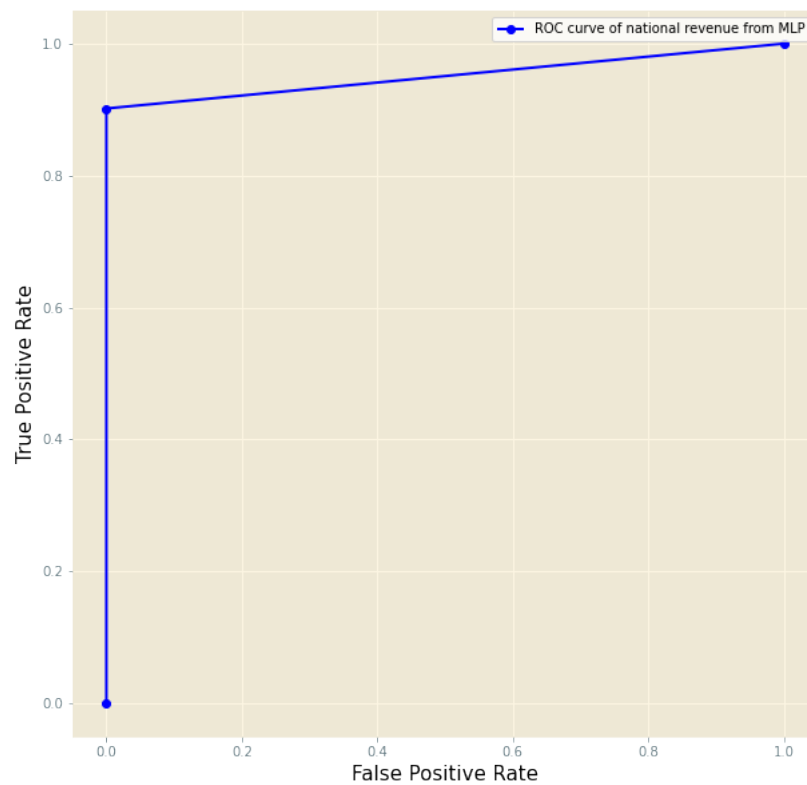


Figure 5.2: ROC of national revenue using MLP

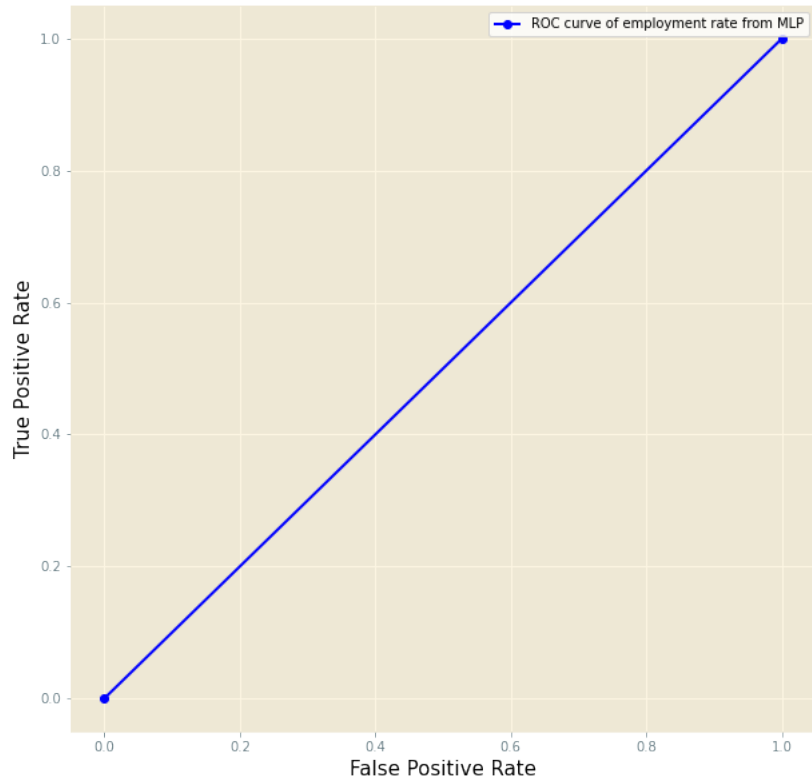


Figure 5.3: ROC of employment rate using MLP

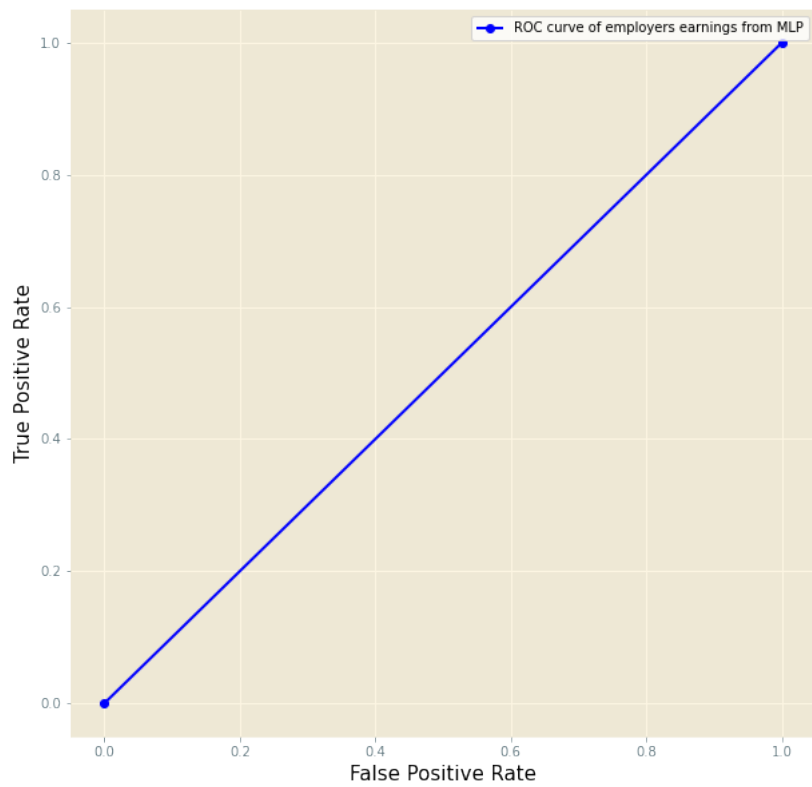


Figure 5.4: ROC of employer's earnings rate using MLP

### 5.1.2 RFC

By using the Random Forest classifier, we got an accuracy of 81% for employment level, 89% for worker earnings, and 85% for national revenue.

Section	Accuracy	Sensitivity	Specificity	Accuracy(cv)
Employment level	81.13%	100%	0%	98.94%
Worker Earnings	89.13%	100%	0%	99.04%
National Revenue	85.93%	85%	100%	98.43%

Table 5.2: Results from Random forest classifier

After applying k-fold cross validation where k=5, suddenly accuracy level showed different results. Taking the mean value of 5 folds we get, 98.94% for employment level, 99.04% for worker earnings, and 98.43% for national revenue.

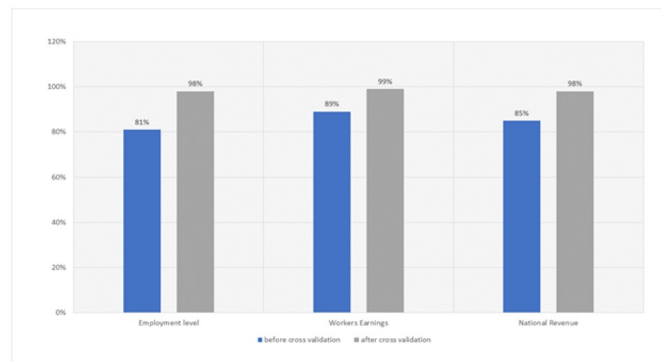


Figure 5.5: Results from Random Forest classifier before and after cross validation (CV)

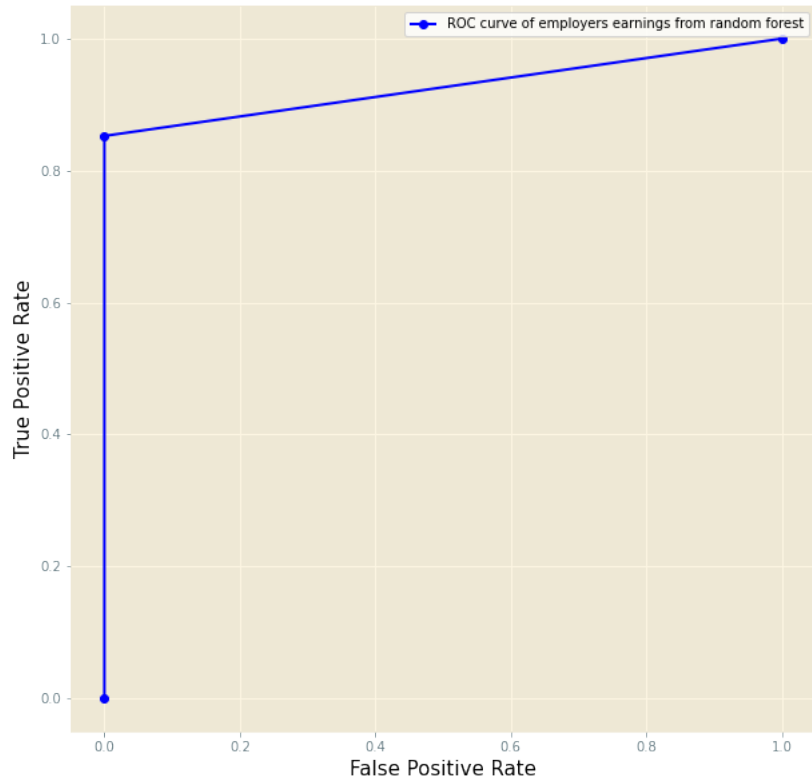


Figure 5.6: ROC of national revenue using random forest

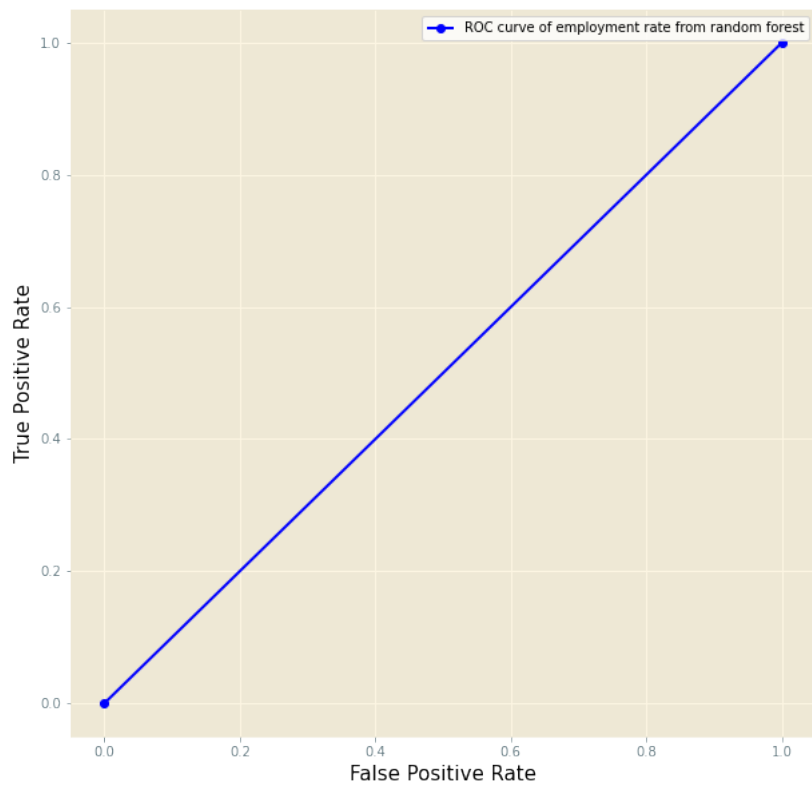


Figure 5.7: ROC of employment rate using random forest

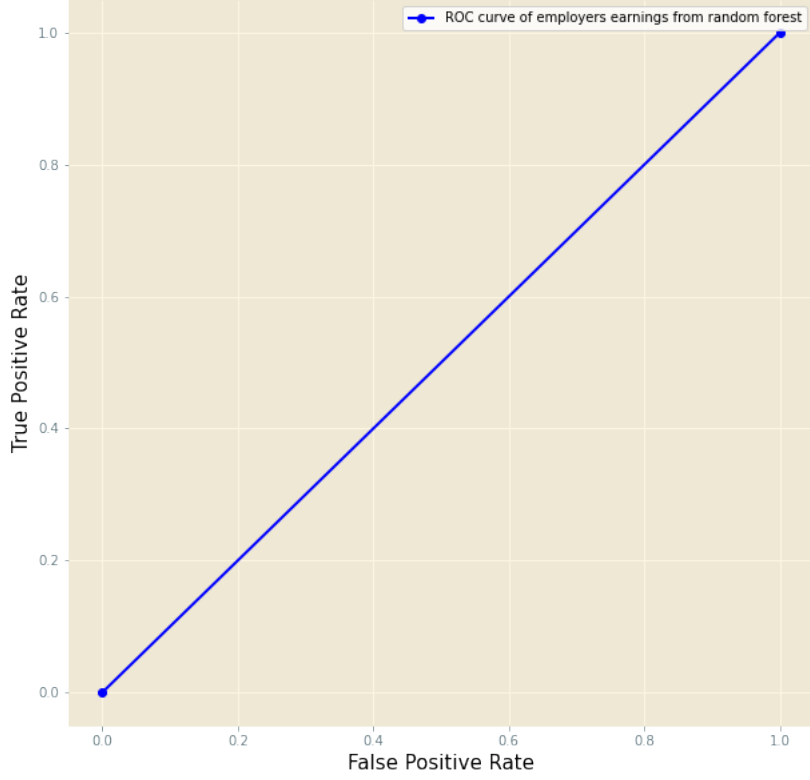


Figure 5.8: ROC of employer's earnings rate using random forest

## 5.2 Result of Regressors

To figure out how much loss a country might face for a pandemic effect we have used two different regressors recurrent neural network and random forest. In this section, we are going to discuss experimental results found from them described in chapter 4. The formulas for calculating the above values for both regressor are,

$$Accuracy = \frac{Correctly\ predicted\ class * 100\%}{Total\ testing\ class} \quad (5.4)$$

$$Mean\ absolute\ error = \left( \sum_{i=1}^n y_i - k_i \right) / n \quad (5.5)$$

Here,  $y_i$ = predicted value  $k_i$ = true value and  $n$ =total number of data points

$$Mean\ Squared\ error = \left( \sum_{i=1}^n (y_i - y_s)^2 \right) / n \quad (5.6)$$

Here  $y_s$ = predicted value  $y_i$ = true value and  $n$  = total number of data points

$$Root\ Mean\ Squared\ error = \sqrt{\left( \sum_{i=1}^n (y_i - y_s)^2 \right) / n} \quad (5.7)$$

Here  $y_s$ = predicted value  $y_i$  = true value and  $n$  = total number of data points

## 5.2.1 RNN

On the other hand, by using the LSTM regression model, we got an accuracy of 90.04% for employment level, 90.33% for worker earnings, and 92.62% for national revenue. In, LSTM revenue model, we get 2.63% mean absolute error, 0.12% mean squared error, and 3.46% root mean squared error. On the other hand, in the LSTM employment rate model, we get 4.87% mean absolute error, 0.24% mean squared error, and 4.89% root means squared error. Finally, in the LSTM earnings model, we get 4.68% mean absolute error, 0.22% mean squared error, and 4.72% root means squared error.

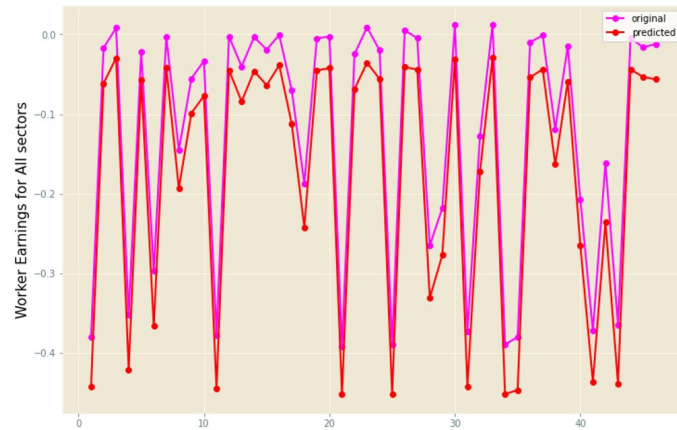


Figure 5.9: Comparison of real and modeled (LSTM) data of worker earnings of all sectors.

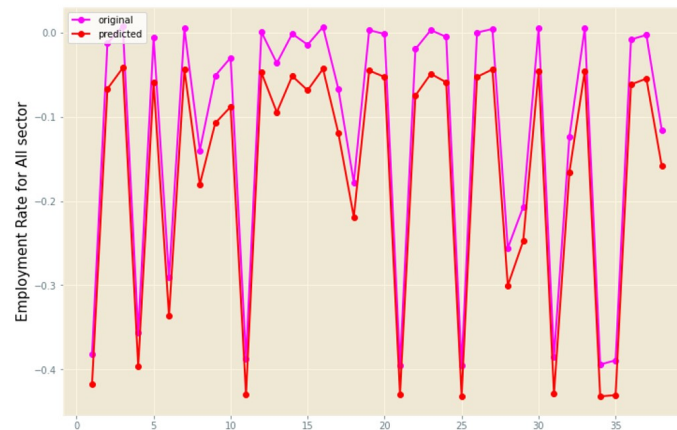


Figure 5.10: Comparison of real and modeled (LSTM) data of employment rate of all sectors.

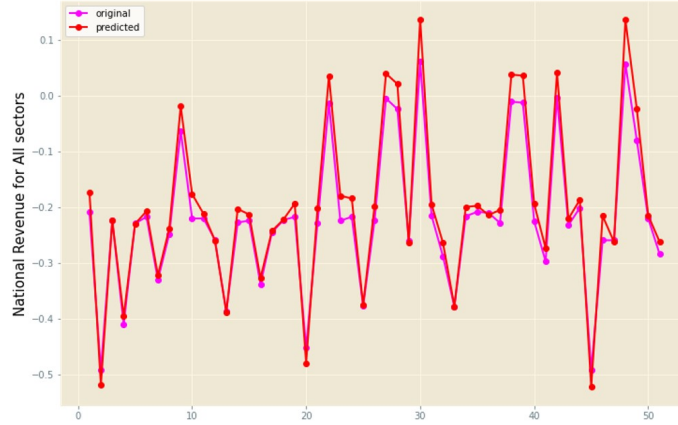


Figure 5.11: Comparison of real and modeled (LSTM) data of national revenue all sectors.

Section	Accuracy	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Employment level	90.04%	4.87%	0.24%	4.89%
Worker Earnings	90.33%	4.68%	0.22%	4.72%
National Revenue	92.62%	2.63%	0.12%	3.46%

Table 5.3: Results from LSTM regression

## 5.2.2 RFR

By following Random Forest regression methodology, we will discuss achieved results in detail. We will also include essential factors for achieving these results. Firstly we import random forest regressor class from the Sk-learn ensemble package then created random forest object and pass n-estimators, which tells us how many trees we need to include in our forest so here we have taken 200 trees, and random-state whose value is kept as 0. After that, we fit the random forest regressor with the training data and do prediction using our test data. We also imputed the missing values using mean. Lastly, we did the cross-validation to assess the effectiveness of our data.

From the Random Forest Regression, we get an accuracy of 93.16% from worker's earnings, 93.18% from employment level and 95.53% from national revenue. We can see that our accuracy of prediction for the employment level, worker earnings, and national revenue is consistent, and the error percentage in these three sectors is less.

Section	Accuracy	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
Employment level	93.18%	3.63%	0.16%	4.05%
Worker Earnings	93.16%	3.65%	0.15%	3.97%
National Revenue	95.53%	2.02%	0.07%	2.69%

Table 5.4: Results from Random forest regression

Lastly, we did the cross-validation by splitting the data-set into five-folds and trained four folds and testing one fold. Finally, we take the average of the results obtained and found that our k fold validation score for revenue is 0.996, for earnings is 0.999, and for employment rate, it is 0.998. We can see that our accuracy of prediction

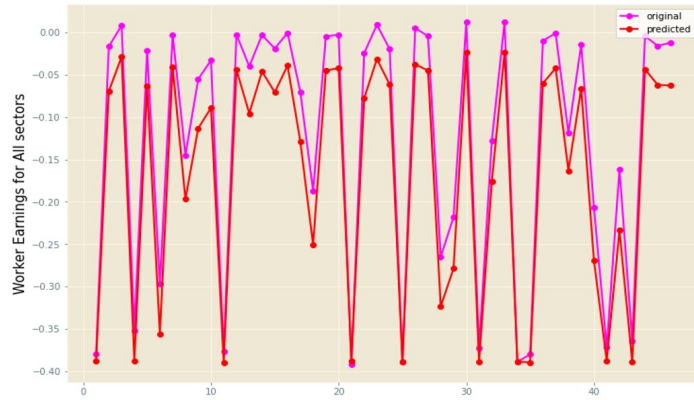


Figure 5.12: Comparison of real and modeled (Random Forest Regression) data of worker earnings of all sectors.

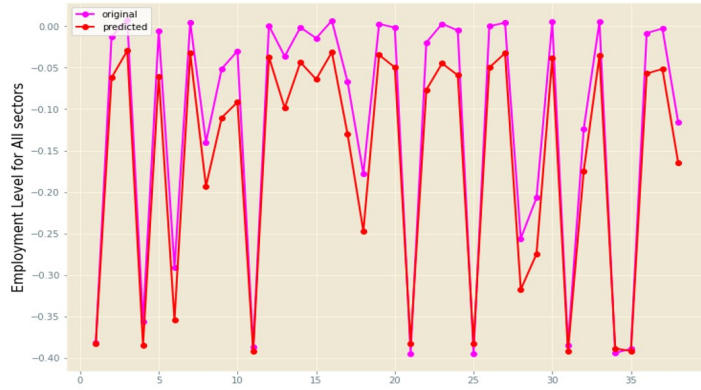


Figure 5.13: Comparison of real and modeled (Random Forest Regression) data of employment rate of all sectors.

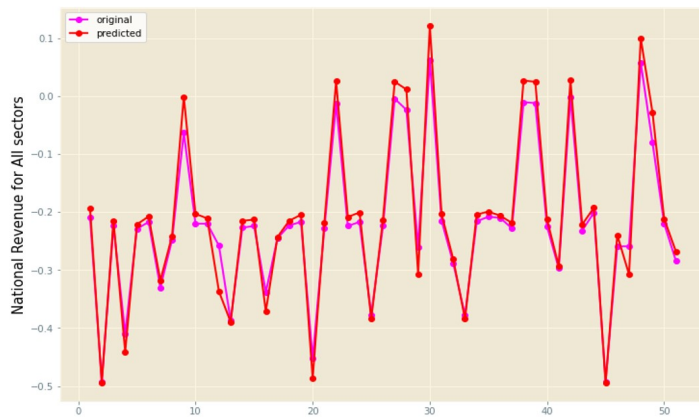


Figure 5.14: Comparison of real and modeled (Random Forest Regression) data of national revenue all sectors.



for the employment level, worker earnings, and national revenue is consistent, and error percentage is low.

### 5.3 Discussion

The goal of our model is to provide approximately 8 months (239 days) prediction of economic indicators. The models achieved demonstrate that it is possible to achieve a quality model using economic indicators as inputs through multilayer perceptron (MLP), long short-term memory (LSTM) RNN, and Random forest classifier and Regressor. The models for revenue, employment rate, and earnings use the same learning rate. When more data is available, a new prediction can be produced with the latest data set by the MLP and LSTM algorithms. The findings demonstrate the potential to use these algorithms in future to model almost the same phenomenon. On all available data, we have developed our model solution, which is incredibly constrained. Although the volume of data is minimal, in all three situations, we have perfect accuracy.

In the MLP classifier model, we got 81% accuracy for employment level, 89% for worker earnings and 95% for national revenue. Although, after applying k-fold employment level model accuracy and worker earnings model accuracy increased but national revenue model accuracy decreased from 95% to 88%. On the other hand, LSTM revenue got 92.62%, LSTM employment rate got 90.04%, and LSTM earnings got 90.33%, which is very much appreciable with this limited amount of data. Our future plan is to include one year of training data and validation data to improve our models so that we can decrease the data limitation problem. Besides, for higher accuracy in the future, we will improve our model by including CNN and bidirectional LSTM recurrent neural networks in our current model.

For the global pandemic situation like COVID-19, we used different sector reports as inputs. Including revenue, employment rate, and earnings data set, we get a proper prediction of the economic recession outbreak. Like the LSTM, MLP model, we use Random forest classifier and regression for learning the same rate of revenue, employment rate, and earnings data. Our model has done an excellent job of predicting the most recent data. The model gives us a positive indicator for foreseeing the economic recession. Though we have used limited data for this prediction, we have got excellent accuracy in all the cases. The advantage of Random Forest the classifier is its tall precision [7] for multi-class classification with a good number of training data-set, which is of the highest need. In addition, as the Random Forest classifier builds different decision trees, and the ultimate result is assessed depending on the voting of these trees, the issue of overfitting happening in a single choice tree approach is killed. The random selection of feature vectors and random choice of features during learning makes the Random Forest classifier and Regressor strong and productive for any sort of dataset [7] Forecast analysis from [31] appears an accuracy of up to 90% in foreseeing classes with an ensemble approach. By using the Random Forest classifier model, we got an accuracy of 81% for employment level, 89% for worker earnings, and 85% for national revenue, and by using the Random forest regression model, we got an accuracy of 95.53% for national revenue, 93.18% for employment level and 93.16% for worker earnings. After applying k-fold cross

validation where  $k=5$ , suddenly accuracy level showed different results for random forest classifier. Taking the mean value of 5 folds we get, 98.94% for employment level, 99.04% for worker earnings, and 98.43% for national revenue, which is very much appreciable with this limited amount of data. Our future plan is to include one year of training data and validation data to improve our model by limiting our problem. Thus, to conclude, Random forest classifier and regression is a useful tool for economists and practitioners dealing with forecasting economic recession prediction.

We did provide a warning model system for increasing awareness of an upcoming shock event. Economic crisis prediction is essential for both practitioners as well as policymakers since it provides an in-depth understanding of economic linkage breakdown after a crisis may significantly facilitate administration strategy selection as well as the development of dependent plans that can be executed, saving the economy to a greater extent. We proposed a system where we selected some crucial indicators that can be used to predict the economic crisis and choose a machine learning technique that is capable of recognizing the consequence of the economic crisis, thus giving scope to policymakers as well as practitioners to rethink their business strategy. As we know that economic conditions are continually changing, and during a pandemic, it changes drastically, so economic crisis prediction will remain an open research issue with many situations and challenges to address. We need to continuously train our data-set with a more diverse set of machine learning algorithms as well as deep learning architectures that will benefit our investigation. The use of a developed and more robust technique will enhance the prediction of the forecasting economy. Finally, a large amount of data-set will allow us to predict more accurately by filtering out noise embedded in our time series data. Furthermore, cross-validation and comparison of training and testing data sets are essential since it helps one truly understand the models' meaning and efficiency and thus improve the outcome's reliability. There is a long way to design models that forecast the global crisis around a pandemic using machine learning. In our prospective research aspirants, the importance of these latest technologies needs to be discussed.

# Chapter 6

## Conclusion and Future research

In this chapter we will give discuss about our future plan about this model and how we will improve this model in order to help out a country from facing recession.

### 6.1 Conclusion

As several studies demonstrate that pandemic outbreaks caused considerable economic changes, we studied the data. We developed early prediction model to forecast the economic crisis with training that can help save substantial capital on land, including random forest regression, classification, ANN, and LSTM. We use different machine learning algorithms in the data set of employment, earning, and revenue, which predicts economic outbreaks. Our factual assessment is preliminary and relies on our limited knowledge about the pandemic outbreak and constrained methodology. With the constrained sum of information, we have succeeded in creating a demonstration that gives an excellent precision rate. Our models could give an excellent prediction for finding the recession of the economy. We use robust algorithms such as random forest regression, classifier, ANN, LSTM, which generate a higher accuracy. It is a profoundly useful tool for economists and practitioners for predicting the economic recession prediction. The results of this study are promising. In addition, more recent articles on the prediction of the financial crisis have been discussed in this article. This methodological analysis has contrasted linked experiments in terms of data sets in particular. An additional significant challenge is careful preparation of the data collection for experiments. It is genuinely a challenging task to construct such an early caution predictive system that can be utilized for essential policy details. Several elective predictive variables may be analyzed, the foremost critical being a short-term outside obligation, for which information is not readily accessible. To conclude, we must go beyond the rough estimates that currently permeate commentary on pandemic disease outbreaks' economic consequences.

### 6.2 Future Work

To improve this model by limiting our problem, we plan to include one year of training data and validation data. Moreover, future inquiries can be made by combining MLP, LSTM, Random forest with other classification models.

# Bibliography

- [1] J. A. Ohlson, “Financial ratios and the probabilistic prediction of bankruptcy,” *Journal of accounting research*, pp. 109–131, 1980.
- [2] R. C. West, “A factor-analytic approach to bank condition,” *Journal of Banking & Finance*, vol. 9, no. 2, pp. 253–266, 1985.
- [3] L. Breiman, “Bagging predictors,” *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [4] F. A. Gers, J. Schmidhuber, and F. Cummins, “Continual prediction using lstm with forget gates,” in *Neural Nets WIRN Vietri-99*, Springer, 1999, pp. 133–138.
- [5] G. Zhang, M. Y. Hu, B. E. Patuwo, and D. C. Indro, “Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis,” *European journal of operational research*, vol. 116, no. 1, pp. 16–32, 1999.
- [6] A. F. Atiya, “Bankruptcy prediction for credit risk using neural networks: A survey and new results,” *IEEE Transactions on neural networks*, vol. 12, no. 4, pp. 929–935, 2001.
- [7] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [8] S. Sarkar and R. S. Sriram, “Bayesian models for early warning of bank failures,” *Management Science*, vol. 47, no. 11, pp. 1457–1475, 2001.
- [9] D. Hanna and Y. Huang, “The impact of sars on asian economies,” *Asian Economic Papers*, vol. 3, no. 1, pp. 102–112, 2004.
- [10] J. H. Min and Y.-C. Lee, “Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters,” *Expert systems with applications*, vol. 28, no. 4, pp. 603–614, 2005.
- [11] K.-S. Shin, T. S. Lee, and H.-j. Kim, “An application of support vector machines in bankruptcy prediction model,” *Expert systems with applications*, vol. 28, no. 1, pp. 127–135, 2005.
- [12] S. Balcaen and H. Ooghe, “35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems,” *The British Accounting Review*, vol. 38, no. 1, pp. 63–93, 2006.
- [13] P. R. Kumar and V. Ravi, “Bankruptcy prediction in banks and firms via statistical and intelligent techniques—a review,” *European journal of operational research*, vol. 180, no. 1, pp. 1–28, 2007.

- [14] M. R. Keogh-Brown and R. D. Smith, “The economic impact of sars: How does the reality match the predictions?” *Health policy*, vol. 88, no. 1, pp. 110–120, 2008.
- [15] C.-F. Tsai and J.-W. Wu, “Using neural network ensembles for bankruptcy prediction and credit scoring,” *Expert systems with applications*, vol. 34, no. 4, pp. 2639–2649, 2008.
- [16] U. Grömping, “Variable importance assessment in regression: Linear regression versus random forest,” *The American Statistician*, vol. 63, no. 4, pp. 308–319, 2009.
- [17] S. Haykin, *Neural Networks and Learning Machines, 3/E*. Pearson Education India, 2010.
- [18] R. Evans, B. Pfahringer, and G. Holmes, “Clustering for classification,” in *2011 7th international conference on information technology in Asia*, IEEE, 2011, pp. 1–8.
- [19] W.-Y. Lin, Y.-H. Hu, and C.-F. Tsai, “Machine learning in financial crisis prediction: A survey,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 421–436, 2011.
- [20] R. M. Simon, J. Subramanian, M.-C. Li, and S. Menezes, “Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data,” *Briefings in bioinformatics*, vol. 12, no. 3, pp. 203–214, 2011.
- [21] R. D. Smith, M. R. Keogh-Brown, and T. Barnett, “Estimating the economic impact of pandemic influenza: An application of the computable general equilibrium model to the uk,” *Social science & medicine*, vol. 73, no. 2, pp. 235–244, 2011.
- [22] L. Fraiwan, K. Lweesy, N. Khasawneh, H. Wenz, and H. Dickhaus, “Automated sleep stage identification system based on time–frequency analysis of a single eeg channel and random forest classifier,” *Computer methods and programs in biomedicine*, vol. 108, no. 1, pp. 10–19, 2012.
- [23] V. Y. Kulkarni and P. K. Sinha, “Pruning of random forest classifiers: A survey and future directions,” in *2012 International Conference on Data Science & Engineering (ICDSE)*, IEEE, 2012, pp. 64–68.
- [24] E. Fedorova, E. Gilenko, and S. Dovzhenko, “Bankruptcy prediction for russian companies: Application of combined classifiers,” *Expert Systems with Applications*, vol. 40, no. 18, pp. 7285–7293, 2013.
- [25] J. Abellán and C. J. Mantas, “Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring,” *Expert Systems with Applications*, vol. 41, no. 8, pp. 3825–3830, 2014.
- [26] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, “Do we need hundreds of classifiers to solve real world classification problems?” *The journal of machine learning research*, vol. 15, no. 1, pp. 3133–3181, 2014.
- [27] Y. E. Cakra and B. D. Trisedya, “Stock price prediction using linear regression based on sentiment analysis,” in *2015 international conference on advanced computer science and information systems (ICACSIS)*, IEEE, 2015, pp. 147–154.

- [28] D. L. Heymann, L. Chen, K. Takemi, D. P. Fidler, J. W. Tappero, M. J. Thomas, T. A. Kenyon, T. R. Frieden, D. Yach, S. Nishtar, *et al.*, “Global health security: The wider lessons from the west african ebola virus disease epidemic,” *The Lancet*, vol. 385, no. 9980, pp. 1884–1901, 2015.
- [29] S. Hegelich, “Decision trees and random forests: Machine learning techniques to classify rare events,” *European Policy Analysis*, vol. 2, no. 1, pp. 98–120, 2016.
- [30] I. Syarif, A. Prugel-Bennett, and G. Wills, “Svm parameter optimization using grid search and genetic algorithm to improve classification performance,” *Telkomnika*, vol. 14, no. 4, p. 1502, 2016.
- [31] J. Kevric and A. Subasi, “Comparison of signal decomposition methods in classification of eeg signals for motor-imagery bci system,” *Biomedical Signal Processing and Control*, vol. 31, pp. 398–406, 2017.
- [32] R. Nyman and P. Ormerod, “Predicting economic recessions using machine learning algorithms,” *arXiv preprint arXiv:1701.01428*, 2017.
- [33] F. Prager, D. Wei, and A. Rose, “Total economic consequences of an influenza outbreak in the united states,” *Risk Analysis*, vol. 37, no. 1, pp. 4–19, 2017.
- [34] B. Wang, L. Gao, and Z. Juan, “Travel mode detection using gps data and socioeconomic attributes based on a random forest classifier,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1547–1558, 2017.
- [35] K. Suri and R. Gupta, “Transfer learning for semg-based hand gesture classification using deep learning in a master-slave architecture,” in *2018 3rd International Conference on Contemporary Computing and Informatics (IC3I)*, IEEE, 2018, pp. 178–183.
- [36] J. Uthayakumar, N. Metawa, K. Shankar, and S. Lakshmanprabu, “Intelligent hybrid model for financial crisis prediction using machine learning techniques,” *Information Systems and e-Business Management*, pp. 1–29, 2018.
- [37] H. Cheng, Z. Xie, L. Wu, Z. Yu, and R. Li, “Data prediction model in wireless sensor networks based on bidirectional lstm,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, p. 203, 2019.
- [38] K. Bluwstein, M. Buckmann, A. Joseph, M. Kang, S. Kapadia, and Ö. Simsek, “Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach,” 2020.
- [39] Z. Car, S. Baressi Šegota, N. Anđelić, I. Lorencin, and V. Mrzljak, “Modeling the spread of covid-19 infection using a multilayer perceptron,” *Computational and Mathematical Methods in Medicine*, vol. 2020, 2020.
- [40] R. Chetty, J. Friedman, N. Hendren, and M. Stepner, “The economic impacts of covid-19: Evidence from a new public database built from private sector data,” *Opportunity Insights*, 2020.
- [41] S. J. Fong, G. Li, N. Dey, R. G. Crespo, and E. Herrera-Viedma, “Composite monte carlo decision making under high uncertainty of novel coronavirus epidemic using hybridized deep learning and fuzzy rule induction,” *Applied Soft Computing*, p. 106 282, 2020.

- [42] W. H. Organization *et al.*, “Health topics. coronavirus,” *Coronavirus: symptoms. World Health Organization, 2020a. Disponivel em: [https://www.who.int/healthtopics/coronavirus#tab=tab\\_3](https://www.who.int/healthtopics/coronavirus#tab=tab_3). Acesso em*, vol. 7, 2020.
- [43] S. Polyzos, A. Samitas, and A. E. Spyridou, “Tourism demand and the covid-19 pandemic: An lstm approach,” *Tourism Recreation Research*, pp. 1–13, 2020.
- [44] M. K. Goyal and A. K. Gupta, *Integrated risk of pandemic: Covid-19 impacts, resilience and recommendations*.