ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 7, Issue 8, August 2020

ANALYSIS AND FORECAST OF COVID-19 RECOVERY CASES USING MACHINE-LEARNING PREDICTIVE MODELS – US CASE STUDY

[1] SrinathVallakirthy

[1] Healthcare Data IT Futurist, Texas Health Resources, Dallas, TX, United States

srivallak.sm@gmail.com

Abstract: This study analyzes COVID-19 data and predicts the recovered cases in the US using machine-learning models. We employ three most common base learners, namely, linear regression, MLP and SVM, for analyzing and predicting COVID-19 recovered cases. Utilizing these linear regression and SVM base learners, we predict the recovered cases that are very satisfying and close to the exact recovered case. However, the MLP model predicted values are lower than that of actual values because of the low number of recovered cases data available. The predicted values are compared for their accuracy in predicting future recovery cases in the US. The linear regression model is found to be most accurate in comparison to the other models in this study. The key findings of this study will help in understanding the trend in coronavirus spread in the US and its recovery rate. It can help by providing a piece of valuable information to healthcare authorities and workers to design appropriate strategies for reducing the death toll and understanding the recovery rate of coronavirus patients in the US.

1. INTRODUCTION

Novel coronavirus has spread to almost all countries across the globe in a short span of period. All countries have reported their hospitals and medical supporting equipment overwhelmed due to the rapid increase in the number of coronavirus patients.

COVID-19 has been observed as a highly infectious disease with a high mortality rate and quick community spread in comparison to its predecessor diseases such as MERS [1]. Over a few days, coronavirus is passed to most of the countries in the world [2]. Up to August 5 2020, more than 18,720,558 cases have been confirmed, 704,645 deaths have been reported, and 11,936,545 patients have recovered from this disease [3]. An abrupt increase in coronavirus patients has been observed in significant coronavirus epicentres, resulting an overwhelm healthcare points, hospitals and medical workers [4]. Despitethe rapid spread of this disease, no vaccine and antiviral have been discovered yet. In this lousy scenario, development of a predictive model that can analyze and forecast the coronavirus patients can help the healthcare authorities and administrator to plan appropriate strategies for medical equipment and staff for handling the current situation.

In this work, we analyze the recovery rate of coronavirus patients in the US using machine-learning techniques. We collected COVID-19 data for recovered cases in the US from Kaggle competition [5]. An exploratory data analysis of the increase in the number of recovery cases in the US is conducted. Based on historical COVID-19 data values, we develop machine-learning models to predict future recovery cases in the US. The key findings of this study can help in understanding the trend in coronavirus spread in the US. It can help by providing a piece of valuable information by predicting recovery case for next six days to healthcare authorities and workers to design appropriate strategies for reducing the death toll and understanding the recovery rate inthe US.

The structure of this paper is organized as follows. Section 2 presents the relevant literature in analyzing and forecasting coronavirus patient data. Section 3 describes COVID-19 data and visualize the trend of coronavirus patients in the US. Section 4 presents the experimental setup for conducting experiments in this work. Section 5 presents the experimental results and discusses the trend of predicted recovery cases in the US. Finally, Section 6 concludes the paper at the end.

ISSN (Online) 2394-6849

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 7, Issue 8, August 2020

2. RELATED WORK

Since the declaration of COVID-19 as a global pandemic by WHO, researchers are attempting very hard to find anti coronavirus. However, it seems to takes time for producing such medicine [6]. Governments and medical staff is making every possible effort to moderate spread of coronavirus and prepare medical equipment for supporting increasing coronavirus patient load. Predicting the number of new coronavirus patients and recovered patients can help to plan medical supporting inventory to some extent. Few papers have been published in recent days on predicting coronavirus patients as below.

Wang et al. [7] suggested a Patient Information Based Algorithm (PIBA) to estimate the death toll due to coronavirus in China. They reported that in Hubei and Wuhan death rate was about predicted 13%. At the same time, the death rate lies in 0.75% to 3% for the rest of China.

Gupta et al. [8] analyzed coronavirus spread in the US and reported a direct relationship between COVID-19 patients and temperature. They expect an abrupt decease in coronavirus patients in the summer season. However, this does not happen as expected.

Ahmar and Val [9] employed ARIMA and Sutte ARIMA to forecast COVID-19 patients over a short span of Spanish stock market. The authors reported a prediction with MAPE of 3.6% till April 16, 2020.

Ceylan [10] also employed ARIMA models to predict the number of positive patients in Italy, Spain and France. MAPE of 4% to 6% has been reported. Fanelli and Piazza in [11] predicted COVID-19 cases in Italy, France and China. The authors predicted ventilation units required in Italy in their study.

Most studies cited above attempted to predict the number of coronavirus patients. In this work, we analyzed COVID-19 data of US and analyzed the recovery cases of coronavirus patients using different machine learning models. We developed machine-learning models and predicted future recovery cases of COVID-19 patients in the US. The findings of this study enable a better understanding of COVID-19 data and plan better strategies for decreasing the increased load of coronavirus patients on medical staff and developing medical supporting equipment required for treating coronavirus patients.

3. CORONAVIRUS DATA SET ANALYSIS

For better understanding the trends of coronavirus patients, their recovery rate and forecasting recovery of COVID-19 patients inthe US, we collected data from Kaggle competition [5]. Kaggle provided data for different countries for cumulative confirmed, recovered and death cases due to coronavirus all over the world.

In this study, we focused onthe US for analyzing coronavirus patients for a period from January 22, 2020 to August 01, 2020. We converted the coronavirus patient data in time-series format. The trend of coronavirus patients from January 22, 2020 to August 01, 2020 for confirmed, recovered and death cases in the US is presented in Fig. 1-3 respectively.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 7, Issue 8, August 2020

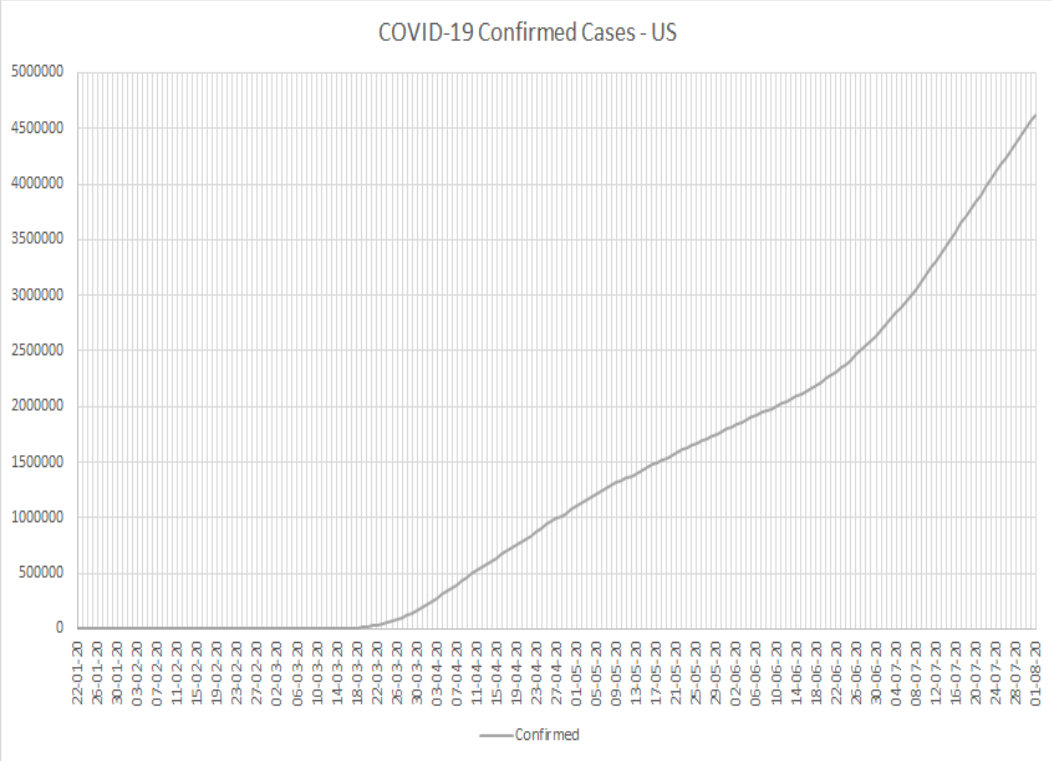

Fig. 1 Trend of COVID-19 confirmed cases in US

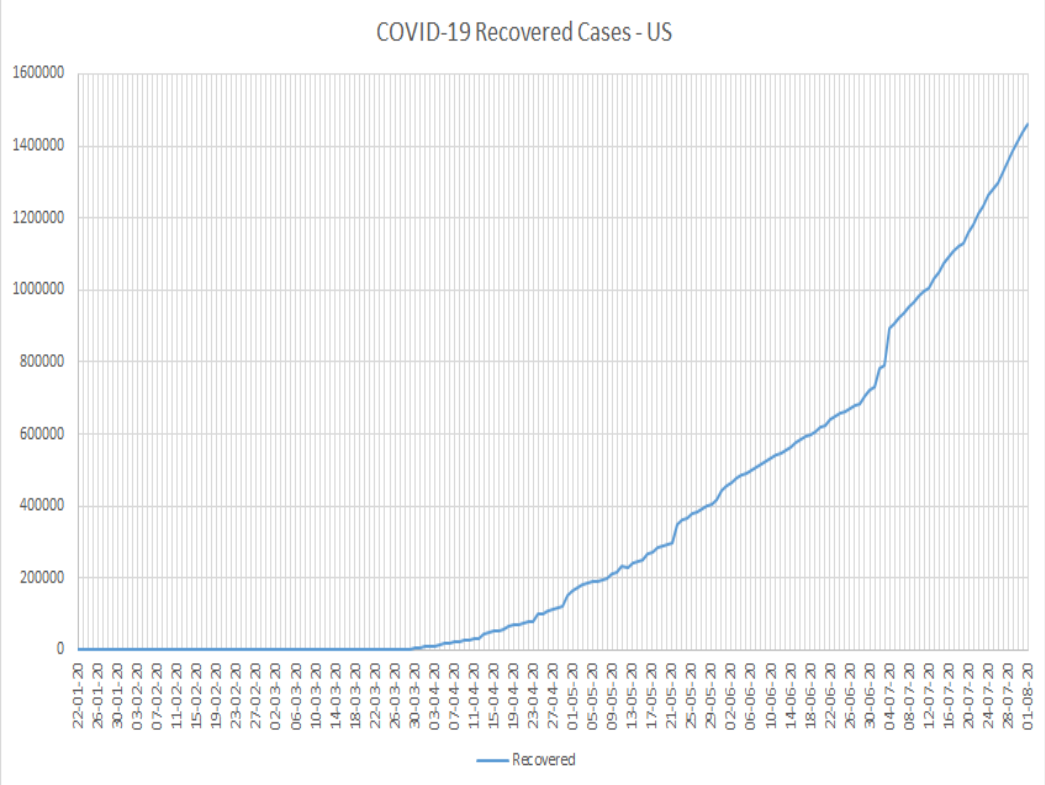

Fig. 2 Trend of COVID-19 recovered cases in US

ISSN (Online) 2394-6849

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 7, Issue 8, August 2020

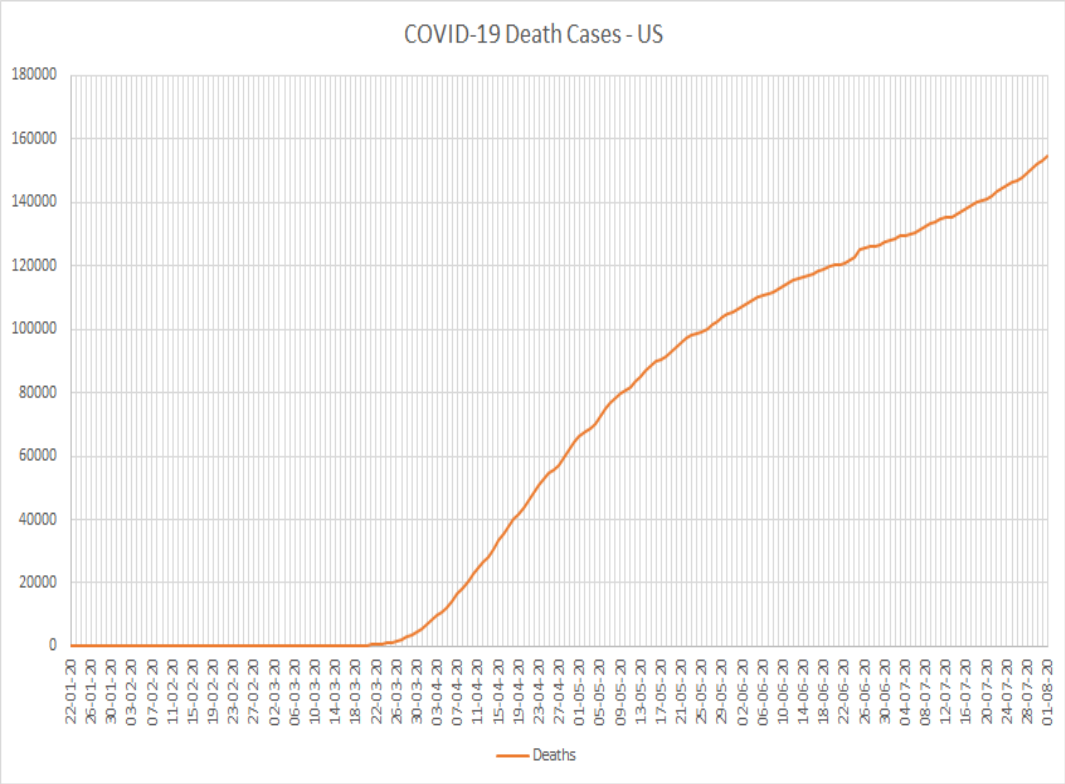

Fig. 3 The trend of COVID-19 Death cases in the US

in the US.

It can be noticed from Fig 1 that during the early days of COVID-19, there are a small number of confirmed cases in the US. However, an abrupt increase has been reported in confirmed cases in the last week of March 2020. The number of confirmed cases of COVID-19 exponentially increased from April 2020 to till date. A similar trend of recovered cases has been observed as presented in Fig. 2. Initially,the recovery rate was slow. However, it has also increased sharply after May 2020. Fig. 3 indicates that death cases in the US have also increased since March 2020. In this study, we used Kaggle COVID-19 data for the US for analyzing and forecasting the number of COVID-19 cases (confirmed, recovered and death cases)

4. EXPERIMENTAL SETUP

For performing a comprehensive analysis recovery statistics and forecasting number of COVID cases (confirmed, recovered and death cases), we used and

WEKA tool [12] for developing analytical models and forecasted future trends of COVID-19 cases in the US. Experiments have been conducted using Intel Core™ i3-2330M CPU @2.20GHz, 4GB RAM PC with Windows 10 Operating System. We used the forecast module of WEKA 3.8.4 for predicting and creating graphs of COVID-19 data as presented in Fig. 4.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 7, Issue 8, August 2020

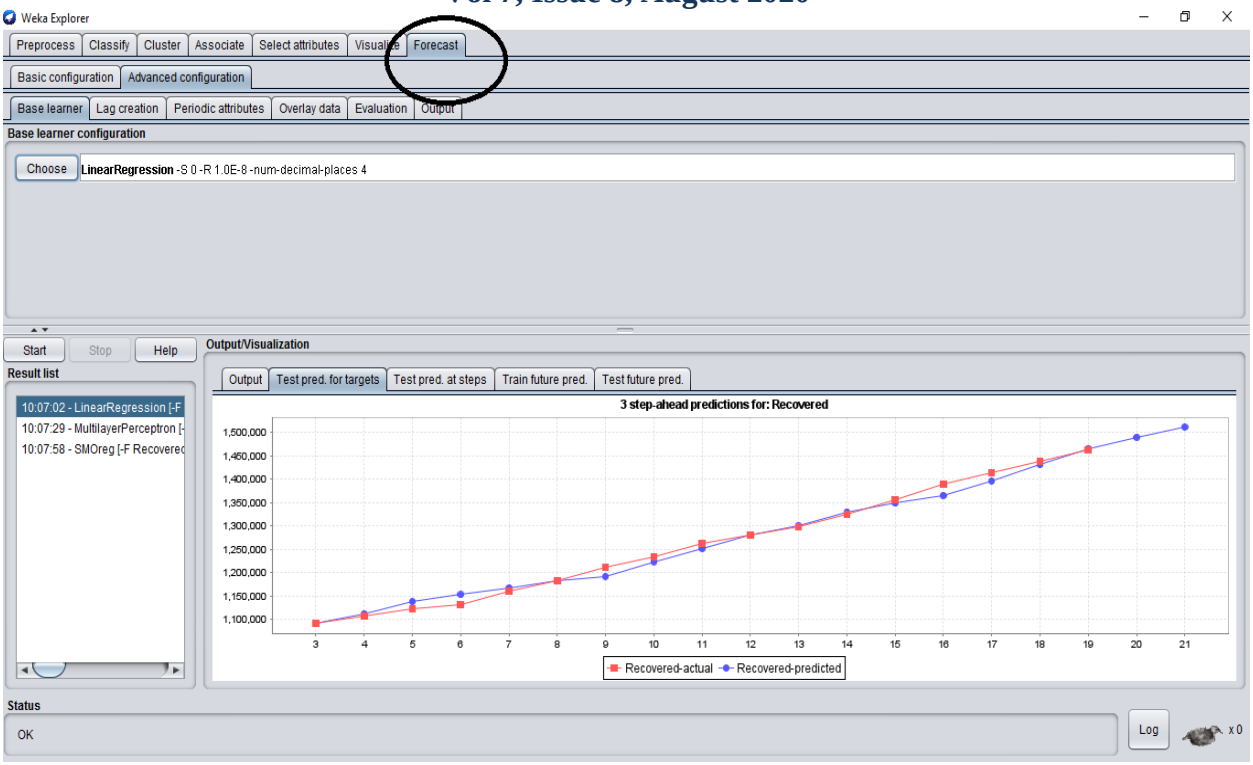

Fig.4 Forecast module of WEKA 3.8.4

To forecast accurately, we conducted three sets of experiments using Linear Regression, Multilayer Perceptron and SVM implementation of WEKA as base learner. We used 90% of available COVID-19 data as training data and rest as test data in this study. We predicted six future values from test data and computed error for different base learners. In this set of experiments, we used a 1-step ahead setting of WEKA package.

5. EXPERIMENTS AND RESULTS

In this work, we conducted three sets of experiments for analyzing and forecasting COVID-19 cases using the most commonly used base learners implemented in WEKA, namely, Linear Regression, Multilayer Perceptron and SVM. The results of these base learners are presented in the following subsections.

1.1 Linear regression-based results:

Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called a simple linear regression. For more than one explanatory variable, the process is called multiple linear regression [13]

In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models[14].In this study, we derived a linear regression model for COVID-19 recovery cases using WEKA-Forecast module based on lagged values and presented in Table 1.

Table 1. Linear regression model

COVID-19 case	Regression formula
Recovered	Recovered = 62.7748 * ArtificialTimeIndex + 0.9305 * Lag_Recovered-1 + 0.3665 * Lag_Recovered-2 + -0.2876 * Lag_Recovered-3 + -0.0996 * Lag_Recovered-4 + 0.0984 * Lag_Recovered-5 + -2128.3915

Based on the regression model presented in Table 1, the forecasted valuesrecovered cases of COVID-19 in the US for test data prediction, and future value prediction (Six values) are presented in Fig. 5 and 6, respectively.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 7, Issue 8, August 2020

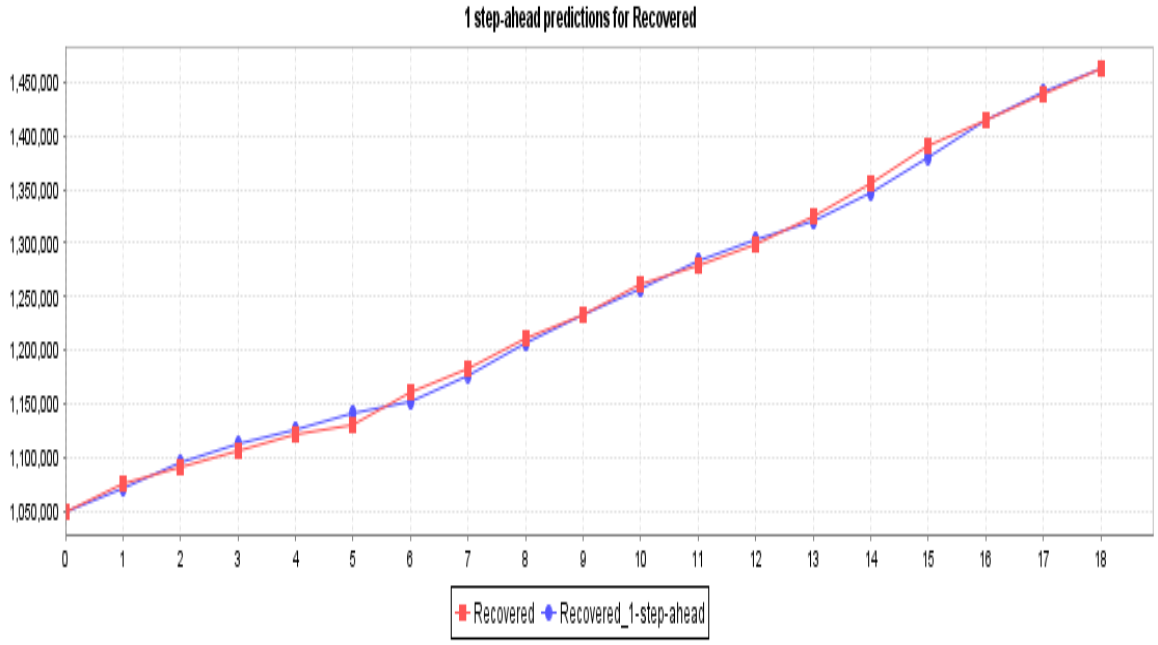

Fig. 5 Test data prediction of COVID-19 recovered cases using linear regression model

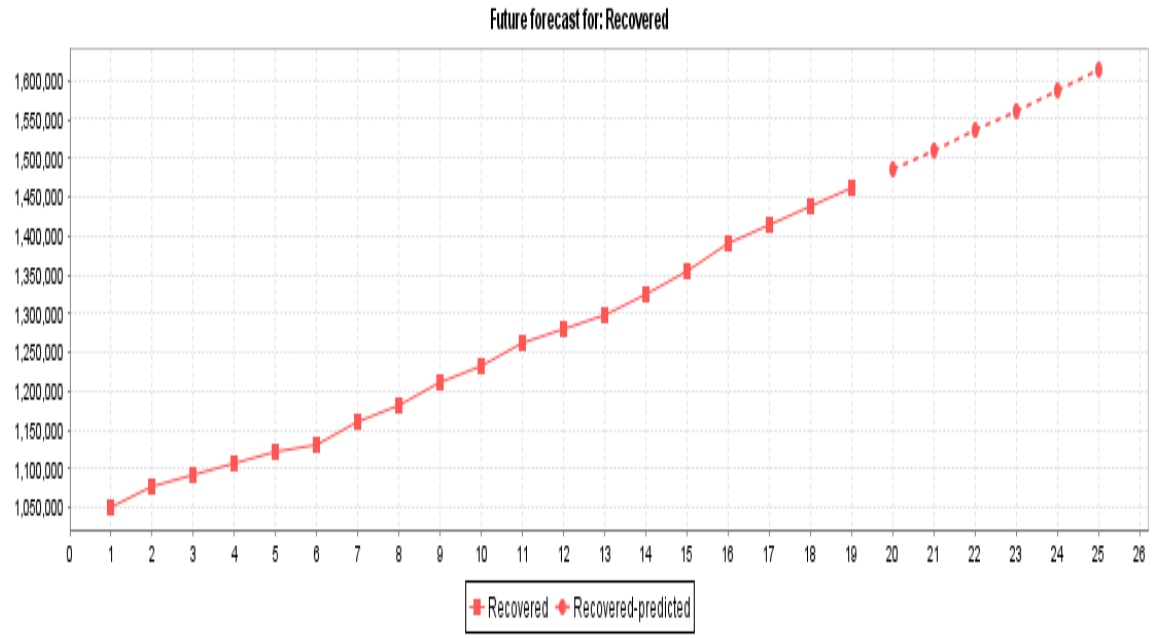

Fig. 6 Future data prediction of COVID-19 recovered cases using Linear regression model

It can be observed from Fig 5 and 6 that using linear regression model correctly predict the test values of COVID-19 data and accordingly forecast six values for recovered cases in the US. The predicted values indicate that number of recovered cases will increase in the next six days as presented in Fig. 6.

5.1 Multilayer Perceptron based results:

A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN). The term MLP is used ambiguously, sometimes loosely to any feedforward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptrons (with threshold

ISSN (Online) 2394-6849

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 7, Issue 8, August 2020

activation)

An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer as presented in Fig. 7. Except for the input nodes, each node is a neuron that uses a non-linear activation function. MLP utilizes a supervised learning technique called backpropagation for training[15][16]. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron.

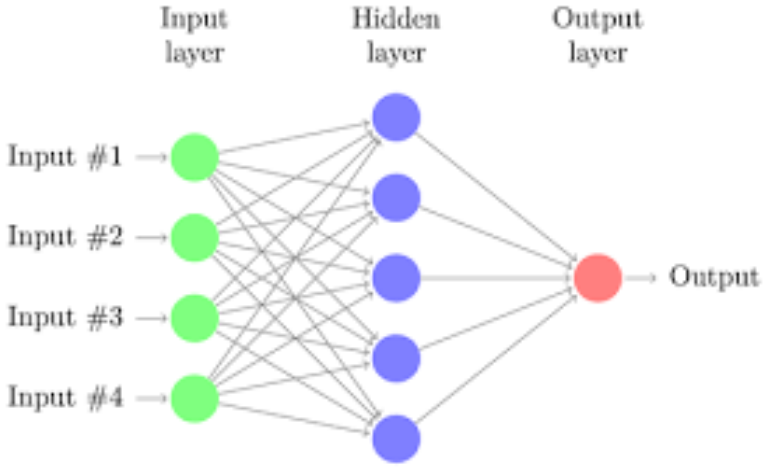

Fig. 7 MLP structure

Based on the MLP model, the forecasted values of recovered cases of COVID-19 in the US for test data prediction and future value prediction (Six values) are presented in Fig. 8 and 9, respectively.

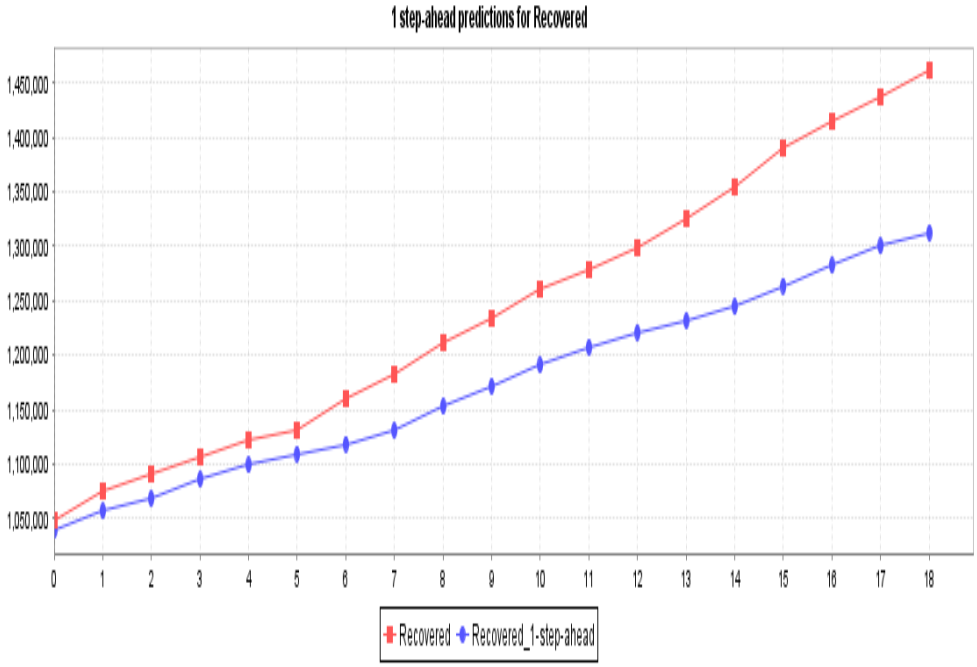

Fig. 8 Test data prediction of COVID-19 recovered cases using MLP model

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 7, Issue 8, August 2020

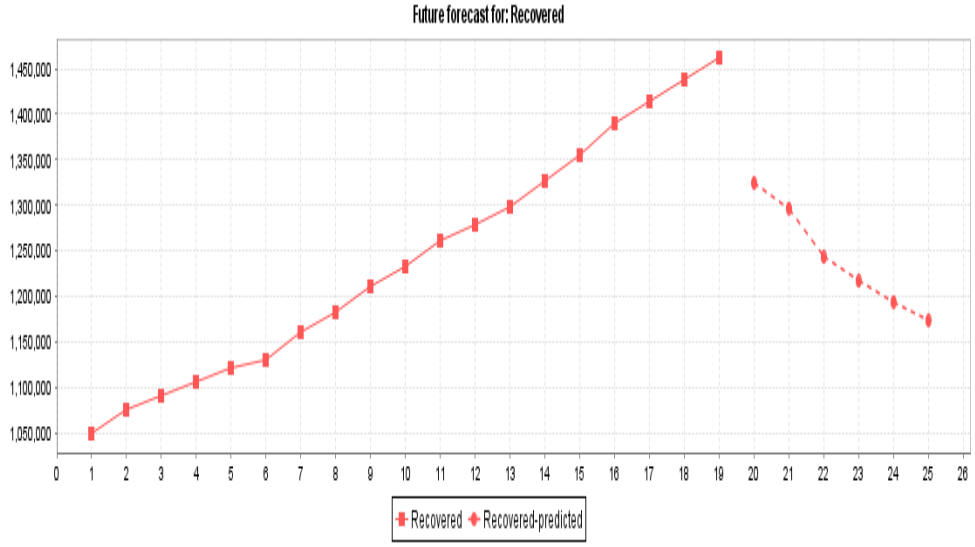

Fig. 9 Future data prediction of COVID-19 recovered cases using MLP model

It can be observed from Fig 8 and 9 that using MLP model predict lower than actual values of recovered COVID-19 cases in the US and accordingly forecasted six values for recovered cases in the US are also less than actual values. The predicted values indicate that number of recovered cases will decrease in the next six days in comparison to previous values as presented in Fig. 9. However, the comparison with actual values proves that MLP is not modelled correctly may be due to the low number of instance in the training dataset.

1.2 SVM Regression-based results

Support-vector machines (SVM [17, [18]) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. The Support Vector Machine (SVM) algorithm is a popular machine learning tool that offers solutions for both classification and regression problems. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible [19][20]. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

WEKA implements support vector machine for regression.

Based on the SVM model, the forecasted values of recovered cases of COVID-19 in the US for test data prediction and future value prediction (Six values) are presented in Fig. 10 and 11 respectively.

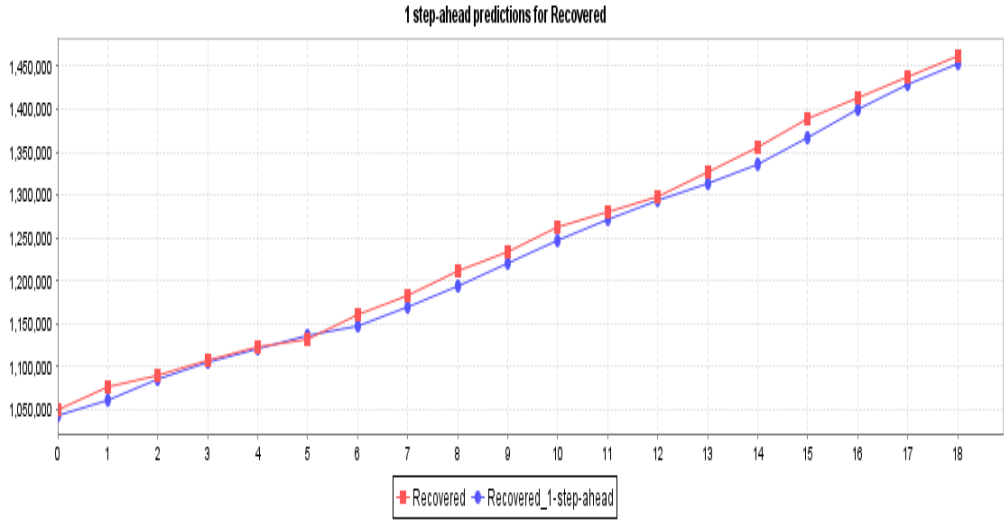

Fig. 10 Test data prediction of COVID-19 recovered cases using SVM model

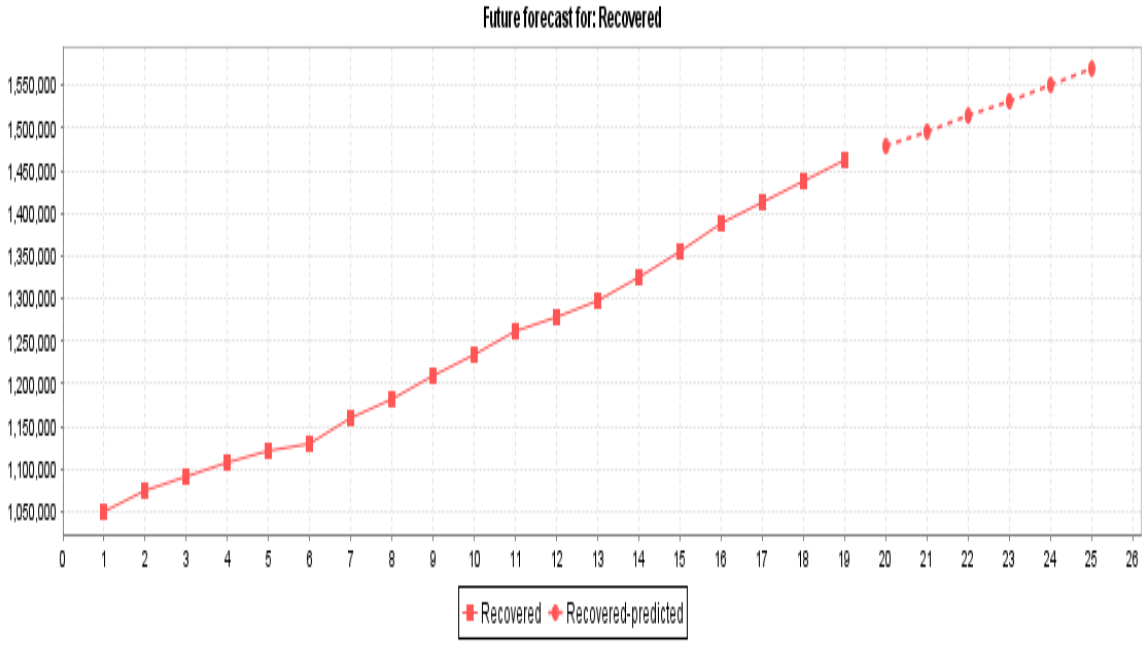

Fig. 11 Future data prediction of COVID-19 recovered cases using SVM model

It can be observed from Fig 10 and 11 that using SVM model predicts recovered COVID-19 cases near to the actual recovered cases and accordingly forecasted six values for recovered cases in the US. The predicted values indicate that number of recovered cases will increase in the next six days in comparison to previous values as presented in Fig. 11.

Table 2 presents comparative values of predictions of recovered cases in the US using linear regression, MLP and SVM along with their six future values. It can be noted from Table 2 that the future predicted values of linear regression and SVM are approximately close to the exact values, indicating an increase in the number of COVID-19 recovered cases in the US for next six days. However, MLP is unable to model the correct future prediction in comparison to linear regression and SVM models.

Table 2. Comparative predicted recovered cases in the US

Date	Actual	Prediction		
		Linear regression	MLP	SVM
19-07-20	1131121	1153055	1066862	1113792
20-07-20	1160087	1174315	1072395	1132819
21-07-20	1182018	1200788	1077991	1157883
22-07-20	1210849	1216652	1086998	1178200
23-07-20	1233269	1231882	1092729	1195200
24-07-20	1261624	1249222	1100487	1212440
25-07-20	1279414	1258545	1104665	1222707
26-07-20	1297863	1290539	1111471	1247577
27-07-20	1325804	1318913	1118518	1272644
28-07-20	1355363	1347500	1123518	1300320
29-07-20	1389425	1370716	1130607	1327085
30-07-20	1414155	1400403	1135973	1354786

31-07-20	1438160	1419506	1143125	1377424
01-08-20	1461885	1437106	1147394	1396776
Future predicted values	?	1468970	1153454	1424177
	?	1504199	1157733	1455011
	?	1540368	1161625	1489138
	?	1565383	1165603	1518523
	?	1588162	1169237	1544124

CONCLUSION

In this study, we analyzed COVID-19 data for recovered cases in the US using data from Kaggle competition. We used the forecast module of WEKA for predicting the future recovered cases in the US. We employed three most common base learners, namely, linear regression, MLP and SVM, for analyzing and predicting COVID-19 recovered cases. Utilizing these linear regression and SVM base learners, we predicted the recovered cases that are very satisfying and close to the exact recovered case. However, the MLP model predicted values are lower than that of actual values because of the low number of recovered cases data available. As per prediction provided by the linear regression model, the recovered patients on August 01 should be 1437106, which is close to the exact cases 1461885. As per the MLP model, the recovered patients on August 01 should be1147394, whereas SVM predicts this value as 1396775. The linear regression model is found to be most accurate in comparison to the other models in this study. The key findings of this study will help in understanding the trend in coronavirus spread in the US and its recovery rate. It can help by providing a piece of valuable information to healthcare authorities and workers to design appropriate strategies for reducing the death toll and understanding the recovery rate of coronavirus patients in the US.

REFERENCES

[1] Mahase, E. (2020). Coronavirus: Covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate. BMJ 368, m641.

[2] World Health Organization (2020). Coronavirus Disease 2019 (COVID-19): Situation Report, 61. https://apps.who.int/iris/handle/10665/331605?show=full.

[3] Coronavirus Update (Live), https://www.worldometers.info/coronavirus/ Last accessed on August 05, 2020

[4] Capricornus EinBlick, COVID-19 Time Series Analysis, https://medium.com/@thecapricornuseinblick/covid-19-time-series-analysis-e6f3f2235e43 last accessed August 05, 2020.

[5] Kaggle competition, Novel Corona Virus 2019 Dataset, https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=time_series_covid_19_confirmed.csv last accessed August 05, 2020.

[6] Grenfell R., Drew T..Here's why it's taking so long to develop a vaccine for the new coronavirus. Science Alert Archived from the original on 28 2020.

[7] Wang L , Li J , Guo S , Xie N , Yao L , Cao Y , et al. Real-time estimation and prediction of mortality caused by COVID-19 with patient information based algorithm. Sci Total Environ 2020:138394 .

[8] Gupta S ,Raghuwanshi GS , Chanda A . Effect of weather on COVID-19 spread in the US: a prediction model for India in 2020. Sci Total Environ 2020:138860.

[9] AhmarAS , del Val EB . Sutte ARIMA: short-term forecasting method, a case: COVID-19 and stock market in spain. Sci Total Environ 2020:138883.

[10] CeylanZ . Estimation of COVID-19 prevalence in Italy, Spain, and France. Sci Total Environ 2020:138817.

[11] FanelliD , Piazza F . Analysis and forecast of COVID-19 spreading in China, Italy and France. Chaos Solitons Fractals 2020;134:109761.

[12] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.
[13] David A. Freedman (2009). Statistical Models: Theory and Practice. Cambridge University Press. p. 26. A simple regression equation has on the right hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression equation has two or more explanatory variables on the right hand side, each with its own slope coefficient.
[14] Hilary L. Seal (1967). "The historical development of the Gauss linear model". Biometrika. 54 (1/2): 1–24. doi:10.1093/biomet/54.1-2.1. JSTOR 2333849
[15] Rosenblatt, Frank. x. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961
[16] Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation". David E. Rumelhart, James L. McClelland, and the PDP research group. (editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundation. MIT Press, 1986.
[17] Cortes, Corinna; Vapnik, Vladimir N. (1995). "Support-vector networks" (PDF). Machine Learning. 20 (3): 273–297. CiteSeerX 10.1.1.15.9362. doi:10.1007/BF00994018.
[18] Ayyagari, MR, Kumar, G. (2019). “An Approach for Facial Emotion Recognition Using Heuristic and Component Analysis”, Journal of Advanced Research in Dynamical & Control Systems 11 (5), 7-14.
[19] Ayyagari, MR, Kumar, G. (2019). “Lung Cancer Detection on Computed Tomography Images Using Digital Image Processing Techniques”, Journal of Advanced Research in Dynamical & Control Systems 11 (04), 677-689.
[20] Kumar, G. (2016). “Denial of service attacks–an updated perspective”, Systems Science & Control Engineering 4 (1), 285-294.