CHAPTER 4

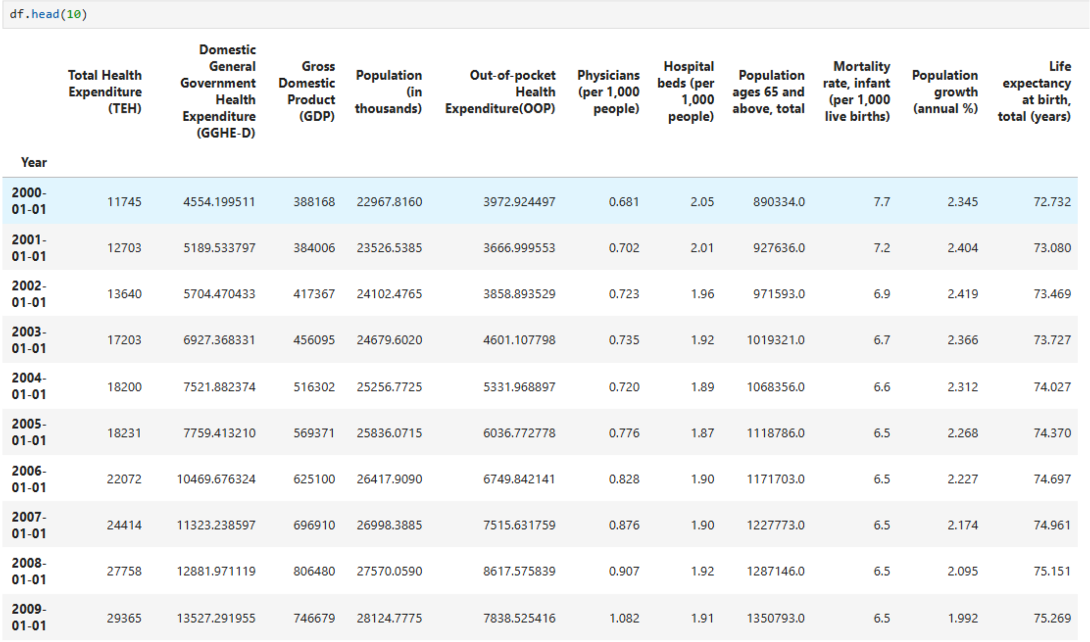
INITIAL FINDINGS

4.1 Introduction

This chapter will discuss the initial findings for the research project. The findings included results from data pre-processing, exploratory data analysis, feature engineering, and initial modelling. In data pre-processing section, the results of data cleaning and data integration are discussed. Exploratory data analysis covers the analysis of the total health expenditure with its components and the relationship between features. Initial findings from machine learning models are presented, evaluated and discussed before ending with a summary section.

4.2 Data Pre-processing Results

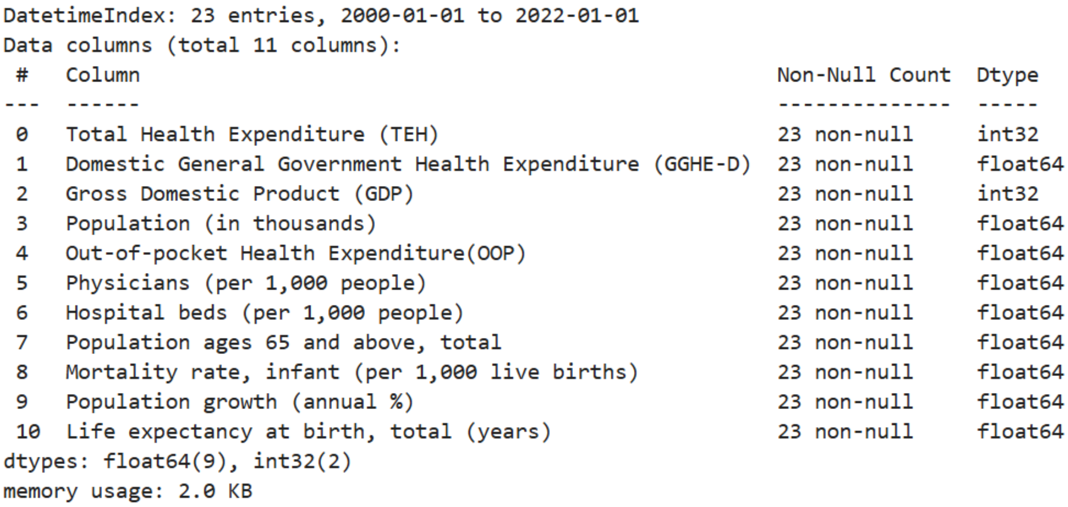
Four datasets obtained in the data collection are cleaned, transformed and integrated into one dataset as shown in Figure 4.1. Its details are shown in Figure 4.2. The final dataset contains 11 columns and 23 rows of data, which represent data values from the year 2000 to 2022. The year is set as the index in DateTime format. There are 2 columns with integer as their datatype; the rest are in float, containing decimal places.



```
df.head(10)
```

	Total Health Expenditure (TEH)	Domestic General Government Health Expenditure (GGHE-D)	Gross Domestic Product (GDP)	Population (in thousands)	Out-of-pocket Health Expenditure(OOP)	Physicians (per 1,000 people)	Hospital beds (per 1,000 people)	Population ages 65 and above, total	Mortality rate, infant (per 1,000 live births)	Population growth (annual %)	Life expectancy at birth, total (years)
Year											
2000-01-01	11745	4554.199511	388168	22967.8160	3972.924497	0.681	2.05	890334.0	7.7	2.345	72.732
2001-01-01	12703	5189.533797	384006	23526.5385	3666.999553	0.702	2.01	927636.0	7.2	2.404	73.080
2002-01-01	13640	5704.470433	417367	24102.4765	3858.893529	0.723	1.96	971593.0	6.9	2.419	73.469
2003-01-01	17203	6927.368331	456095	24679.6020	4601.107798	0.735	1.92	1019321.0	6.7	2.366	73.727
2004-01-01	18200	7521.882374	516302	25256.7725	5331.968897	0.720	1.89	1068356.0	6.6	2.312	74.027
2005-01-01	18231	7759.413210	569371	25836.0715	6036.772778	0.776	1.87	1118786.0	6.5	2.268	74.370
2006-01-01	22072	10469.676324	625100	26417.9090	6749.842141	0.828	1.90	1171703.0	6.5	2.227	74.697
2007-01-01	24414	11323.238597	696910	26998.3885	7515.631759	0.876	1.90	1227773.0	6.5	2.174	74.961
2008-01-01	27758	12881.971119	806480	27570.0590	8617.575839	0.907	1.92	1287146.0	6.5	2.095	75.151
2009-01-01	29365	13527.291955	746679	28124.7775	7838.525416	1.082	1.91	1350793.0	6.5	1.992	75.269

Figure 4.1 Pre-processed Dataset



```
DatetimeIndex: 23 entries, 2000-01-01 to 2022-01-01
Data columns (total 11 columns):
 #   Column                                                    Non-Null Count  Dtype  
---  ------                                                    --------------  -----  
 0   Total Health Expenditure (TEH)                            23 non-null     int32  
 1   Domestic General Government Health Expenditure (GGHE-D)   23 non-null     float64
 2   Gross Domestic Product (GDP)                              23 non-null     int32  
 3   Population (in thousands)                                 23 non-null     float64
 4   Out-of-pocket Health Expenditure(OOP)                     23 non-null     float64
 5   Physicians (per 1,000 people)                             23 non-null     float64
 6   Hospital beds (per 1,000 people)                          23 non-null     float64
 7   Population ages 65 and above, total                       23 non-null     float64
 8   Mortality rate, infant (per 1,000 live births)            23 non-null     float64
 9   Population growth (annual %)                              23 non-null     float64
 10  Life expectancy at birth, total (years)                   23 non-null     float64
dtypes: float64(9), int32(2)
memory usage: 2.0 KB
```

Figure 4.2 Details of the Cleaned Dataset

4.3 Exploratory Data Analysis Results

4.3.1 Health Expenditure

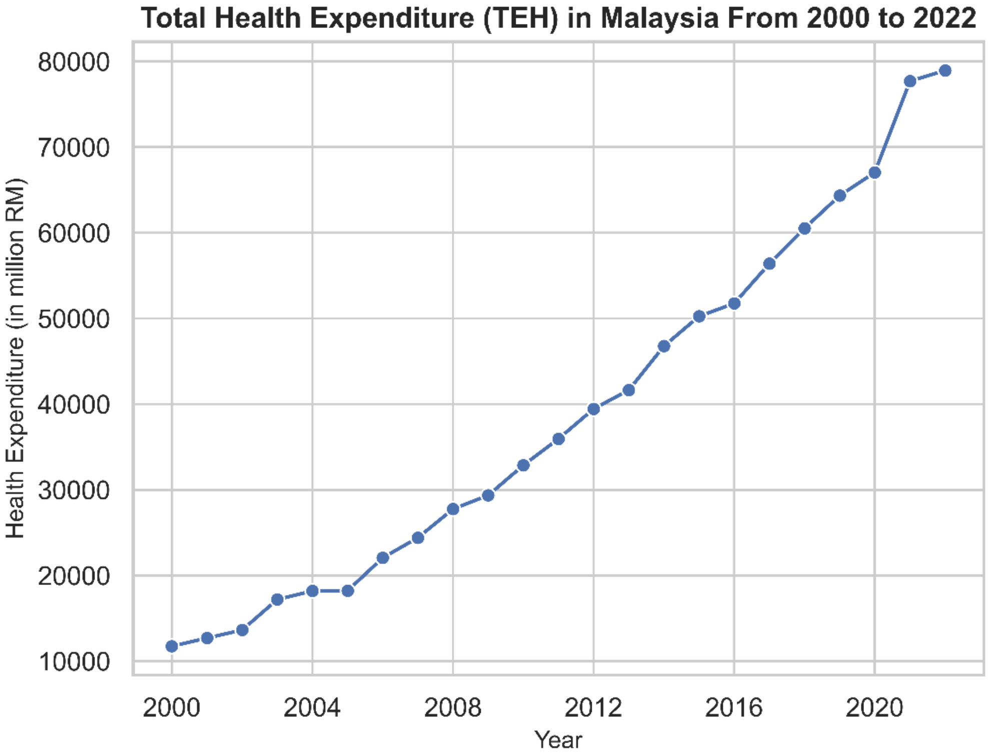


Figure 4.3 Total Health Expenditure in Malaysia over Year

The total health expenditure in Malaysia is plotted over year from 2000 to 2022. The line chart shows that there is a gradual increase in health expenditure over 23 years, except that there is a steep increase from 2020 to 2021 (RM 67 million to RM 77 million). This is a result of increased health expenditure during COVID-19 pandemic, which includes testing, treatment, contact tracing, vaccination, medical equipment and other COVID-19-related spending (MOH, 2024). Since there is a strong positive trend observed from the line chart, the time series is not stationary. Therefore, differencing has to be applied to stabilise the mean when carrying out ARIMA modelling.

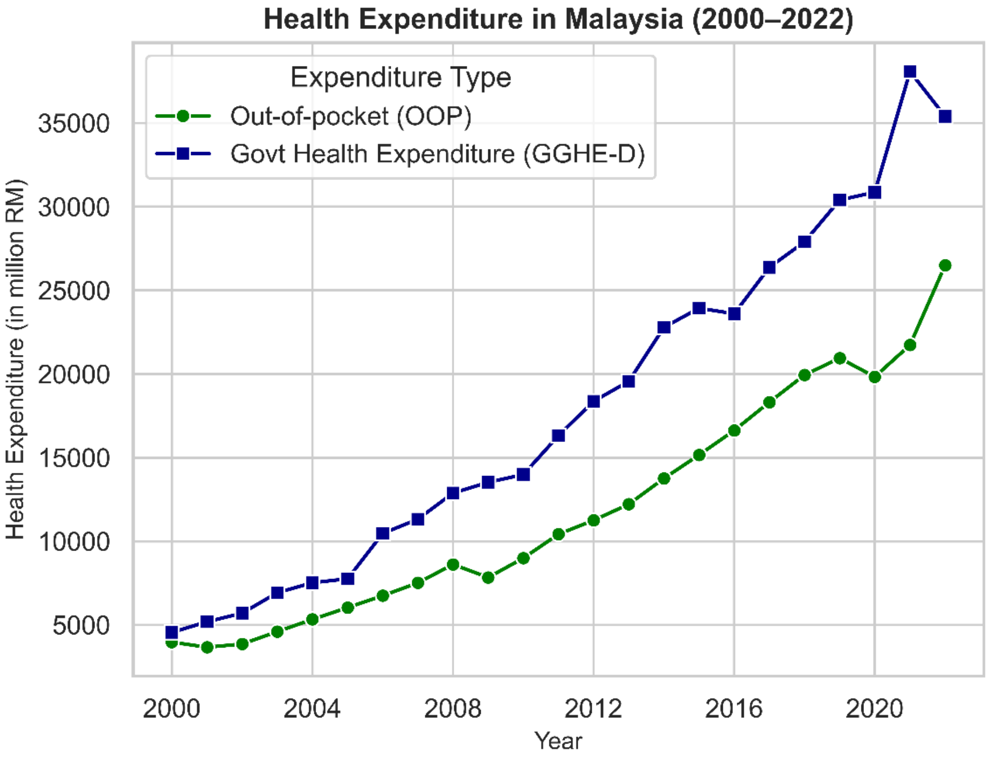


Figure 4.4 Health Expenditure by Type in Malaysia over Year

The two main health expenditure types are plotted on the line chart to show their trend from 2000 to 2022. It can be seen that Domestic General Government Health Expenditure shows a steeper upward trend when compared to Out-of-pocket Health Expenditure, despite both beginning at a similar starting point at 2000 (around RM 4,000 million to RM 5,000 million). There is a steady growth in both expenditure types from 2000 to 2018. It is noticeable that out-of-pocket health expenditure slightly reduced during 2008 to 2009, which is likely related to the 2008 economic crisis, leading to a reduction in individuals' or household health spending.

Health expenditure exhibited fluctuation from 2019 to 2022. GGHE-D shows a sharp rise from 2020 to 2021, acting as the main contributor to the overall increase in total health expenditure, before a slight decline in 2022. The OOP decreased slightly to RM 20,000 million in 2019, then increased steeply to around RM 27,000 million in 2022. This can be suggested by the initial impact on the economy that leads to reduced income and increased unemployment rate due to the lockdown, which is reflected in reduced household healthcare spending. As the number of COVID-19 cases in Malaysia increased between 2020 and 2022, this led to a rise in OOP during the pandemic, due to an increased demand for private healthcare services, for instance, private hospitals, private medical clinics and private pharmacies. (MOH, 2024).

4.3.2 Correlation Between Health Expenditures and Features

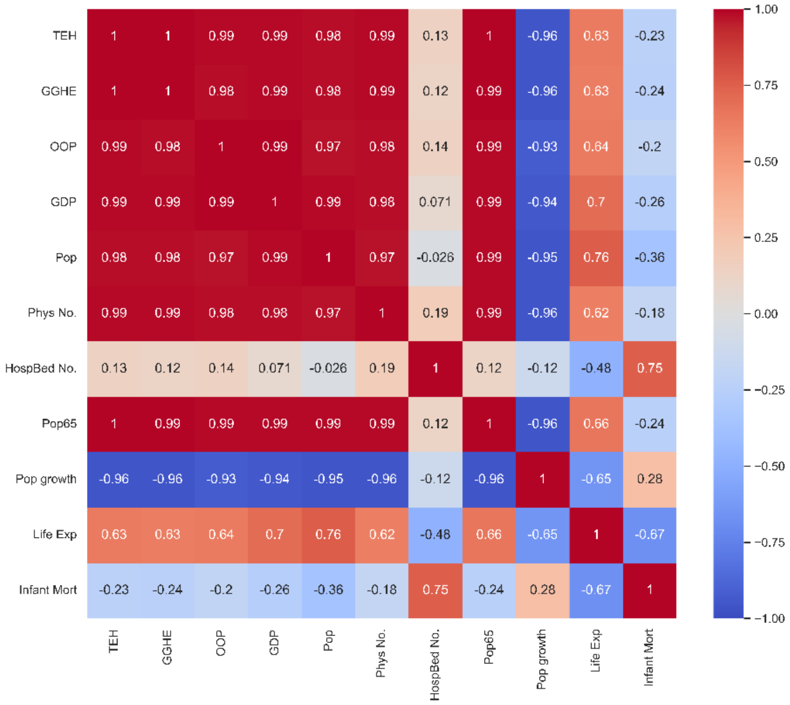


Figure 4.5 Correlation Heatmap

Correlation between the variables is computed and plotted into a heatmap by using the seaborn library. From the correlation heatmap, there is a strong positive correlation between total health expenditure (TEH), domestic general government health expenditure (GGHE-D) and out-of-pocket health expenditure (OOP), due to the fact that GGHE-D and OOP are the components that form TEH. Aside from that, gross domestic product (GDP), population in thousands (Pop), total population aged 65 years old (Pop 65) and number of physicians per 1000 people have a strong positive correlation with all three of the health expenditures. Life expectancy at birth (Life Exp) has a moderate positive correlation with health expenditures, ranging between 0.63 to 0.64.

Population growth in annual % (Pop growth) has a strong negative correlation with health expenditures (-0.96 to -0.93). This indicates that as the population growth rate in Malaysia continues to decline, the health expenditures are still increasing. The rise in health expenditures may be explained by other factors like improvement in healthcare quality or population ageing, as shown by the strong positive correlation with number of physicians and the total population aged 65 years old discussed above. The results aligned with the findings from Khan et. al. (2016) who conducted ADRL test on GDP, life expectancy and population growth from 1981 to 2014. However, their study revealed population above 65 years old have a negative correlation with health expenditures, which contradicts this project's initial findings. This might be explained by the changes in population structure in Malaysia in the recent 20 years, which changed its relationship with health expenditure.

Furthermore, the number of hospital beds has a weak positive correlation to total health expenditure (0.13), general government health expenditure (0.12) and out-of-pocket expenditure (0.14). There is also a weak negative correlation between infant mortality rate and health expenditure, ranging from -0.2 to -0.24. This result is in line with the result from Yap & Selvaratnam (2018), which suggested infant mortality rate is negatively associated with per capita public health spending in Malaysia.

4.4 Feature Engineering

As discussed in the correlation computed, the number of hospital beds and the infant mortality rate weakly correlate with health expenditures. Therefore, these two columns are not selected as the features in the machine learning models for the prediction of health expenditure. This step is to ensure the accuracy of the prediction. Also, the feature reduction can reduce the complexity of the model and reduce computational and time resources.

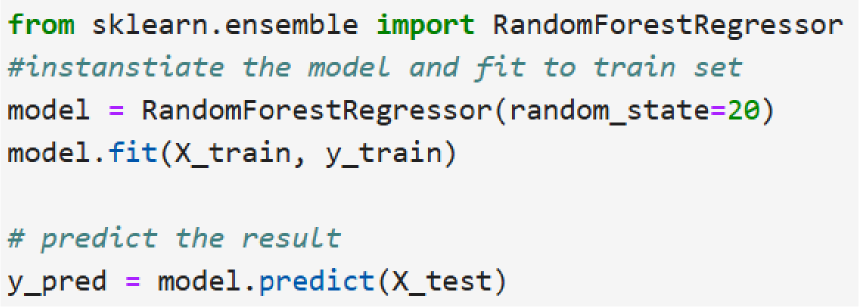
Other features, including GDP, population in thousands, total population aged 65 years old, number of physicians per 1000 people, life expectancy at birth and population growth, are chosen as the predictive indicators for health expenditures to use in Random Forest model, supported by the literature and exploratory data analysis done on these features.

4.5 Initial Modelling

The initial modelling was conducted only on total health expenditure, without investigating deeper into its expenditure types (GGHE-D and OOP). Random Forest and ARIMA are conducted using the pre-processed dataset. For Random Forest, features discussed in the feature engineering section are used for prediction, while for ARIMA, it is modelled based on its lagged data points.

4.5.1 Random Forest

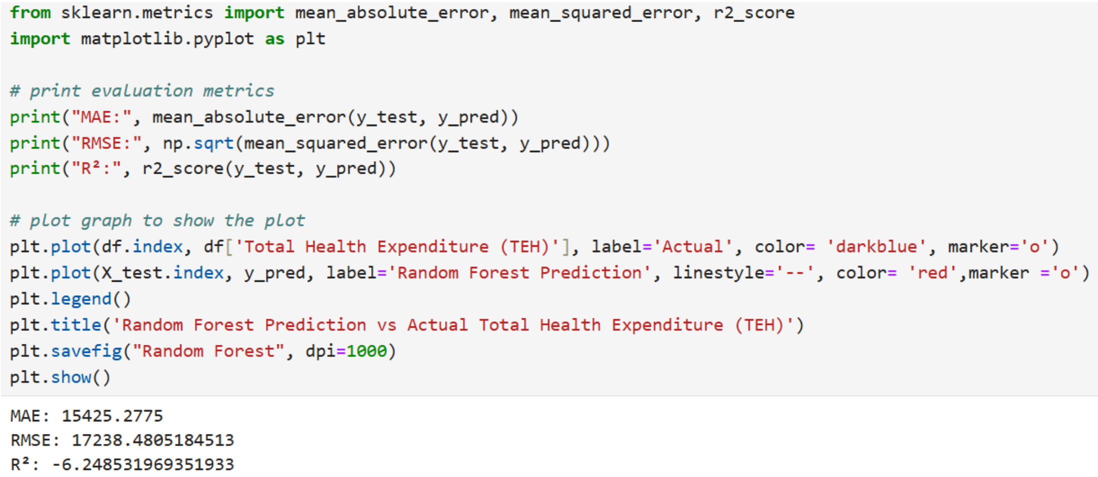
Random Forest Regressor is imported from sklearn.ensemble library. The model is instantiated with default parameters and fit to the training set. Prediction is done by using the X_test and the result is saved as y_pred. The process described is shown in Figure 4.6. Then, the evaluation metrics are imported from the sklearn library and calculated from the difference between the prediction result with the actual total health expenditure, as shown in Figure 4.7. A line chart for comparison is plotted as shown in Figure 4.8.



```python
from sklearn.ensemble import RandomForestRegressor
#instanstiate the model and fit to train set
model = RandomForestRegressor(random_state=20)
model.fit(X_train, y_train)

# predict the result
y_pred = model.predict(X_test)
```

Figure 4.6 Random Forest Modelling



```python
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
import matplotlib.pyplot as plt

# print evaluation metrics
print("MAE:", mean_absolute_error(y_test, y_pred))
print("RMSE:", np.sqrt(mean_squared_error(y_test, y_pred)))
print("R²:", r2_score(y_test, y_pred))

# plot graph to show the plot
plt.plot(df.index, df['Total Health Expenditure (TEH)'], label='Actual', color= 'darkblue', marker='o')
plt.plot(X_test.index, y_pred, label='Random Forest Prediction', linestyle='--', color= 'red',marker ='o')
plt.legend()
plt.title('Random Forest Prediction vs Actual Total Health Expenditure (TEH)')
plt.savefig("Random Forest", dpi=1000)
plt.show()
```

```
MAE: 15425.2775
RMSE: 17238.4805184513
R²: -6.248531969351933
```

Figure 4.7 Compute Evaluation Metrics and Plotting Line Chart for Prediction

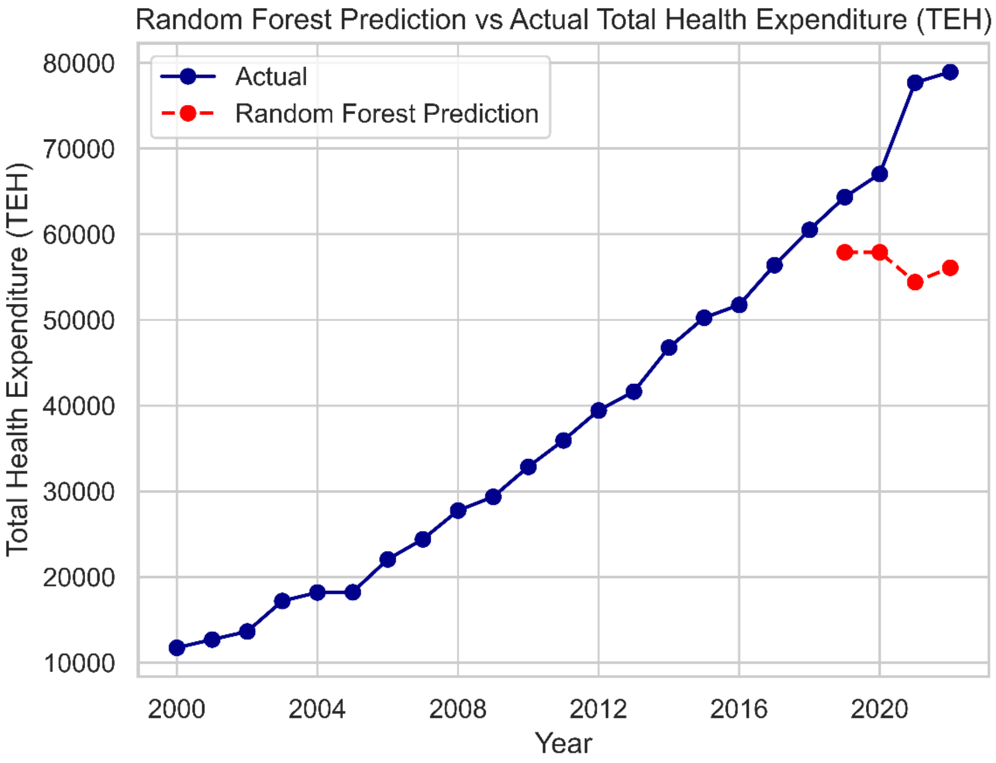
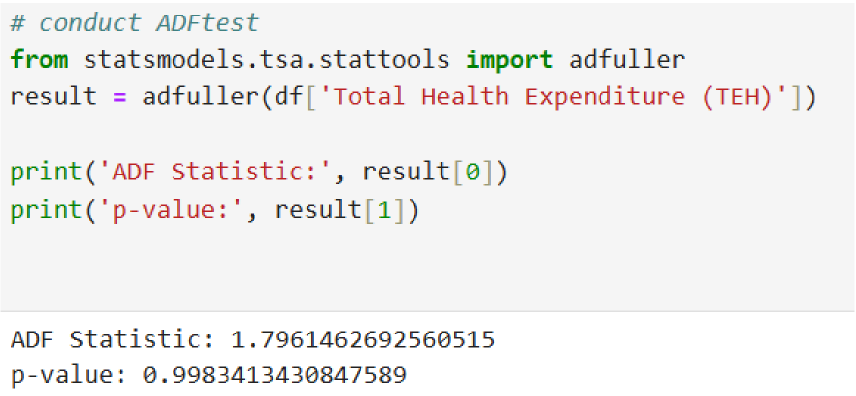


Figure 4.8 Random Forest Prediction versus Actual Total Health Expenditure (2000 to 2022)

4.5.2 ARIMA

Before modelling with ARIMA, the stationarity of the data must be confirmed. From the exploratory data analysis, it can be seen that there is a trend of increase for total health expenditure. The ADF test is run to confirm stationarity of the time series data to determine the need for differencing. Adfuller is imported from statsmodels and applied to the total health expenditure column. The results in Figure 4.9 show that the p-value is 0.9983, which means differencing needs to be done. The time series data is differenced once and saved as ‘TEH_diff1’.



```
# conduct ADFtest
from statsmodels.tsa.stattools import adfuller
result = adfuller(df['Total Health Expenditure (TEH)'])

print('ADF Statistic:', result[0])
print('p-value:', result[1])
```

```
ADF Statistic: 1.7961462692560515
p-value: 0.9983413430847589
```

Figure 4.9	Augmented Dickey-Fuller Test Result

Next, ACF and PACF plot is conducted on the ‘TEH_diff1’ column to determine the order for ARIMA *p* and *q* terms. From the plot shown in Figure 4.10, it can be seen that the cut-off point is at 0. Therefore, p and q are set as 0 for the ARIMA model.

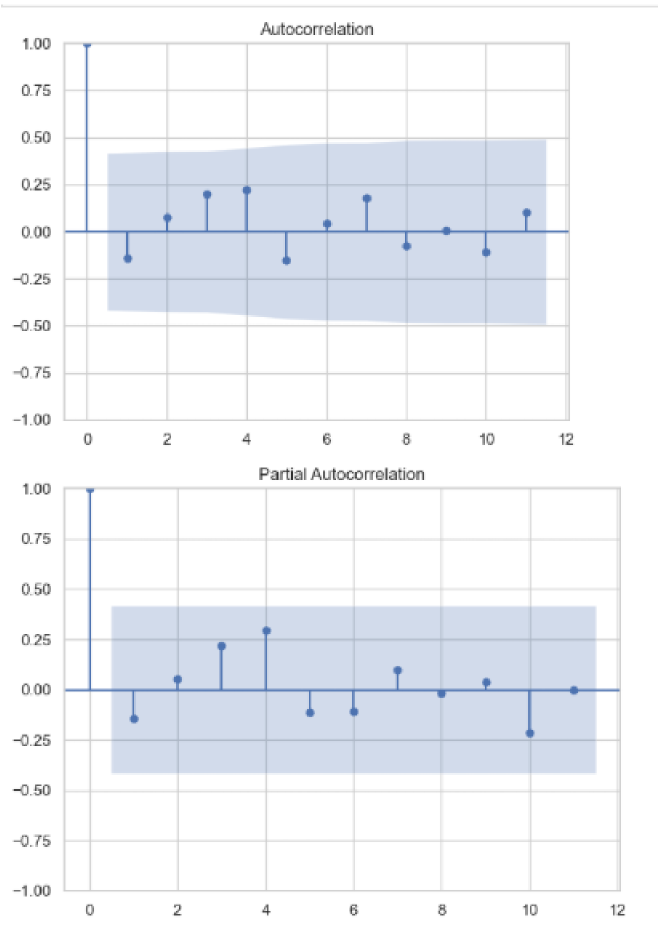
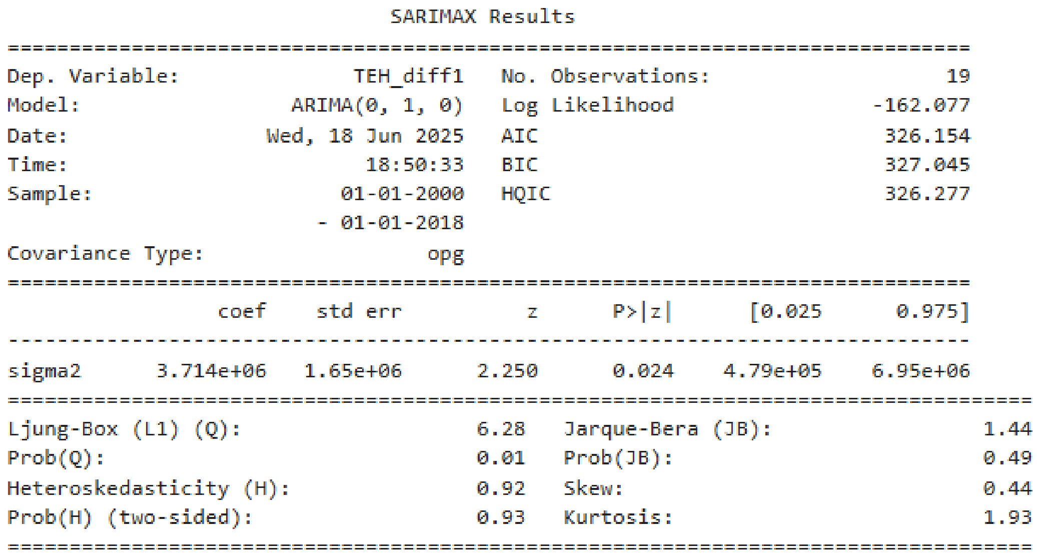


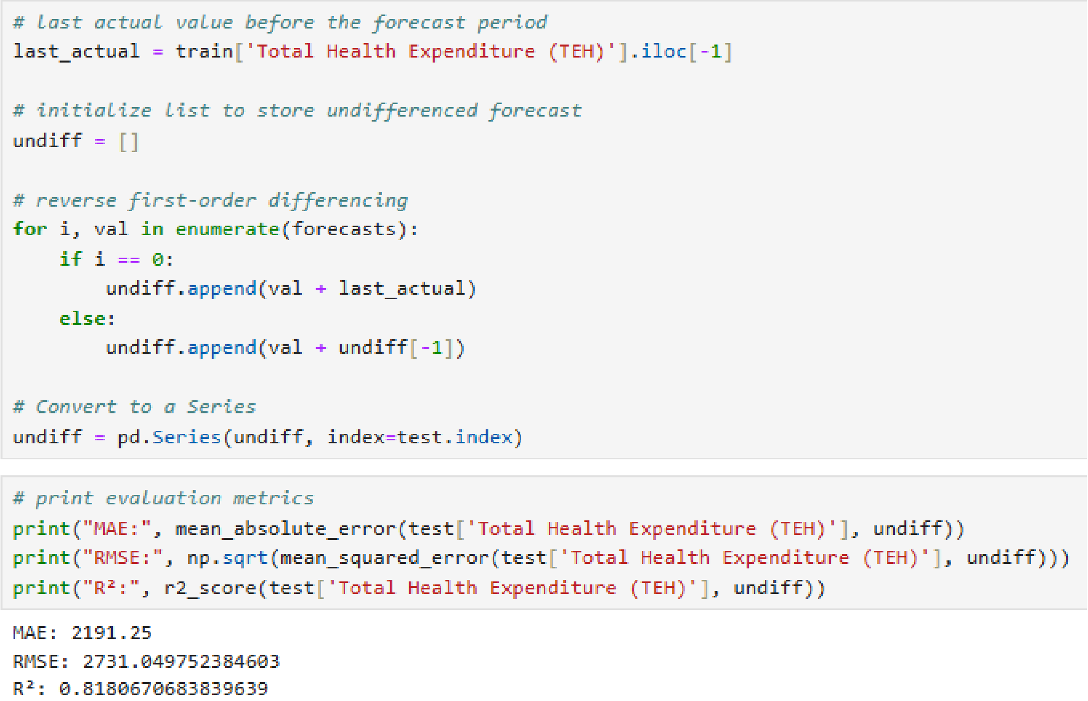
Figure 4.10	ACF and PACF Plot Result

The next step is to split the data into a training set and testing set. For this project, the data is split into 80% training data and 20% testing data in a similar way to that conducted for Random Forest. The training data is fit into the ARIMA (0, 1, 0) model, and a summary of the model is printed. Then, a forecast is made by using the test set, with evaluation metrics generated and the forecasted result plotted to compare with the actual total health expenditure in Figure 4.11. The results are discussed in the next section.



```
                               SARIMAX Results
==============================================================================
Dep. Variable:              TEH_diff1   No. Observations:                   19
Model:                 ARIMA(0, 1, 0)   Log Likelihood                -162.077
Date:                Wed, 18 Jun 2025   AIC                            326.154
Time:                        18:50:33   BIC                            327.045
Sample:                    01-01-2000   HQIC                           326.277
                         - 01-01-2018
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2      3.714e+06   1.65e+06      2.250      0.024    4.79e+05    6.95e+06
===================================================================================
Ljung-Box (L1) (Q):                   6.28   Jarque-Bera (JB):                 1.44
Prob(Q):                              0.01   Prob(JB):                         0.49
Heteroskedasticity (H):               0.92   Skew:                             0.44
Prob(H) (two-sided):                  0.93   Kurtosis:                         1.93
===================================================================================
```

Figure 4.11 ARIMA Modelling and Summary



```
# last actual value before the forecast period
last_actual = train['Total Health Expenditure (TEH)'].iloc[-1]

# initialize list to store undifferenced forecast
undiff = []

# reverse first-order differencing
for i, val in enumerate(forecasts):
    if i == 0:
        undiff.append(val + last_actual)
    else:
        undiff.append(val + undiff[-1])

# Convert to a Series
undiff = pd.Series(undiff, index=test.index)
```

```
# print evaluation metrics
print("MAE:", mean_absolute_error(test['Total Health Expenditure (TEH)'], undiff))
print("RMSE:", np.sqrt(mean_squared_error(test['Total Health Expenditure (TEH)'], undiff)))
print("R²:", r2_score(test['Total Health Expenditure (TEH)'], undiff))
```

```
MAE: 2191.25
RMSE: 2731.049752384603
R²: 0.8180670683839639
```

Figure 4.12 Reverse First-order Differencing and Compute Evaluation Metrics

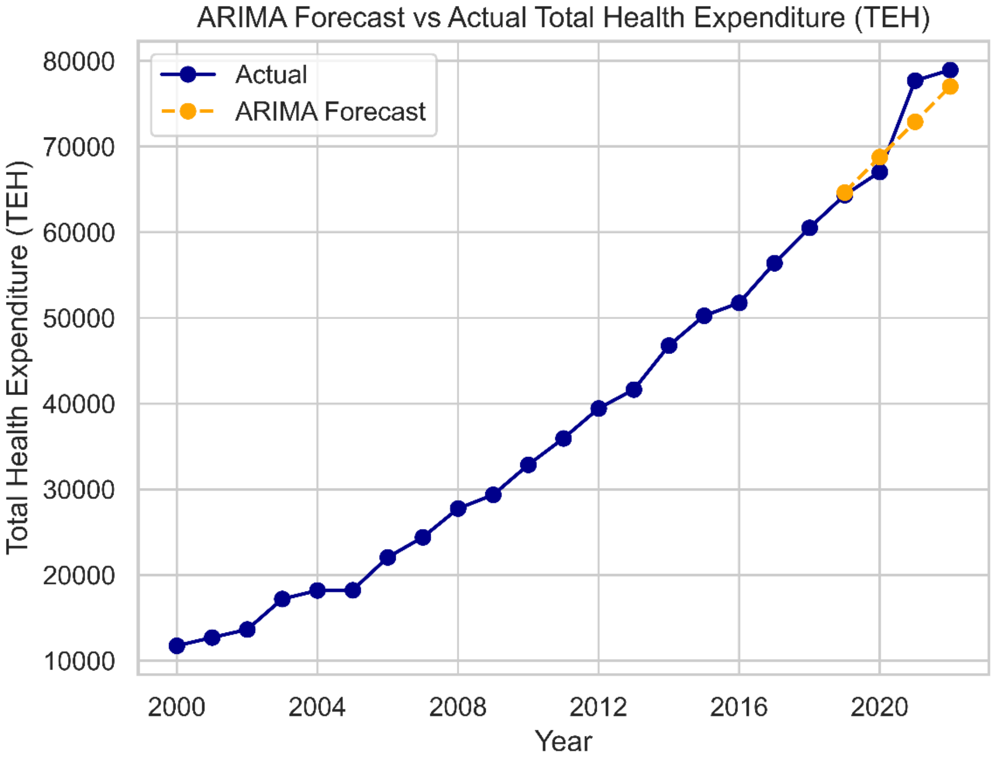


Figure 4.13 ARIMA Forecast versus Actual Total Health Expenditure (2000 to 2022)

4.6 Result Evaluation and Discussion

Table 4.1 Result Evaluation

Models	Mean Absolute Error (MAE)	Root Mean Squared Error	Coefficient of Determination
Random Forest	15314	16951	-6
ARIMA	2191	2731	0.818

ARIMA has an acceptable forecast result with the actual TEH, with a small root mean squared error of RM 2731 million and a high R^2 of 0.818. The result in Figure 4.13 shows that aside from a larger gap with the actual TEH in 2021, the prediction data points lie closely with the actual TEH. This outcome is justifiable as the increased health expenditure caused by the impact of the pandemic was unexpected and difficult to predict.

The predicted result from the random forest is far from accurate when compared with the actual values for 2019 to 2022, as shown by RMSE of RM 16,951 million and negative R^2 of -6. This is likely due to lagged values of health expenditure not being provided to the model as a feature, which can be included in future to improve the model result. Also, due to COVID-19, there is a steep increase in health expenditure from 2020 to 2021, likely due to increased health budget allocated for COVID-19-related health expenses, which is an aspect not learned by the model due to a lack of learning data. Furthermore, cross-validation and hyperparameter tuning were not done to utilise the model's full power. The small data size might be a limiting factor for random forest as not enough features are provided for the model to learn.

4.7 Summary

Results from exploratory data analysis show that there is a gradual increase in health expenditures from 2000 to 2022, with some fluctuation during 2020 to 2022 due to the COVID-19 pandemic. The initial result of machine learning models suggests that ARIMA outperform random forest without tuning in terms of total health expenditure prediction and makes an accurate prediction despite the fluctuation in health expenditures during the pandemic. Several improvements can be made to increase the models' accuracy, validate the results and extend the research outcomes. These will be further discussed in the future works section in the next chapter.