Predictive Modeling of Covid-19 Spread with Machine Learning: A Focus on Decision Tree Accuracy

Amalia Shifa Aldila*[1], Lawrence Adi Supriyono[2]
[1,2]Universitas Jakarta Internasional, Jalan Letjen S. Parman No.1AA.
Slipi, Jakarta Barat, 081119167363
e-mail: *[1]amalia.shifa@uniji.ac.id, [2]lawrence.supriyono@uniji.ac.id

Abstrak

Virus Sars CoV-2 merupakan penyebab utama wabah Covid-19 yang pertama kali terdeteksi di Wuhan, Tiongkok, pada Desember 2019 dan dengan cepat menyebar ke seluruh dunia. Penelitian ini bertujuan untuk memprediksi jumlah kasus terkonfirmasi dan tingkat keparahan wabah dalam rentang 23 Januari hingga 10 Juni 2020. Data yang digunakan adalah dataset terbuka dari Kaggle berjudul "Global Forecasting Covid-19 Week 5". Untuk menghasilkan prediksi yang optimal, penelitian ini menguji berbagai algoritma pembelajaran mesin dan pembelajaran mendalam, yaitu Random Forest, XGBoost, Polynomial Regression, Decision Tree, ANN, dan LSTM. Kinerja model dinilai melalui skor R^2 dan Root Mean Square Error (RMSE). Hasil terbaik dicapai oleh model Decision Tree dengan skor R^2 sebesar 0,97 dan RMSE 52,57, menunjukkan akurasi tinggi dalam prediksi kasus Covid-19. Penelitian ini mengindikasikan bahwa model Decision Tree unggul dalam prediksi Covid-19 dibandingkan algoritma lain dan menawarkan potensi signifikan untuk pengembangan strategi mitigasi yang lebih efektif di masa mendatang.

***Kata kunci**: Covid-19 Forecasting, Machine Learning Models, Decision Tree Accuracy, Predictive Modeling.*

Abstract

The Sars CoV-2 virus is the primary cause of the Covid-19 outbreak, first detected in Wuhan, China, in December 2019, and rapidly spreading worldwide. This study aims to predict the number of confirmed cases and the severity of the outbreak within the period of January 23 to June 10, 2020. The data used is an open-access dataset from Kaggle titled "Global Forecasting Covid-19 Week 5". To achieve optimal predictions, this study tested various machine learning and deep learning algorithms, including Random Forest, XGBoost, Polynomial Regression, Decision Tree, ANN, and LSTM. Model performance was assessed using R^2 scores and Root Mean Square Error (RMSE). The best results were achieved by the Decision Tree model, with an R^2 score of 0.97 and an RMSE of 52.57, indicating high accuracy in predicting Covid-19 cases. This research suggests that the Decision Tree model excels in Covid-19 forecasting compared to other algorithms, offering significant potential for developing more effective mitigation strategies in future outbreaks.

***Keywords:** Covid-19 Forecasting, Machine Learning Models, Decision Tree Accuracy, Predictive Modeling.*

1. INTRODUCTION

Covid-19 is a disease caused by the Sars CoV-2 virus, which first emerged at the end of December 2019 in Wuhan, China, and began spreading worldwide by the end of February 2020. This virus originates from animals and is transmitted to humans through intermediary saliva. As a result, humans who have direct contact with animals carrying the Covid-19 virus have a high chance of being exposed to it. The increasing spread has led the World Health Organization (WHO) to classify this virus as a global pandemic, due to the rising number of cases in nearly all parts of the world [1].

Since its spread began, researchers from various fields, especially health and technology, have conducted studies to analyze the trend of this virus's spread. The use of mathematical models and technology applications, such as Artificial Intelligence (AI), including Machine Learning and Deep Learning, has been very helpful in providing accurate results, particularly in predicting the spread of the Covid-19 virus [2][3][4]. These advanced models allow for more precise projections of future outbreaks, helping authorities implement more effective public health measures.

Research on this topic is diverse, ranging from the analysis of cases in specific regions of a country to predicting the global spread of Covid-19. For instance, Fanelli and Piazza [5] analyzed the temporal spread of the Covid-19 virus in 2019 in three countries China, Italy, and France. Machine learning algorithms, particularly Random Forest, have been used by Yesilkanat [6] to estimate the future spread of Covid-19 in 190 countries, providing valuable insights for global forecasting. Wang et al. [7] also conducted studies using logistic models and machine learning techniques to predict Covid-19 cases and trends. Additionally, deep learning time-series forecasting has been applied in studies examining cases in India and the USA [8][9].

The integration of Machine Learning (ML) and Deep Learning (DL) methods has proven to be effective in predicting the Covid-19 trajectory. ML models, such as Polynomial Regression, Random Forest, XGBoost, and Decision Tree, are employed in many studies to predict the course of the virus with increasing accuracy [10][11]. The Artificial Neural Network (ANN), as a deep learning model, has also gained significant attention for its ability to handle complex non-linear data patterns and improve prediction outcomes [12][13].

In this paper, the focus is on the data related to the spread of Covid-19, including data from various regions. Both machine learning and deep learning algorithms are employed to predict the virus's spread, using models like Polynomial Regression, Random Forest, XGBoost, Decision Tree, and ANN. Performance evaluation of each model is conducted through metrics like RMSE and R^2 scores, which measure prediction accuracy and error rates [14][15].

2. RESEARCH METHODOLOGY

In this study, data was obtained from the kaggle competition COVID-19 Global Forecasting (Week5) which can be downloaded at the following link www.kaggle.com/competitions/covid19-global-forecasting-week-5. The data obtained is data on daily cases of Covid-19 from several countries with a total data of 969,640 from January 23 2020 to June 10 2020.

2.1. Processing Data

The data consists of 9 columns namely Id, County, Province State, Country Region, Population, Weight, Date, Target, and Target Value.

Table 1. Data Covid-19 Globar Forecasting Week 5

	Id	County	Province_State	Country_Region
0	1	NaN	NaN	Afghanistan
1	2	NaN	NaN	Afghanistan
2	3	NaN	NaN	Afghanistan
3	4	NaN	NaN	Afghanistan
4	5	NaN	NaN	Afghanistan

Table 2. Data Population Covid-19

Population	Weight	Date	Target	TargetValue
27657145	0.058359	2020-01-23	ConfirmedCases	0
27657145	0.583587	2020-01-23	Fatalities	0
27657145	0.058359	2020-01-24	ConfirmedCases	0
27657145	0.583587	2020-01-24	Fatalities	0
27657145	0.058359	2020-01-25	ConfirmedCases	0

Based on the results of statistical processing, it is known that in general the majority of data comes from the US, with details as follows:

```
US                895440
China               9520
Canada              3640
United Kingdom      3080
France              3080
                     ...
Uruguay              280
Timor-Leste          280
Senegal              280
Argentina            280
Zimbabwe             280
Name: Country_Region, Length: 187, dtype: int64
```

Figure 1. Total Data Base on Country Region

Considering data completeness only data from US were used in the model. Next, check for missing data.

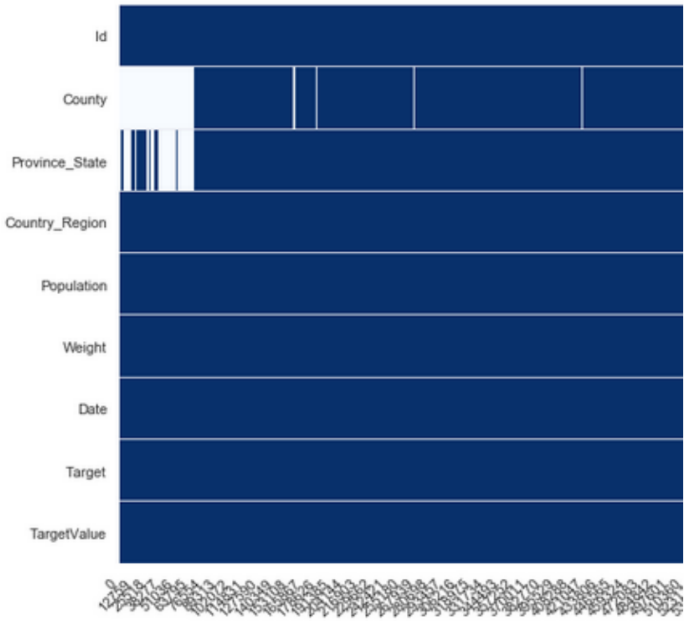

Figure 2. Lost Data Check

In this process it is known that there is a lot of missing data in the County and Province State columns, so these 2 columns are not used in making the model and only data with complete rows and Target Value >= 0 are entered into the model. Based on the target column, the modeling is divided into 2, namely the confirmed cases and fatalities modeling, so that there are 2 datasets that will be used in the modeling. Furthermore, after processing the data, the Id, Country Region, and Target columns are removed because they are considered to no longer have an effect on the model. Furthermore, for each dataset, a train-test data split process was carried out with a test_size ratio of 0.22 and scaling was also carried out on the data using the standard scaler method.

Table 3. Dataset after Processing

	Population	Weight	Date	TargetValue
67760	55869	0.091485	20200123	0
67762	55869	0.091485	20200124	0
67764	55869	0.091485	20200125	0
67766	55869	0.091485	20200126	0
67768	55869	0.091485	20200127	0

2.2. Training Model

In modeling confirmed cases and fatalities, regression models are used from several machine learning algorithms with the best tuning hyperparameters as follows:

2.2.1. Random Forest

n_estimators=100,
criterion='mse',
max_depth=None,
min_samples_split=2,
min_samples_leaf=1,
max_features='auto',
bootstrap=True

2.2.2. XGBoost

base_score=0.5,
booster='gbtree',
colsample_bytree=0.7,
learning_rate=0.01,
max_depth=10,
n_estimators=800,
objective='reg:linear',
reg_alpha=6e-05,
subsample=0.7

2.2.3. Linier Regression (Polynomial Regression)

LinearRegression(copy_X=True,
fit_intercept=True, n_jobs=None,
normalize=False, degree=2)

2.2.4. Decision Tree

DecisionTreeRegressor(ccp_alpha=0.0, criterion='mse', max_depth=None,
max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, presort='deprecated',
random_state=None, splitter='best')

2.2.5. ANN (Artificial Neural Network)

```
def create_model(num_neuron,act_func,optim):
model = Sequential()
model.add(Input(shape=(n_features,)))
model.add(Dense(num_neuron, activation=act_func))
model.add(Dense(num_neuron, activation=act_func))
model.add(Dense(num_neuron, activation=act_func))
model.add(Dense(1))
model.compile(optimizer=optim, loss='mse')
return model
optim = [“adam”]
num_neuron = [“64”]
act_func = [“tanh”]
```

2.2.6. LSTM (Long Short Term Memory)

```
generator = TimeseriesGenerator(scaled_train, scaled_train, length=n_input,
batch_size=1)
model = Sequential()
model.add(LSTM(100, activation='relu', input_shape=(n_input, n_features)))
model.add(Dense(1))
```

```
model.compile(optimizer='adam', loss='mse')
hist= model.fit(generator,epochs=epoch,verbose=1)
```

2.3. Model Evaluation

The model evaluation uses RMSE (Root Mean Square Error) and R^2 (Coefficient of Determination) calculations, with the following calculation formula:

$$RMSE_{fo} = \sqrt{\sum_{i=1}^{N} (z_{fi} - z_{oi})^2 / N}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (z_{oi} - z_{fi})^2}{\sum_{i=1}^{N} (z_{oi} - \underline{z}_o)^2} \quad \text{..... (1)}$$

With :

f = prediction score
o = observation score
$\boldsymbol{zo}$ = Mean score of observation
$\boldsymbol{zfi}$ = Prediction result
$\boldsymbol{zoi}$ = Observation result
N = Number of data

3. RESULT AND DISCUSSIONS

At the data processing stage, a correlation plot is carried out between features to show the relationship of each feature in general, for each confirmed cases and fatalities dataset, the correlation results are as follows:

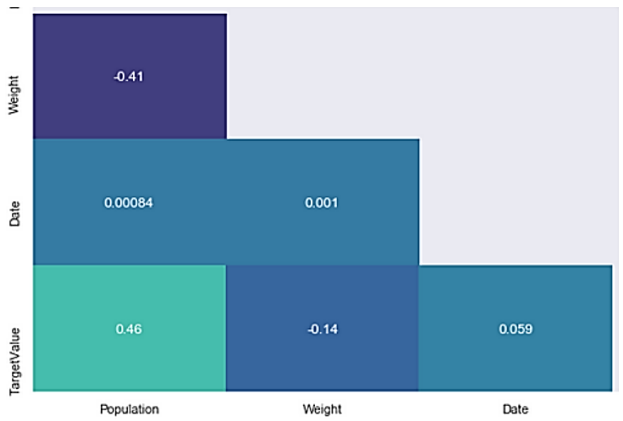

Figure 3. Confirmed Cases Dataset Correlation Plot

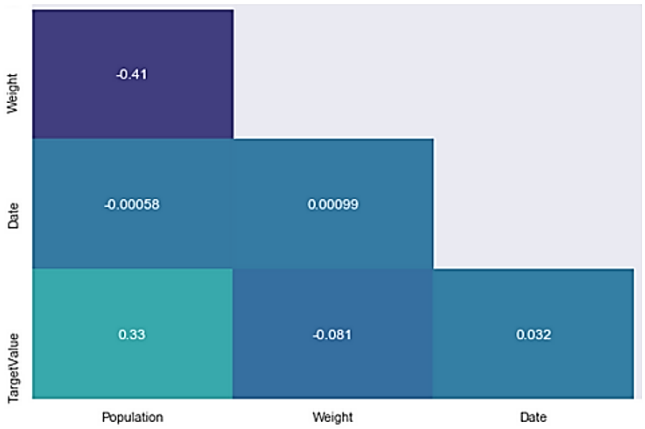

Figure 4. Fatalities Dataset Correlation Plot

Based on the results of the correlation plot, it can be seen that the two datasets (confirmed cases and

fatalities) show a similar trend of correlation where the highest correlation with the target value is achieved by the population, then followed by the weight which is negatively correlated and the lowest correlation is achieved by the date, this correlation result becomes a benchmark for determine the features that affect changes in the target value, which then the importance of these features will be validated with the results of running the model. the results of the running model for each machine learning algorithm are as follows:

3.1. Random Forest

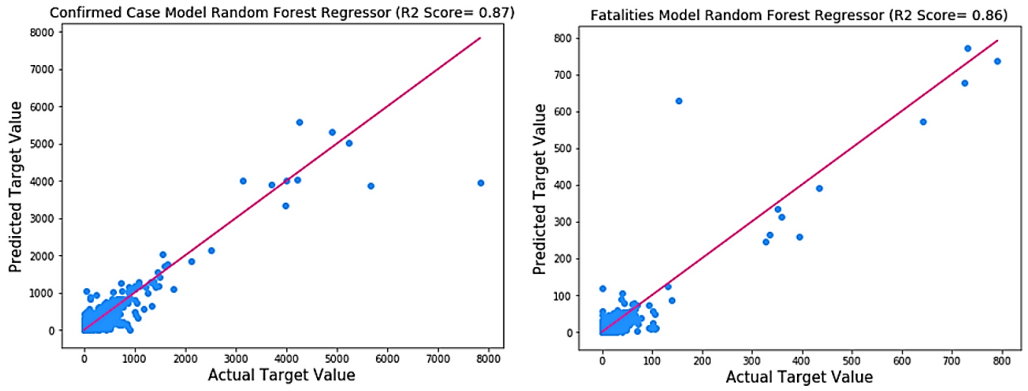

Figure 5. Random Forest R^2 Score Confirmed Cases and Fatalities

Based on the results of running the random forest model, the model accuracy values were 0.87 and 0.86 for the confirmed cases and fatalities dataset with an RMSE of 21.51 and 2.19.

3.2. XGBoost

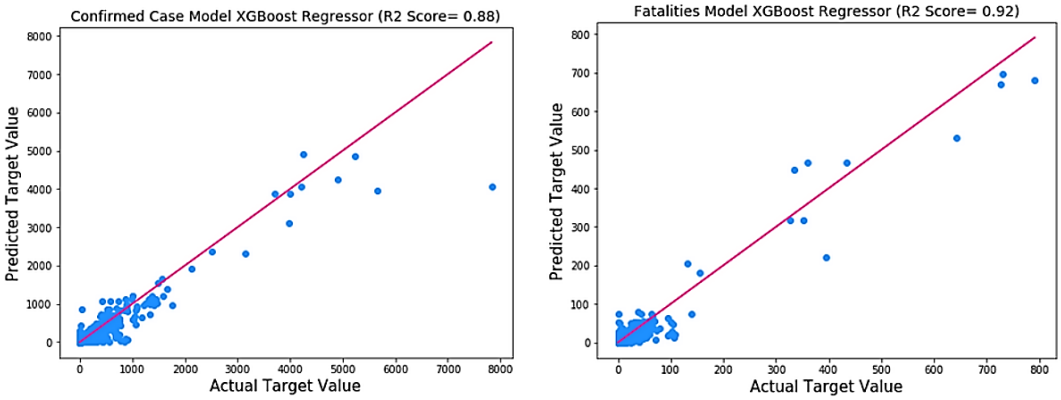

Figure 6. XGBoost R^2 Score Confirmed Cases and Fatalities

The accuracy of the XGBoost model is 0.88 and 0.92 for the confirmed cases and fatalities dataset with an RMSE of 20.65 and 1.61.

3.3. Linear Regression (Polynomial Regression)

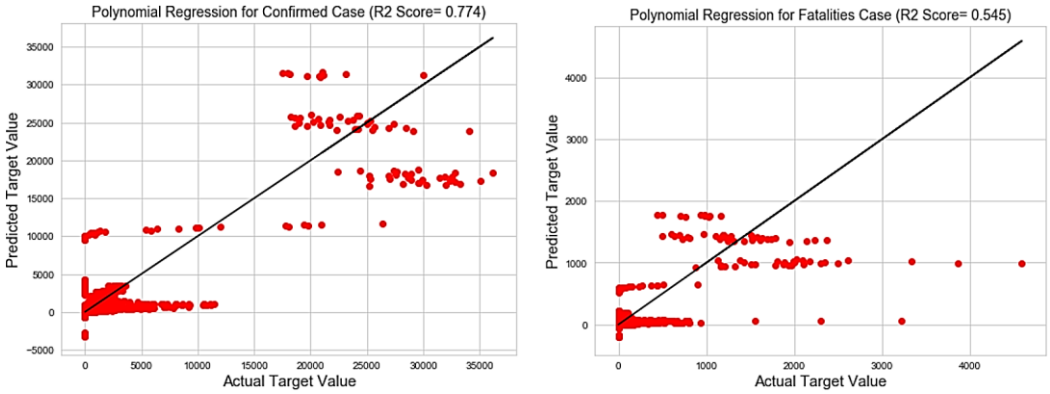

Figure 7. Polynomial Regression R^2 Score Confirmed and Fatalities

Based on the results of running the polynomial regression model, the model accuracy values were 0.77 and 0.55 for the confirmed cases and fatalities dataset with an RMSE of 169.73 and 16.22.

3.4. Decision Tree

Figure 8. Decision Tree R^2 Score Confirmed and Fatalities

Based on the results of running the decision tree model, the model accuracy values were 0.97 and 0.91 for the confirmed cases and fatalities dataset with an RMSE of 52.57 and 7.24.

3.5. ANN (Artificial Neural Network)

In the ANN model, the Confirmed Cases feature will be used as an input besides date and population. The results of the model training history and the R^2 results of the predictions can be seen in the figure 9.

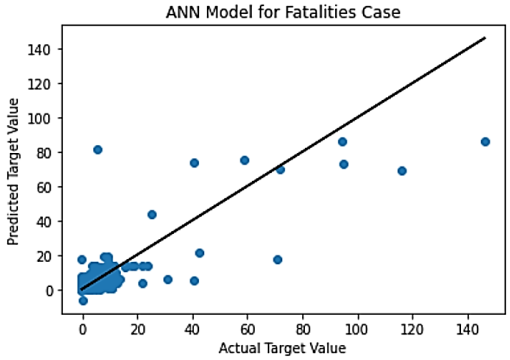

Figure 9. ANN R^2 Score Confirmed and Fatalities

The results of the test R2 and RMSE values obtained from the model training were 0.7468 and 125.025.

3.6. LSTM (Long Short-Term Memory)

To train the LSTM model, the data will first be transformed into a time series that has a length of 8 days. After that, predictions are made for the next 1 day and predictions for the 2nd day will be predicted based on the previous first prediction. The results of R2 from predictions of case fatalities and confirmed cases can be seen in Figure 10.

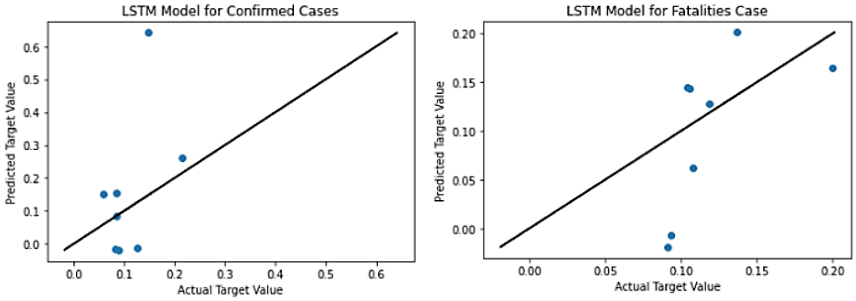

Figure 10. LSTM R^2 Score Confirmed and Fatalities

Based on the results of the LSTM model training, the model accuracy values were 0.22 and 0.28 for the confirmed cases and fatalities dataset with an RMSE of 9869.07 and 624.32. The following results are far worse than all other methods due to very limited data when the data is formed as a time series. This limitation can be seen from the amount of data which initially lasted 3 months and was formed into a time series data with 8 days as the length of one series. Based on the above model, the mean loss decrease and mean score decrease are calculated to determine the importance of each feature to the target value, and for each datasheet the following results are obtained:

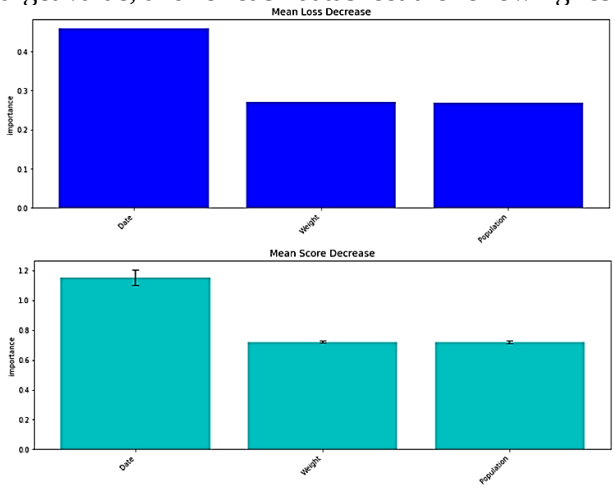

Figure 11. Mean Loss and Score Decrease Dataset Confirmed Cases

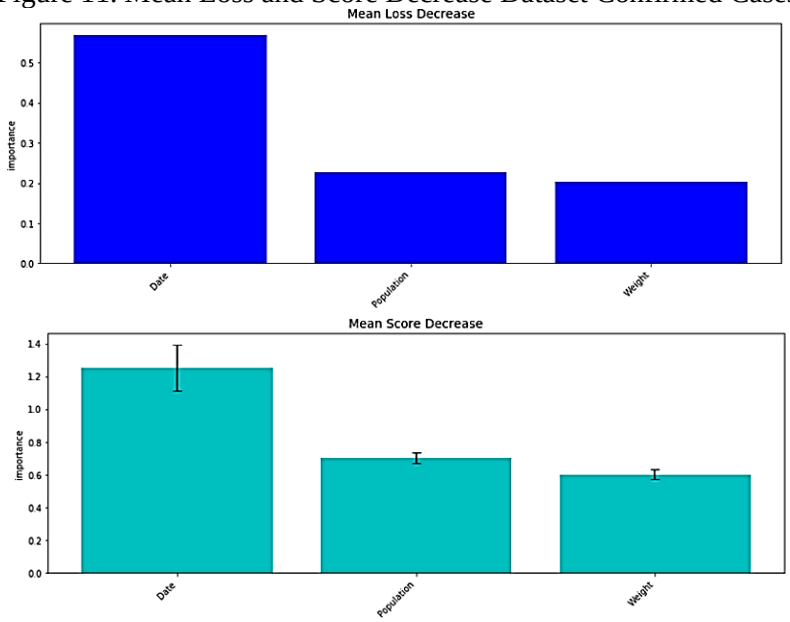

Figure 12. Mean Loss and Score Decrease Dataset Fatalities

Based on the results of the feature importance calculation for the confirmed cases and fatalities dataset, it is known that the feature date has the greatest influence on the accuracy of the model, this is inversely proportional to the correlation results where the feature date has a low correlation with the target value. thus it can be concluded that features with high correlation values not necessarily have a major influence on the accuracy of the model and vice versa.

4. CONCLUSION

Based on the R^2 and RMSE results obtained from all the methods used, the Decision Tree method has the best results with an R^2 value of 0.97 but has an RMSE value that is no better than the Random Forest model (0.88), namely 52.57 for the DT and 21.51 for the RF model. In theory, a good model has a high R^2 score and a low RMSE, but in the results our group obtained, the DT model has a higher R^2 than the RF model. However, in contrast to the RMSE value. Then, the XGboost model with R^2 is around 0.88. ANN has a relatively lower R^2 value with an R^2 value of around 0.75. The R^2 value obtained from the LSTM method is the worst with a value of 0.28.

REFERENCES

[1] Balli S. (2021). Data Analysis of Covid-19 Pandemic and Short-term Cumulative Case Forecasting using Machine Learning Time Series Method. Chaos, Solitons, and Fractals 142:110512.

[2] Fanelli, D. and Piazza, F. (2020). Analysis and Forecast of Covid-19 Spreading in China, Italy, and France. Chaos, Solitons, and Fractals 134:109761.

[3] Yeshilkanat, CM. (2020). Spatio-temporal Estimation of the Daily Cases of Covid-19 in Worldwide using Random Forest Machine Learning Algorithm. Chaos, Solitons, and Fractals 140:110210.

[4] Wang, J., Zhang, J., & Zhang, H. (2020). Logistic Models and Machine Learning for Covid-19 Prediction. International Journal of Environmental Research and Public Health, 17(8), 2905.

[5] Ahmed, M., & Younis, M. (2020). Time Series Forecasting for Covid-19 Cases Using Deep Learning Algorithms. Journal of Computational Biology, 27(11), 1556-1568.

[6] Sharma, A., & Rani, M. (2020). Deep Learning for Forecasting Covid-19 Spread. Journal of Machine Learning Research, 21(124), 1-15.

[7] Zhang, Y., & Wang, M. (2020). Covid-19 Forecasting Using XGBoost and Machine Learning Models. Advances in Science, 15(3), 202-208.

[8] Raj, A., & Gupta, A. (2020). Machine Learning Approaches for Epidemic Prediction. Computational Biology and Chemistry, 86, 107282.

[9] Chen, M., & Zhao, Y. (2020). Predicting Epidemic Outbreaks Using Random Forest. International Journal of Environmental Research and Public Health, 17(1), 101.

[10] Verma, A., & Gupta, D. (2020). Polynomial Regression for Modeling Covid-19 Growth. Mathematical Methods in the Applied Sciences, 43(6), 3211-3220.

[11] Brown, G., & Maugis, S. (2020). Analyzing Covid-19 Trends with Machine Learning. The Lancet, 395(10223), 877-884.

[12] Singh, D., & Kumar, S. (2020). AI in Public Health: Modeling and Predictions for Covid-19. Artificial Intelligence in Medicine, 105, 101864.

[13] Wang, L., & Deng, X. (2020). AI-Assisted Epidemiological Forecasting for Covid-19. Journal of Epidemiology and Community Health, 74(10), 819-825.

[14] Zhao, Z., & Zhang, T. (2020). Impact of Social Distancing on the Covid-19 Spread: A Machine Learning Study. Scientific Reports, 10, 21348.

[15] Kumar, A., & Patil, A. (2020). Covid-19 Spread Forecasting Using AI Algorithms: A Comparative Study. Journal of Computational and Graphical Statistics, 29(3), 651-667.