PREDICTION OF THE NUMBER OF FOREIGN TOURISTS POST-COVID-19 IN INDONESIA WITH LINEAR REGRESSION AND RANDOM FOREST REGRESSION ALGORITHM

1st Dave Cendra Wirawan
00000036759
Sistem Informasi
Universitas Multimedia Nusantara
Tangerang, Indonesia
dave.cendra@student.umn.ac.id

2nd Gladys Patricia
00000042400
Sistem Informasi
Universitas Multimedia Nusantara
Tangerang, Indonesia
gladys.patricia@student.umn.ac.id

3rd Muhammad Rafii Haditomo E
00000039904
Sistem Informasi
Universitas Multimedia Nusantara
Tangerang, Indonesia
muhammad.rafii@student.umn.ac.id

4th T.Angel Caroyallita
00000034662
Sistem Informasi
Universitas Multimedia Nusantara
Tangerang, Indonesia
angel58@student.umn.ac.id

***Abstract*—Tourism industry is one of the fields that brings a lot of income to Indonesia. However, this sector cannot run optimally due to the COVID-19 pandemic. Tourist destinations is limited to minimize COVID-19 case. This research is conducted to determine the effect of COVID-19 cases and the number of vaccinations on the number of foreign tourists who came to Indonesia. In addition, it is intended to determine the number of foreign tourists visiting tourist destinations in Indonesia in the post-pandemic era through predictions. The methods used are Linear Regression and Random Forest Regression models. The prediction model made using Linear Regression has RMSE value of 671,556. The RMSE value of the Linear Regression model is better than the model with Random Forest with RMSE value of 2257.8687. Therefore, it can say that the linear regression algorithm is more suitable for predicting the number of tourists who come to Indonesia after the pandemic.**

***Index Terms*—COVID-19, Linear Regression, Random Forest Regression, tourist, Vaccination.**

I. Introduction

Indonesia is a country with abundant natural wealth and beauty. Indonesia has a variety of tourist destinations that are ready to welcome visitors with their beauty. As an archipelagic country, each island in Indonesia has its own characteristics and charm. The beauty offered by nature has succeeded in attracting the attention of foreign tourists to visit. The tourism industry is also one of the fields that provide a large income for the country. But on the other hand, gathering and vacationing in tourist destinations is not supportive. This is caused by the COVID-19 pandemic, COVID-19 (Coronavirus Disease) is a disease caused by the SARS-CoV-2 virus. Based on the National COVID-19 Handling Task Force, this virus works by infecting the lower respiratory tract, and developing into acute respiratory disorders that can cause organ failure and death [1].

COVID-19 has attacked dozens of countries and killed tens of millions of people, including in Indonesia. Due to the COVID-19 attack, the World Health Organization (WHO) officially declared a global world health emergency on January 31, 2020 [2]. This virus was first detected in Wuhan on December 31, 2019. The transmission and spread of this virus is through droplets that come out when someone coughs, sneezes, or talks. This virus is growing every day around the world. One of the efforts made by the government to prevent the spread of COVID-19 in Indonesia is to apply new normal rules. This new normal rule results in restrictions on activities and the number of people involved in an activity, as well as implementing health protocols [3]. The government has set a policy to close foreign tourist receipts when the COVID-19 case curve begins to increase. The reopening of tourist destinations will be carried out again if the COVID-19 case starts to sag. This policy is carried out to control the spread of this disease so as not to endanger the lives of the Indonesian people. However, in reality, the new variant of COVID-19 continues to grow and spread in Indonesian territory. The tourism sector is planned to run normally after COVID-19 cases are reduced and the health conditions of countries around the world are improving. Through this research, the number of foreign tourists will be predicted in the post-COVID-19 era.

There are several previous studies that were used as a reference in the implementation of this research. These are:

Previous research was conducted by Muhammad Ridwan in 2017 to predict the number of foreign tourists arriving

through international airports in Indonesia using the Linear Regression algorithm. The level of accuracy of research conducted using the Linear Regression algorithm is quite high, which is 78% with an error value of 22% [4].

In 2021, the Linear Regression algorithm is used by Wisnu Hatta Nugroho to predict the addition of new cases of COVID-19 in Jakarta. Based on the accuracy test using the Coefficient of Multiple Determination (R Square), the accuracy rate is 94% [5].

In 2022, a study was conducted by Syakirah Fachid and Agung Triayudi to compare the Linear Regression and Random Forest Regression algorithms in predicting positive cases of COVID-19. The study resulted in a lower level of accuracy for linear regression compared to random forest regression, which is 94% vs 97.7%. So it can be said that the model formed by Random Forest Regression is better suited to make predictions from the research conducted compared to Linear Regression [6].

Based on previous research references, this study will analyze the number of COVID-19 cases in Indonesia and the COVID-19 vaccine data with the number of foreign tourists coming to Indonesia. Researchers will use the Linear Regression and Random Forest Regression models to predict the number of foreign tourists who come to Indonesia after COVID-19. The use of these two algorithms is to find out which of the two algorithms is the most suitable for predicting the number of foreign tourists who come to Indonesia after COVID-19. The comparison is based on the comparison RMSE values of the two algorithms.

II. Literature Review

A. Theoretical Basis

This study uses several literature studies and theories to support the process of predicting data on foreign tourists to Indonesia. The theory used is Linear Regression, Random Forest Regression, and Pearson Correlation.

1) Linear Regression: Linear Regression is an analytical technique used to see the correlation or relationship between one variable and another. Linear regression itself was first introduced by Sir Francis Galton in 1894 [7]. Linear regression consists of at least 2 variables, namely the dependent variable and the independent variable. In this study, the linear regression implemented was Simple Linear Regression. This linear regression has only one determinant and one independent variable. This method is used to predict future data based on past data. Prediction of the value of the dependent variable can be done if the dependent variable is known [8]. The relationship between the independent variable and the dependent variable in the form of a linear regression equation can take several forms, including linear, exponential, and multiple relationships.

   There are two types of linear regression, namely simple linear regression and multiple linear regression. In simple terms, it can be concluded that simple linear regression only uses one independent variable, while multiple linear regression involves more than one independent variable [9]. For simple linear regression can be calculated using the following formula:

$$Y = a + bX \tag{1}$$

Description :
Y = dependent variable
a = constant
b = variable coefficient x
X = independent variable

Meanwhile, to calculate multiple linear regression can use the following formula:

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \tag{2}$$

Description:
Y = dependent variable
a = constant
b1, b2, .., bn = regression coefficient value
X1 ,X2 ,., Xn = independent variable

The workings of linear regression are as follows:

a) Determine the purpose of analyzing simple linear regression.
b) Perform data collection.
c) Identify predictor variables and response variables.
d) Calculates X^2, Y^2, XY, and the total of each variable.
e) Calculate a and b based on the formula.
f) Create a linear regression equation model .
g) Make predictions on predictor variables or response variables.
h) Perform a significance test using t-test.

In its application, linear regression has advantages and disadvantages. Reporting from the site edureka [10], the advantages of linear regression are:

a) Linear regression can work very well for data that can be separated linearly.
b) Easy to implement and efficient to train.
c) Can handle overfitting problems well using dimension reduction, regulation, and cross-validation techniques.

While the disadvantages of using linear regression are as follows:

a) Limited to linear relationships.
b) Fairly sensitive to outliers.
c) There must be an independent variable.

2) Random Forest Regression: Random Forest Regression is one part of the algorithm that belongs to supervised learning which in this algorithm utilizes ensemble learning. Ensemble learning is a process that utilizes the use of several models trained with the same data train to

then produce a prediction and classification. The purpose of using ensemble learning is to find a model that has better prediction and classification accuracy than the other models formed [6].

The advantage of the ensemble learning method in Random Forest Regression is that it is able to produce many models from the same data but each has a different level of accuracy, so that a single model can be determined which is the most suitable for implementation. The implementation process of Random Forest Regression uses bootstrapping to perform random sampling. So that the model formed uses a number of variable samples that are processed with a certain number of iterations. It can be said that Random Forest Regression is a combination of ensemble learning and a decision tree framework that produces a number of decision trees with different results or outputs.

3) Pearson Correlation: The Pearson Correlation test is a simple correlation measurement technique, where the measurements carried out involve the dependent and independent variables [11]. The purpose of doing the Pearson Correlation test is to measure the level of correlation or strength of the relationship between two variables linearly [12]. The correlation value between the tested variables can be known by the following equation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Description:
x = first variable
y = second variable
n = number of tests performed

The degree of correlation of the two variables is indicated by the correlation coefficient. The correlation coefficient is a unit of measurement used to determine the relationship between two variables. The correlation coefficient is between -1 ¡ 0 ¡ 1, where -1 indicates a strong correlation that is inversely proportional. While 1 shows a perfect positive correlation with a very strong relationship between two variables in the same direction. How the Pearson correlation test works, namely:

a) Determine the initial hypothesis of the test. The hypothesis can be whether there is a significant correlation between the variables being tested
b) Determine the level of significance to perform correlation testing
c) Doing statistical tests

In its use, the Pearson correlation test has advantages and disadvantages. Reporting from the Statistical Solutions site [13], the advantages of the Pearson correlation are:

a) Pearson correlation test is very good to be used in determining the linear relationship of two variables
b) The Pearson correlation test is able to provide detailed information from the results of the tests carried out, such as the magnitude of the correlation level to the direction of the relationship between the two variables

Meanwhile, some of the shortcomings of the Pearson correlation, among others [14]:

a) The Pearson correlation test is not suitable for attributive research hypothesis testing that includes only one variable.
b) Correlation does not always provide information about the cause and effect of the data being tested.
c) The results of a large correlation coefficient do not always indicate a high linear relationship between two variables.

4) Root Mean Square Error: Root Mean Square Error (RMSE) is a measurement of the error value of the prediction results made. RMSE which is the square root of Mean Square Error (MSE) helps in knowing and measuring the error rate of the model built [15]. In this case, every model formation carried out has a target to be able to produce a model with a minimum RMSE value and even close to a value of 0. The closer to the value of 0, then the model made can be said to be good and minimal errors.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \quad (3)$$

Description:
n = number of records
i = index from record start from first record
y = real record based on testing data
$\hat{y}$ = prediction score from many of y value of the model

III. Research Methods

A. Research methods

The research method used is literature study sourced from books and journals that have a relationship with the research topic. The literature study method is a series of activities related to the process of collecting library lists, taking notes, and managing research materials.

B. Research Variable

There are two research variables used in this study, namely:

1) Dependent Variable
The dependent variable is a variable that changes based on independent variable. The dependent variable in this study is named as 'Total Wisatawan' or the number of tourist.
2) Independent Variable
The independent variable is the variable that causes changes to occur in the dependent variable. The independent variables in this study are 'month', 'Negatif', 'Total kasus', 'Total sembuh', 'total_vaccinations', 'people_vaccinated', 'people_fully_vaccinated', 'Total Wisatawan'.

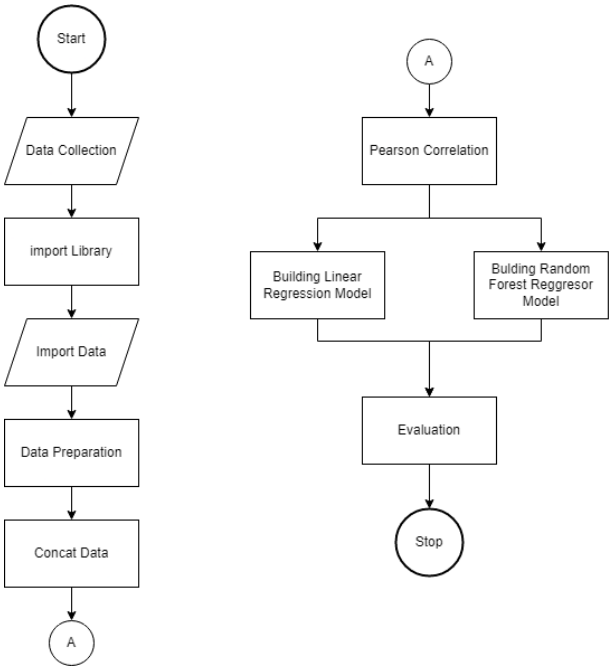


Fig. 1. Research's Flow

C. Data Collection Technique

In this study, datasets were collected by scraping data from several different sites. There are 3 datasets used in this study. The dataset is the "COVID-19 Case" data obtained from the website https://kawalcovid19.id/. In addition, there is a "Vaccination" dataset obtained from the website https://github.com/. The last dataset used is the "Monthly Foreign Tourist Visits" dataset which was obtained from the website https://www.bps.go.id/.

D. Data Analysis Technique

1) Data Collection
   There are three datasets used in this research, all of them are obtained from websites such as Badan Pusat Statistik, Kawal COVID-19, and GitHub. Each dataset has a different file format, number of columns, and rows
2) Import Library
   Import Library is a list of libraries that will be used to support the creation of research models. The libraries used in this research are pandas, numpy, seaborn, and matplotlib. The pandas library is used for data manipulation such as reading datasets, removing null values, and so on. The numpy library is used for numerical computation. numpy has the ability to create N-dimensional array objects. The seaborn library is used to create graphs and statistics. The matplotlib library is used to perform data visualizations such as plotting a graph for one or more axes.
3) Import Data
   Import Data is the process of retrieving datasets that have been collected. In this study, the data imported were cases of COVID-19, vaccinations, and tourists.
4) Data Preparation
   After importing data, the next step that needs to be done is data preparation. Data preparation is the process of cleaning and changing raw data before it is processed and analyzed. This process is very important so that the data used in the research get the best modeling results
5) Concat Data
   Concat is one of the techniques used to merge datasets. The process of merging datasets with concat is nailed to the same variable. At this stage, the vaccination and covid datasets are combined based on the date and month. The results of the merging are then combined with the tourist dataset, resulting in 12 variables in the dataset that are used in the next stage.
6) Pearson Correlation
   Pearson correlation is a step to test the correlation of several variables that have passed the data preparation stage in the feature selection section. This correlation shows the relationship or relationship between variables, whether the correlation is quite high or low, along with the direction of the correlation.
7) Exploratory Data Analysis
   Exploratory Data Analysis (EDA) is the process of testing data to identify a pattern, find anomalies, test a hypothesis, and find out the statistical value of each relevant variable. The EDA process makes it possible to get some in-depth findings about the variables in the data set and their relationships. At this stage, describe the data to see a statistical description of each variable from the data that has been obtained, then detect outliers using a box plot, and visualize the data to see the findings from the dataset.
8) Building Model
   Two types of models were made, namely Linear Regression and Random Forest Regression. The purpose of making these two types of models is to compare which model can make better predictive data. Because the two models have different ways of making predictions.
9) Evaluation
   Evaluation is the final stage of the machine learning process. Evaluation is used to evaluate the performance of the model used. One form of a good evaluation can be seen from the accuracy based on the RMSE value of each model that is formed. If the RMSE value is close to 0, then the model applied can be said to be good.

IV. Result and Discussion

A. Pearson Correlation

Based on the correlation matrix above, it can be concluded that there are various types of relationships or correlations between variables. The thing that needs to be considered in this correlation is actually the Total Tourist variable with other variables. It can be seen that the correlation results are as follows:

Total travelers with:

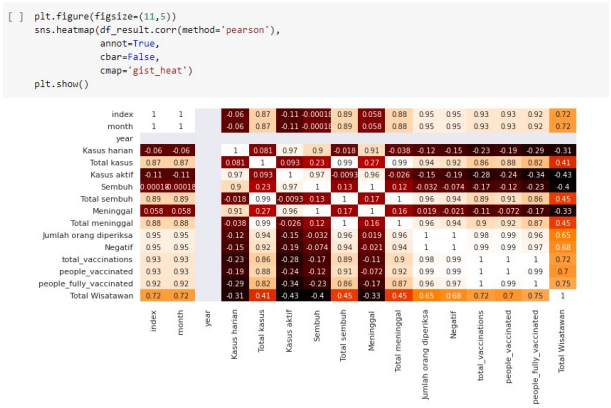


Fig. 2. Correlation between variable

- Month of 0.72 with a strong correlation
- Daily case of -0.31 with weak inverse correlation
- The total case is 0.41 with a fairly strong correlation
- Active cases are -0.43 with a fairly strong correlation inversely
- Healed by -0.4 with a fairly strong correlation inversely
- Total recovery is 0.45 with a fairly strong correlation
- Died by -0.33 with low correlation and inversely
- The total death is 0.45 with a fairly strong correlation
- The number of people examined was 0.65 with a strong correlation
- Negative 0.68 with a strong correlation
- Total vaccination is 0.72 with strong correlation

Based on this correlation, it can be concluded that there is a fairly strong relationship between COVID-19 cases and COVID-19 vaccinations on the number of foreign and local tourists to Indonesia. The highest correlation was obtained by vaccination, where the higher the number of vaccinations, the higher the number of tourists in Indonesia. Meanwhile, active cases have a fairly strong and reverse effect, where the higher the active cases, the fewer visitors. This fairly strong correlation actually shows the fact that in the midst of COVID-19, people are still visiting and traveling.

B. Describe Data

At this stage, the researcher explores the dataset that is neat and ready to use. The first step is to look at the statistical description of all the required variables using the describe() function. From this function, it is found that each variable used has different statistical values. The statistical values that can be seen are count, mean, std, min, 25%, 50%, 75%, and max as shown on Fig. 3

C. Box Plot

The next step to explore the data is to make a box plot with the aim of finding out if there are outliers in the data. After visualization, it can be seen that there are no outliers in the data used. In addition, you can see the maximum value, middle value, and minimum value of the total tourists. The

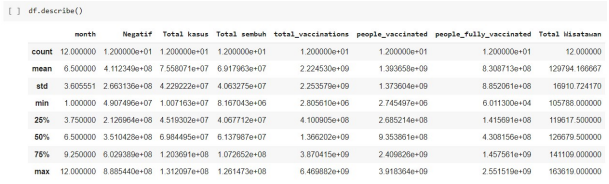


Fig. 3. Describe Data

maximum value for total tourists is 163619. The median value for total tourists is 126679.5, and the minimum value for total tourists is 105788.

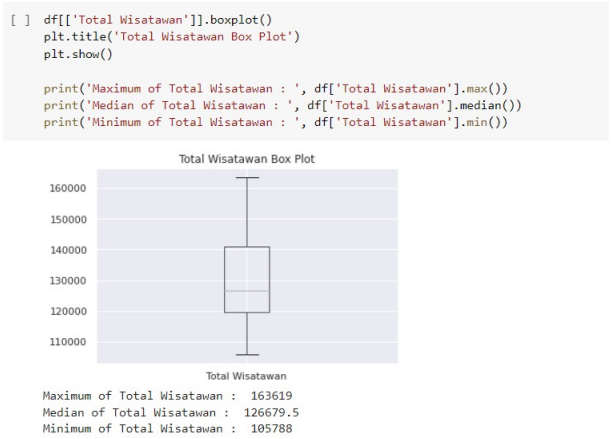


Fig. 4. Box Plot

D. Visualize Data

The next step is to visualize the data to see interesting findings about the data used. The first visualization is a bar chart to see the 5 highest number of tourists by month as shown on Fig. 5. After visualization, it was found that the most total tourists were obtained in December 12 or December with a total of more than 160000 tourists. Then the second position was obtained in 11th or November with a total of 150000 tourists. In the third position was the month of 10 or October with a total of tourists of approximately 145000. The fourth position is on the 5th month or May with a total of 140000 tourists. Then the 5th position is the 7th month or July with a total of 130000 tourists.

The second visualization is a scatter plot. Scatter plots can show the relationship between variables. At this stage, the variables used to be visualized are total case data and total recovered data as shown on Fig. 6. After visualization, it can be concluded that the total cases have a fairly strong relationship with the total recovered, which indicates that if the total number of cases increases, the total cure will also increase. In addition, the color of the plot is based on the total cure variable where the closer the color is to yellow, the higher the total number of cures.

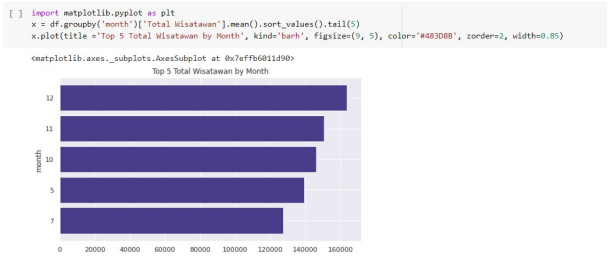


Fig. 5. *Bar Chart*

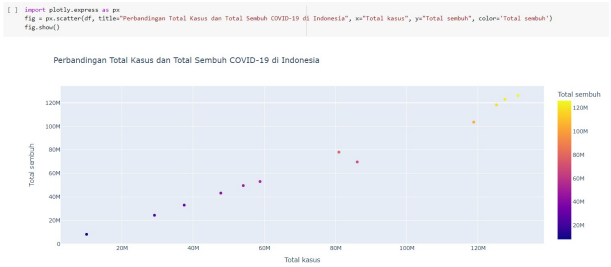


Fig. 6. Scatter Plot

The next visualization is the pairplot. The result of this visualization is the correlation and distribution of data between one variable and another as shown on Fig. 7. After visualization, the following results are obtained:

Fig. 7. Pair Plot

E. Linear Regression

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```

```
x = df.drop(columns='Total Wisatawan')
y = df['Total Wisatawan']
```

Fig. 8. Split feature and target

From the implementation of the linear regression model, the results of the coefficient of determination are obtained. Coefficient of determination is a value that states how close the regression function we use is to the data used as a reference. The result of the coefficient of determination obtained from

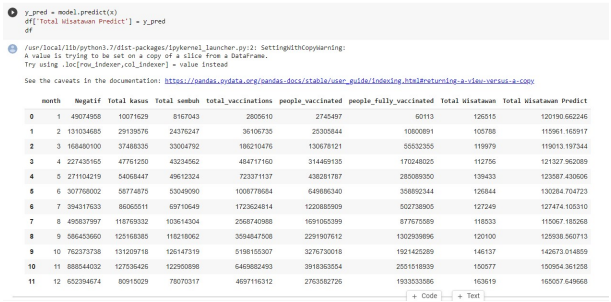
```
model = LinearRegression().fit(x, y)
r_sq = model.score(x, y)
print('coefficient of determination:', r_sq)
```

```
coefficient of determination: 0.8279603017180894
```

Fig. 9. Coefficient of determination of the model

this dataset is 0.827. From the regression model that is formed, it can be concluded several value results, namely:

- R-squared: 0.828
- Adj. R-squared: 0.622
- F-statistic: 4.011
- Prob: 0.0745
- Log-likelihood: -122.77
- AIC: 259.5
- BIC: 262.9



	month	Negatif	Total kasus	Total sembuh	total_vaccinations	people_vaccinated	people_fully_vaccinated	Total Wisatawan	Total Wisatawan Predict
0	1	49074958	10071629	8167043	2805610	2745497	60113	126515	120190.662246
1	2	131034685	29139576	24376247	36106735	25305844	10800891	105788	115961.165917
2	3	168480100	37488335	33004792	186210476	130678121	55532355	119979	119013.197344
3	4	227435165	47761250	43234562	484717160	314469135	170248025	112756	121327.962089
4	5	271104219	54068447	49612324	723371137	438281787	285089350	139433	123587.430606
5	6	307760002	58774875	53049090	1006779684	649886340	356892344	126844	130284.704723
6	7	394317633	86065511	69710649	1723624814	1220885909	502738905	127249	127474.105310
7	8	495837997	118769332	103614304	2568740988	1691065399	877675589	118533	115067.185268
8	9	586453660	125168385	118218062	3594847508	2291907612	1302939896	120100	125938.560713
9	10	762373738	131209718	126147319	5198155307	3276730018	1921425289	146137	142673.014859
10	11	888544032	127536426	122908898	6469882493	3918363554	2551518939	150577	150954.361258
11	12	652394674	80915029	78070317	4697116312	2763582726	1933533586	163619	165057.649668

Fig. 10. Assign prediction result to dataframe

The prediction model is formed from the model built using model.predict(x), where x is a feature. After forming the y_pred prediction model, then enter the y_pred variable into the dataframe. The chart below is the comparison between the

```
def plotGraph(y_test,y_pred,regressorName):
    if max(y_test) >= max(y_pred):
        my_range = int(max(y_test))
    else:
        my_range = int(max(y_pred))
    plt.scatter(range(len(y_test)), y_test, color='blue')
    plt.scatter(range(len(y_pred)), y_pred, color='red')
    plt.title(regressorName)
    plt.show()
    return
```

```
plotGraph(df['Total Wisatawan'],df['Total Wisatawan Predict'], "Comparison")
```

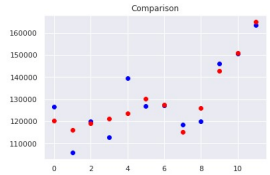


Fig. 11. Comparison of actual vs predictive data

actual data (red) and the predicted data (blue).

F. Random Forest Regression

The making of this model used n_estimators of 500 and random_state of 0. n_estimators is the number of trees that you

```
[ ]  from sklearn.ensemble import RandomForestRegressor

     regressor = RandomForestRegressor(n_estimators=500, random_state=0)
     regressor.fit(x_train, y_train)
     y_pred = regressor.predict(x_test)
```

Fig. 12. Random Forest Regression code

want to build before taking the maximum vote or prediction average.

```
[ ]  from sklearn import metrics
     print('Mean Absolute Error: {}' .format(round(metrics.mean_absolute_error(y_test, y_pred)*0.1,4)))
     print('Mean Squared Error: {}' .format(round(metrics.mean_squared_error(y_test, y_pred)*0.01,4)))
     print('Root Mean Squared Error: {}' .format(round(np.sqrt(metrics.mean_squared_error(y_test, y_pred))*0.1,4)))

     Mean Absolute Error: 1929.1085
     Mean Squared Error: 5097965.9907
     Root Mean Squared Error: 2257.8676
```

Fig. 13. Model evaluation

After making the model, the next step is to evaluate the model made. At the evaluation stage, several assessment metrics from the model formed will be seen, such as MAE, MSE, and RMSE. It can be seen that the RMSE value in the random forest regression model is 2257.8676.

V. Conclusion and Suggestion

Based on the research conducted, several conclusions were obtained, namely:

1) Judging from the results of using Pearson correlation to measure the correlation between variables in the dataset, it was found that the total Covid-19 cases had a strong enough effect on the total number of foreign tourists who came to Indonesia with a correlation value of 0.41.
2) Total vaccination has a strong effect on the total number of foreign tourists who come to Indonesia with a correlation value of 0.75.
3) The linear regression model formed to predict the number of tourists who came to Indonesia after the pandemic showed good results with an R-squared value of 0.828 (83%) and an RMSE value of 671,556.
4) The random forest regression model that was formed to predict the number of tourists who came to Indonesia after the pandemic showed quite good results with the RMSE value formed was 2257.8676.
5) From the results of the modeling formed between linear regression and random forest regression, it is concluded that the linear regression algorithm is more suitable to be used in predicting the number of tourists who come to Indonesia after the pandemic with a significant difference in the RMSE value and is supported by an R-squared value of 83%.

Suggestion:

1) Retrieve data from more diverse sources with complete data. It is recommended to take the dataset in daily time and up to date until the research is completed.
2) Try using another algorithm or type of prediction, not just regression. Clustering can also be done based on existing data.

References

[1] S. COVID-19, "Pengendalian Covid-19," 2021.
[2] P. D. P. Indonesia, "Jurnal Respirologi Indonesia," *Jurnal Respirologi Indonesia*, vol. 20, no. 2, pp. 120–129, 2020.
[3] L. L. Sunnah, I. & Indrayati, "Edukasi New Normal Sebagai Upaya Pencegahan Penyebaran Covid-19 melalui G 5M dan CTPS," *Indonesian Journal of Community Empowerment (IJCE)*, vol. 3, no. 1, pp. 56–60, 2021.
[4] M. Ridwan and H. Himawan, "Implementasi algoritma linear regresi untuk prediksi jumlah wisatawan mancanegara melalui bandara internasional indonesia."
[5] W. H. Nugroho, "Sistem prediksi jumlah kasus covid-19 di jakarta menggunakan metode linear regression," Ph.D. dissertation, Universitas Pembangunan Nasional Veteran Jakarta, 2021.
[6] Fachid and Triayudi, "Perbandingan Algoritma Regresi Linier dan Regresi Random Forest Dalam Memprediksi Kasus Positif Covid-19," *Jurnal Media Informatika BudiDarma*, vol. 6, no. 1, pp. 68–73, 2022.
[7] S. Kumari, K. & Yadav, "Linear regression analysis study," *J. Pract. Cardiovasc. Sci*, vol. 4, no. 1, p. 33, 2018.
[8] G. N. Ayuni and D. Fitrianah, "Penerapan Metode Regresi Linear Untuk Prediksi Penjualan Properti pada PT XYZ," *Jurnal Telematika*, vol. 14, pp. 79–86, 2018.
[9] S. Disa, "Penerapan Metode Regresi Linear dalam Pembuatan Perangkat Lunak Simulasi Target Penjualan," *Sinta*, vol. 5, pp. 82–89, 2015.
[10] Edureka, "Linear regression for machine learning: Intro to ML algorithms," 2022. [Online]. Available: https://www.edureka.co/blog/linear-regression-for-machine-learning/advantages
[11] W. R. Safitri, "Analisis Korelasi Pearson dalam Menentukan Hubungan Antara Kejadian Demam Berdarah Dengue dengan Kepadatan Penduudk di Kota Surabaya Pada Tahun 2012-2014: Pearson Correlation Analysis to Determine The Relationship Between City Population Density with Incident Dengue Fever of Surabaya in The Year 2012-2014," *jikep*, vol. 2, no. 2, pp. 21–29, 2022.
[12] I. S. Miftahuddin Miftahuddin, Ananda Pratama Sitanggang, "Analisis Hubungan Antara Kelembaban Relatif Dengan Beberapa Variable Iklim Dengan Pendekatan Korelasi Pearson Di Samudera Hindia," *Jurnal Siger Matematika*, 2021.
[13] S. Solution, "Pearson's Correlation Coefficient - Statistics Solutions," 2021.
[14] B. Davis, "What are the advantages and disadvantages of correlation?" 2019. [Online]. Available: https://www.mvorganizing.org/what-are-the-advantages-and-disadvantages-of-correlation/
[15] W. Hastomo, A. S. B. Karno, N. Kalbuana, E. Nisfiani, and E. T. P. Lussiana, "Optimasi Deep Learning untuk Prediksi Saham di Masa Pandemi Covid-19," *Jurnal Edukasi dan Penelitian Informatika*, vol. 7, no. 2, pp. 133–140, 2021.