PROJECT PROPOSAL PRESENTATION

PREDICTION OF HEALTH EXPENDITURE IN MALAYSIA USING MACHINE LEARNING

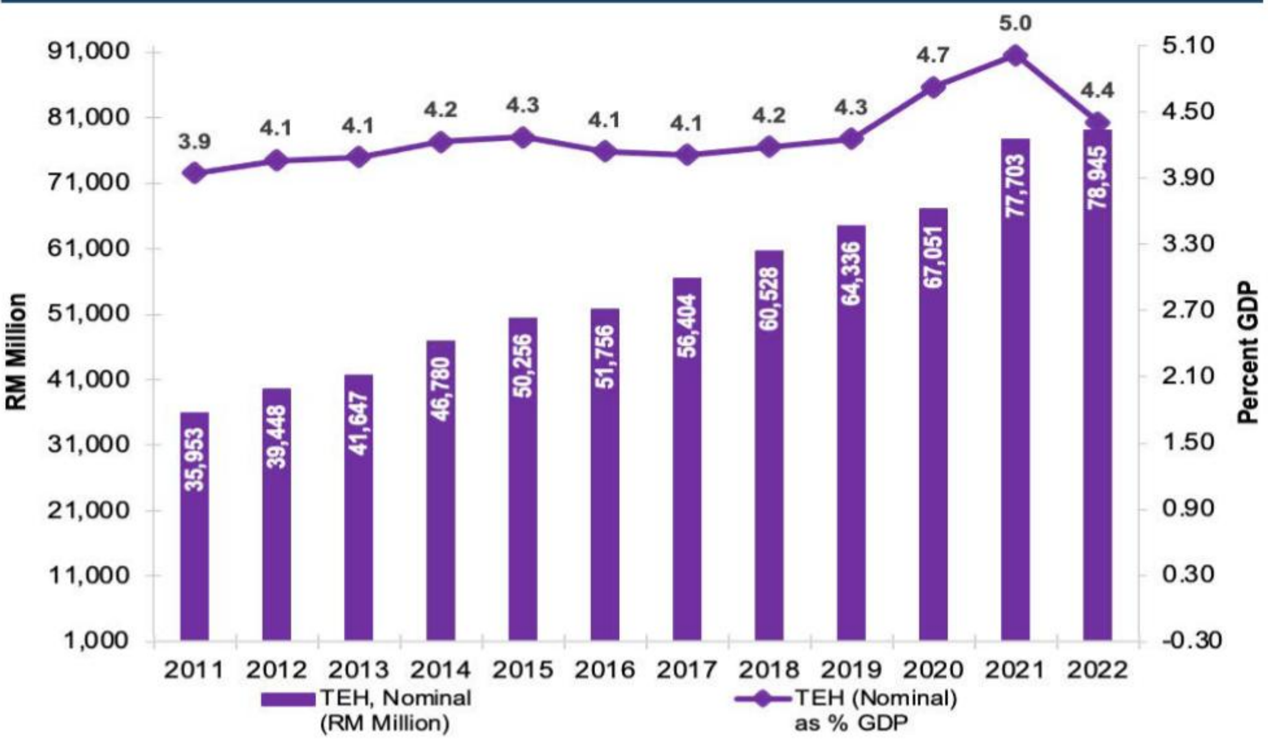
CANDIDATE : LOCK CHUN HERN MCS241047
LECTURER : ASSOC. PROF. DR MOHD SHAHIZAN BIN OTHMAN
VENUE : ONLINE
DATE : 29 JUNE 2025
VIDEO LINK : https://youtu.be/RfOSxvLuB18

Table of Contents

Research Introduction

Problem Background

FIGURE 4.1: Trend for Total Expenditure on Health, 2011-2022 (RM Million & Percent GDP)



Growing medical costs present major challenges for healthcare sector

2024 @ 4:31pm

malaymail

HOME MALAYSIA SINGAPORE MONEY WORLD LIFE EAT/DRINK SHOWBIZ OPINION SPORTS TECH/GADGETS

MALAYSIA

Malaysia's soaring medical inflation: How rising costs are straining public, private healthcare

Malaysia's health expenditure to grow by 8.3% CAGR until 2028

Its medium-term health expenditure growth is expected to be one of the fastest in the ASEAN.

Health expenditure in Malaysia is projected to grow at a compound annual growth rate (CAGR) of 8.3% until 2028 on the back of strong government support for the sector, according to a BMI report.

According to a 2023 report from Bank Negara Malaysia (BNM), Malaysia recorded medical inflation at 12.6 percent, significantly higher than the global average of 5.6 percent.

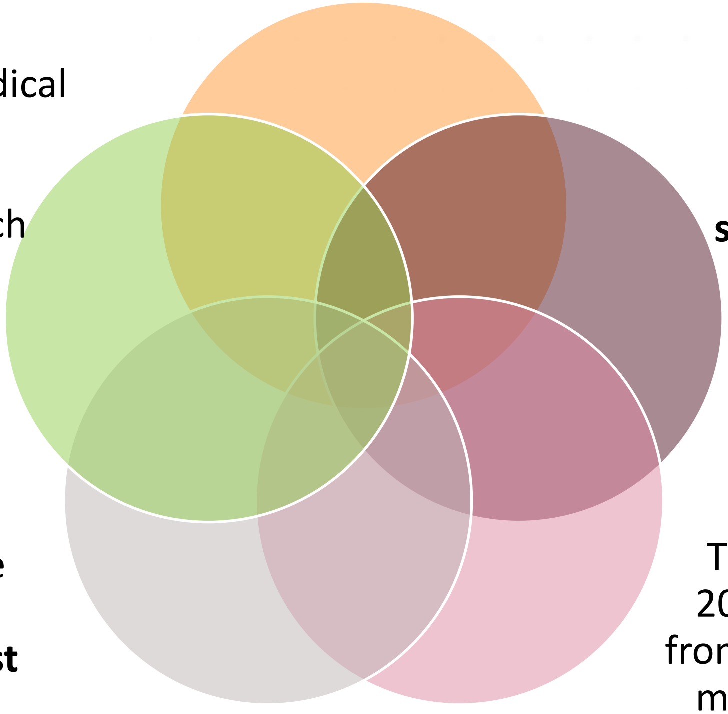
KUALA LUMPUR: Health Minister Datuk Seri Dr Dzulkefly Ahmad said that the rising cost of medical care in Malaysia is a concerning trend that highlights significant challenges within the healthcare sector. —BERNAMA

(Ministry of Finance Malaysia, 2024) (Healthcare Asia, 2024) (Parzi, 2024) (Hassandarvish, 2024)

Problem Background

Health expenditure is defined as all the money spent on health goods and services, including preventative measures, promotion and provision of health services, nutrition, pharmaceuticals, and emergency aid. (World Health Organisation [WHO], 2025).

With the inflation in medication prices, medical expenses, and the ageing population in Malaysia, it is anticipated that **healthcare expenditure will continue to increase**, which post challenges to government and people.



Health financing in Malaysia is largely funded by **public funding** 52.3% (RM 41,257 million) of total health expenditure, followed by **private sources of financing**, accounting for RM 37,688 million (47.7%)

Malaysia government has allocated **RM45.3 billion in Budget 2025** for the Ministry of Health for spending on healthcare, which is the **second highest after education** (Ministry of Finance Malaysia, 2024)

The TEH of Malaysia is gradually increasing from 2011 to 2022, showing more than **2-fold increase** from RM 35,953 million (3.94% as GDP) to RM78,945 million (4.41% as GDP), and a **significant increase** can be seen in 2022 compared to pre-COVID-19 pandemic value (MOH, 2024).

Problem Statement & Research Question

UTM UNIVERSITI TEKNOLOGI MALAYSIA

Problem Statements	Research Question
Determinants of healthcare expenditure are complex and varies depending on the prediction models. Selecting the most appropriate determinants is vital for optimal health spending prediction.	What are the **key determinants** of health expenditure that contribute to accurate prediction in machine learning algorithms?
Growing health expenditure in Malaysia is a significant challenge for government. Accurate prediction for each components of health expenditure is required for informed decision making.	What is the **predicted health expenditure** of Malaysia from 2026 to 2035 using machine learning algorithms ?
Machine learning algorithms perform differently depending on the context of available dataset. Assessment of different models is required to determine the best model in predicting health expenditure of Malaysia.	How do different machine learning models perform on Malaysia's health expenditure data, and which model demonstrates the **highest accuracy**?

Aim and Objectives

Research Aim:

The aim of this research is to **predict health expenditure** in Malaysia using machine learning techniques to provide insight for health financing and policy planning.

Research Objectives:

a) To identify the **key determinants** of health expenditure to use as features for machine learning algorithms

b) To apply **Random Forest** and **ARIMA** for predicting health expenditure in Malaysia from 2026 to 2035

c) To evaluate and compare the **performance metrics** of the machine learning models and to identify the model with the highest **accuracy** in forecasting health expenditure in Malaysia

SCOPES OF STUDY

SOURCE OF DATA

KEMENTERIAN KESIHATAN MALAYSIA

World Health Organization

THE WORLD BANK

SCOPE OF DATA

- Data related to health economics
- Demographic Data
- Open Source
- No individual data (e.g. medical history and medication history)

TIME FRAME OF DATA

From year 2000 to year 2022

MACHINE LEARNING TECHNIQUES

- ARIMA
- Random Forest(RF)

Expected Research Contribution

- Provide insights for policymakers in the country in planning health expenditures and allocation of budgets
- Ensure the long-term sustainability of health financing
- Contribute to better health outcomes for the patients and people in Malaysia
- Provide insights for other countries with similar composition of healthcare systems or income levels

Literature Review

Literature Review

Health expenditure can be represented by Total Health Expenditure (**TEH**), Current Health Expenditure (**CHE**), which excludes health-related expenditure (e.g., personnel training, research and development), General Government Health Expenditure (**GGHE**), and household Out-Of-Pocket health expenditure (**OOP**). (WHO, 2025).

Ku Abd Rahim et al. (2020) conducted a systematic review on the economic evaluation of healthcare in Malaysia and highlighted that publications related to health economics are **sparse and inadequate to meet stakeholders' and policymakers' needs.**

Public spending on health in Malaysia remains **lower** when compared to the average 6.3% of GDP spent in middle-income countries (WHO, 2020).

Low public health spending may contribute to a range of issues **like chronic understaffing, high workload, and critical infrastructure shortages.**

At the same time**, rising OOP payments and increased pharmaceutical costs** create a potential risk to the healthcare system, justifying the need for economic evaluation for health policy planning (Khor et. al, 2024).

By using **predictive analytics** and focusing on cost savings, initiatives can be taken to improve patient access to affordable healthcare services, reduce healthcare costs, improve efficiency in the healthcare system, and ultimately improve patient outcomes (Devi & Bansal, 2024).

Research Gaps

01

Limited academic research on Malaysia health expenditure prediction model despite needs. In contrast, many countries have researches on health expenditure using advanced forecasting techniques to support health financing decisions.

02

Health expenditure is affected by **multiple factors** and traditional models struggle to capture complex and non-linear relationship.

03

Different determinants (features) of total health expenditure are used in the studies from different countries and there are **no general consensus** between studies

Existing Models

Exponential Smoothing & Regression

- Time series analysis and exponential smoothing models [ETS] are employed in the forecasting processes **(Sahoo et al.,2023)**
- Vector autoregressive regression (VAR) and ordinary least squares (OLS) regression models **(Kazemian et al., 2022)**

Grey Model

- Trend prediction of China's THE to 2030 using GM (1,1) model **(Li & Zhang, 2024)**
- New Structure of the Multivariate Gray Prediction Model NSGM(1,10) **(Jia et al.,2021)**
- Optimized GM (1,1) and nonlinear grey Bernoulli model (NGBM) (1,1) **(Özcan & Tüysüz, 2018)**

Other Machine Learning

- Random Forest, Support Vector Regression (SVR) and traditional statistical forecasting methods (logistic regression, ARIMA) **(Wang et al., 2024)**
- Neural network models, the Adaptive Neuro-Fuzzy Inference System (ANFIS) and the Hybrid Neural Fuzzy Inference System (HyFIS) in Jordan **(Saleh et al., 2023)**
- Five AI models (RF, ANN, Multiple linear regression (MLR), SVR, Relevance Vector Machine (RVM) with combination of genetic-algorithm feature selection in Turkey **(Ceylan & Atalan, 2020).**

ARIMA

- Autoregressive Integrated Moving Average (ARIMA) model **(Klazoglou & Dritsakis, 2018)**
- ARIMA model for forecasting health expenditure 5 years forward China (**Zheng et al. , 2020)**
- ARIMA for the Brazil, Russia, India, China, and South Africa (BRICS) countries to forecast until 2030 **(Jakovljevic et al., 2022)**

Strength and Weakness of Models

Models	Strength	Limitation	References
Grey Model	A small amount of data is needed Low data distribution requirement Predict better than a back propagation neural network when the data is fewer	Poor long-term forecasting Univariate prediction model does not capture complex patterns in data Prediction performance of a multivariable model may be affected by the correlation among variables	(Li & Zhang, 2024), (Jia et al., 2021)
Exponential Smoothing Model	Gives more weight to the recent outcomes than past observations Automatically select the best-fitting model based on data error, trend, and seasonal components	Limited incorporation of external factors Assumption of continuity in historical pattern	(Sahoo et al.,2023)

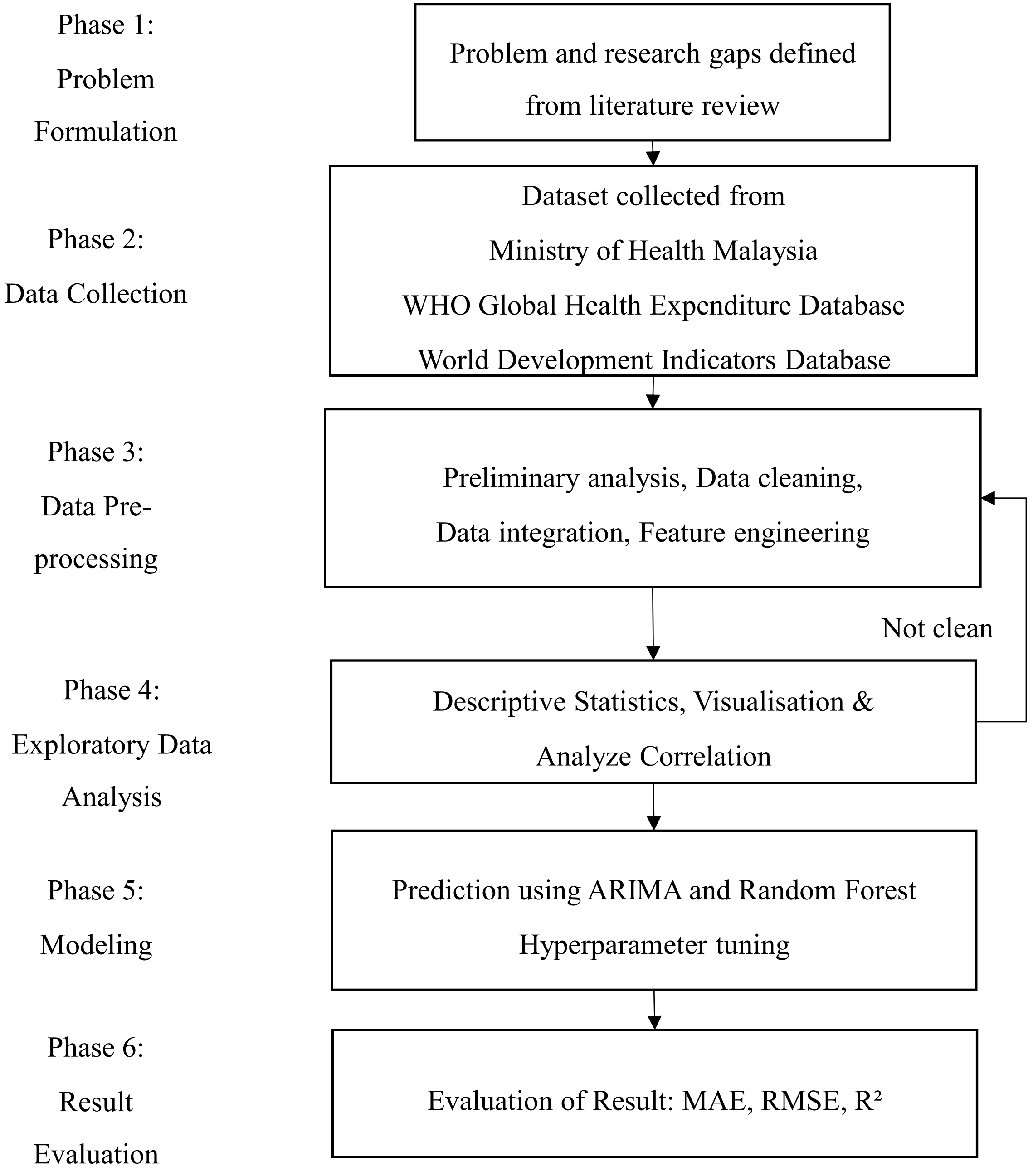
Models	Strength	Limitation	References
Autoregressive Integrated Moving Average (ARIMA)	Explainability Flexibility Better performance for the small dataset Suitable for short-term forecasting Smaller computational requirements	Univariate modelling Vulnerable to changes in other fields Difficulty in forecasting complex real-world problems More sensitive to outliers Uncertainty if the prediction interval is large	(Zheng et al., 2020), (Jakovlje et al., 2022), (Kontopoulou et al., 2023)
Artificial Neural Network	Able to manage large and complex data Non-linear time dependencies Can combine the forecasts of multiple time series	Require a large amount of training data Require optimization Computationally expensive Time-consuming Low explainability	(Kontopoulou et al., 2023), (Ahmed et al, 2023)
Random Forest	Able to incorporate multiple factors Less affected by missing values Lower risk of overfitting and bias Lower overall variance Better generalization capability	Computationally expensive Higher memory usage Time-consuming Less interpretable than an individual decision tree	(Wang et.al., 2024), (Ceylan & Atalan, 2020), (Muremyi et. al, 2020)

Determinants of Health Expenditure

References	Determinants of Health Expenditure	Data sources
(Ceylan & Atalan, 2020)	**GDP** per capita**, Life expectancy at birth**, Unemployment rate, Crude Birth rate, **No. of Hospital** and **No. of physician**	OECD library
(Saleh et al., 2023)	No. of physicians, No. of beds in hospitals**, population size**, and consumer price index	WHO, Jordan's Ministry of Health, Central Bank of Jordan.
(Jia H. et al, 2021)	Number of people **> aged 65**, Population, GDP, number of medical personnel, No. of beds in hospital, GGHE, OOP**, infant mortality rate**, household consumption expenditure.	National data sourced from China Statistical Yearbook and China NHA Report
(Lorenzoni, 2019)	Percentage of population over 65 years old, GDP per capita elasticity, Baumol coefficient (wage over productivity), technology progress (country research and development spending as a share of GDP), and mortality	National sources of the countries in the OECD and the Eurostat HEDIC (Health Expenditure by Disease and Condition) report

Methodology

Research Framework



Data Collection

No	Dataset	Year	File size	Columns	Rows	Variables	Sources
1	MHNA_2022.csv	2011 - 2022	1KB	3	13	Total Health Expenditure **(TEH)** in million (MYR)	(MOH, 2024)
2	MHNA_2017.csv	1997 - 2017	1KB	3	22	Total Health Expenditure **(TEH)** in million (MYR)	(MOH, 2019)
3	NHA indicators.xlsx	2000-2022	8KB	26	8	Current Health Expenditure (CHE) in million (MYR), Domestic General Government Health Expenditure (GGHE-D) in million (MYR), Domestic Private Health Expenditure (PVT-D) in million (MYR), Out-of-pocket (OOPS) as % of Current Health Expenditure (CHE), Population Size (in thousands)	(World Health Organization, 2025b)
4	P_Data_Extract_From_World_Development_Indicators.csv	2000-2022	2KB	27	12	Gross Domestic Product (GDP), Number of Physicians (per 1000 people), Number of Hospital Beds (per 1000 people), Population aged 65 years old and above (total), Infant Mortality Rate (per 1000 live births), Population Growth (annual %), Life Expectancy at birth (total years)	(The World Bank, World Development Indicators, 2025)

Data Pre-processing: Preliminary analysis

pd.read_csv OR pd.read_excel on all datasets and view head and info

Explore & understand the data and decide next steps

```
# import dataset
df1_extracted = pd.read_csv('mhna_2017.csv', index_col=0)
df2_extracted = pd.read_csv('mhna_2022.csv', index_col=0)

# view info for both csv
print(df1_extracted.head())
print(df2_extracted.head())
```

```
   Year  TEH_Nominal
0  1997        8,550
1  1998        9,156
2  1999        9,953
3  2000       11,745
4  2001       12,703
   Year  TEH_Nominal
0  2011       35,953
1  2012       39,448
2  2013       41,647
3  2014       46,780
4  2015       50,256
```

```
#read the NHA indicator xlsx
who_nha_ind = pd.read_excel('NHA indicators.xlsx')
who_nha_ind.head()
```

	Countries	Indicators	Unnamed: 2	2000	2001	2002	2003	2004
0	NaN	NaN	NaN	Value	Value	Value	Value	Value
1	Malaysia	Current Health Expenditure (CHE)	Million NCU	9761.00293	10273.686523	11138.546875	13336.99707	14767.112305

```
# read data from WDI
wdi_data= pd.read_csv('WDI_Data.csv')
wdi_data
```

	Country Name	Country Code	Series Name	Series Code	2000 [YR2000]	2001 [YR2001]	2002 [YR2002]	2003 [YR2003]	2004 [YR2004]	2005 [YR2005]	...	2013 [YR2013]
0	Malaysia	MYS	Physicians (per 1,000 people)	SH.MED.PHYS.ZS	0.681000	..	0.723000	7.350000e-01	7.200000e-01	7.760000e-01	...	1.557000e+00
1	Malaysia	MYS	Hospital beds (per 1,000 people)	SH.MED.BEDS.ZS	2.050000	2.01	1.960000	1.920000e+00	1.890000e+00	1.870000e+00	...	1.910000e+00

Data Cleaning

-Correct data types

-Identify and removing errors (incorrect values, missing data, duplicates)

-Fixing format (round, remove ‘ ‘ or ,)

```
#change datatype of GDP to integer
who_nha_ind['Gross Domestic Product (GDP)']= who_nha_ind['Gross Domestic Product (GDP)'].astype(int)
```

```
# Rename TEH_Nominal column for easier view
df_combined = df_combined.rename(columns={'TEH_Nominal': 'Total Health Expenditure (TEH)'})

# Replacing the comma in between the number
df_combined['Total Health Expenditure (TEH)'] = df_combined['Total Health Expenditure (TEH)'].str.replace(',', '')
df_combined
```

```
# fill in missing value for physician data
# fill first missing value, interpolating from the back and forward value
wdi_data.loc[1, 'Physicians (per 1,000 people)'] = (wdi_data.loc[0, 'Physicians (per 1,000 people)'] + wdi_data.loc[2, 'Physicians (per 1,000 people)'])

# for missing value in 2022, fill using forward fill
wdi_data.loc[22, 'Physicians (per 1,000 people)'] = wdi_data.loc[21, 'Physicians (per 1,000 people)']

# fill in missing value for hospital beds data, using forward fill
wdi_data.loc[22, 'Hospital beds (per 1,000 people)'] = wdi_data.loc[21, 'Hospital beds (per 1,000 people)']
```

Data Cleaning

-Transform the data into suitable structure
-Fix header, dropping unused row & columns
-Setting year as index

```
#inverse the row and column after setting first column index
wdi_data= wdi_data.set_index(wdi_data.columns[0])
wdi_data = wdi_data.T
```

```
# setting year as index
df['Year'] = pd.to_datetime(df['Year'], format='%Y')
df.set_index('Year', inplace=True)
```

```
#fix structure of wdi_data by dropping unused column
wdi_data = wdi_data.drop(columns= ['Country Name','Country Code','Series Code'])

#drop unused rows
wdi_data = wdi_data.drop(wdi_data.index[6:])

wdi_data
```

Data Integration

- Data integration refers to the compilation of datasets from various sources into a unified dataset.
- Concatenation is done for Total Health Expenditure (TEH)
- The datasets are merged together into a single dataset after the cleaning process.
- Pandas' merge method will be used for this function, with an inner join chosen and merge on the 'Year' column.

```
#combine 2 table into one according to year using concat
df_combined = pd.concat([df1_2000_2010, df2_extracted])
print(df_combined)
```

```
#merging data
df= df_combined.merge(who_nha_ind, on='Year', how= 'outer').merge(wdi_data, on='Year', how= 'outer')
df
```

Feature Engineering

- Feature refers to variable in the dataset. In this step, the new feature was calculated from the existing features.
- Feature selection was carried out at the end of the data pre-processing step. This step is important in improving model performance and reducing overfitting. The feature selection step selects the appropriate variables that are being used for the predictive modelling.

```
# transformation of OOPS into actual number instead of percentages
who_nha_ind['Out-of-pocket Health Expenditure (OOP)'] = (
    who_nha_ind['Out-of-pocket (OOPS) as % of Current Health Expenditure (CHE)'] / 100 *
    who_nha_ind['Current Health Expenditure (CHE)']
)
```

Exploratory Data Analysis

- Descriptive Analysis of Health Expenditures
- Matplotlib
- Seaborn library
- Visualization for features
- Correlation heatmap

```
# import necessary libraries for visualization
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Create a line plot showing the total health expenditure over time
sns.set(style="whitegrid")
sns.lineplot(data= df, x= 'Year', y='Total Health Expenditure (TEH)',marker='o')
plt.title('Total Health Expenditure (TEH) in Malaysia From 2000 to 2022', fontsize=12, fontweight='bold')
plt.xlabel('Year', fontsize=10)
plt.ylabel('Health Expenditure (in million RM)', fontsize=10)
plt.savefig("TEH.png", dpi=1000)
plt.show()
```

Correlation Heatmap

```
#rearrange column
df = df[['Total Health Expenditure (TEH)','Domestic General Government Health Expenditure (GGHE-D)','Out-of-pocket Health Expenditure(OOP)',
         'Gross Domestic Product (GDP)', 'Population (in thousands)', 'Physicians (per 1,000 people)',
         'Hospital beds (per 1,000 people)', 'Population ages 65 and above, total',
         'Population growth (annual %)', 'Life expectancy at birth, total (years)',
         'Mortality rate, infant (per 1,000 live births)']]

#rename the column using shortform for easier views
short_name= {'Total Health Expenditure (TEH)':'TEH','Domestic General Government Health Expenditure (GGHE-D)':'GGHE',
             'Gross Domestic Product (GDP)' :'GDP', 'Population (in thousands)' :'Pop', 'Out-of-pocket Health Expenditure(OOP)':'OOP',
             'Physicians (per 1,000 people)' :'Phys No.', 'Hospital beds (per 1,000 people)':'HospBed No.',
             'Population ages 65 and above, total':'Pop65', 'Mortality rate, infant (per 1,000 live births)':'Infant Mort',
             'Population growth (annual %)' :'Pop growth', 'Life expectancy at birth, total (years)' :'Life Exp'}

df_acronym= df.rename(columns= short_name)

# view correlation between variables
corr= df_acronym.corr()
corr
```

```
# generate heatmap
plt.figure(figsize=(12,10))
sns.heatmap(data=corr, cmap= 'coolwarm', vmin= -1, vmax= 1,annot= True, annot_kws={"size": 12})
plt.savefig("correlation heatmap.png", dpi=1000)
```

Algorithms Selection

ARIMA

-Univariate

-Better performance for small dataset

-Explainability

-Flexibility

-Suitable for short-term forecasting

-Smaller computational requirements

Random Forest

-Multivariate

-Less affected by missing values

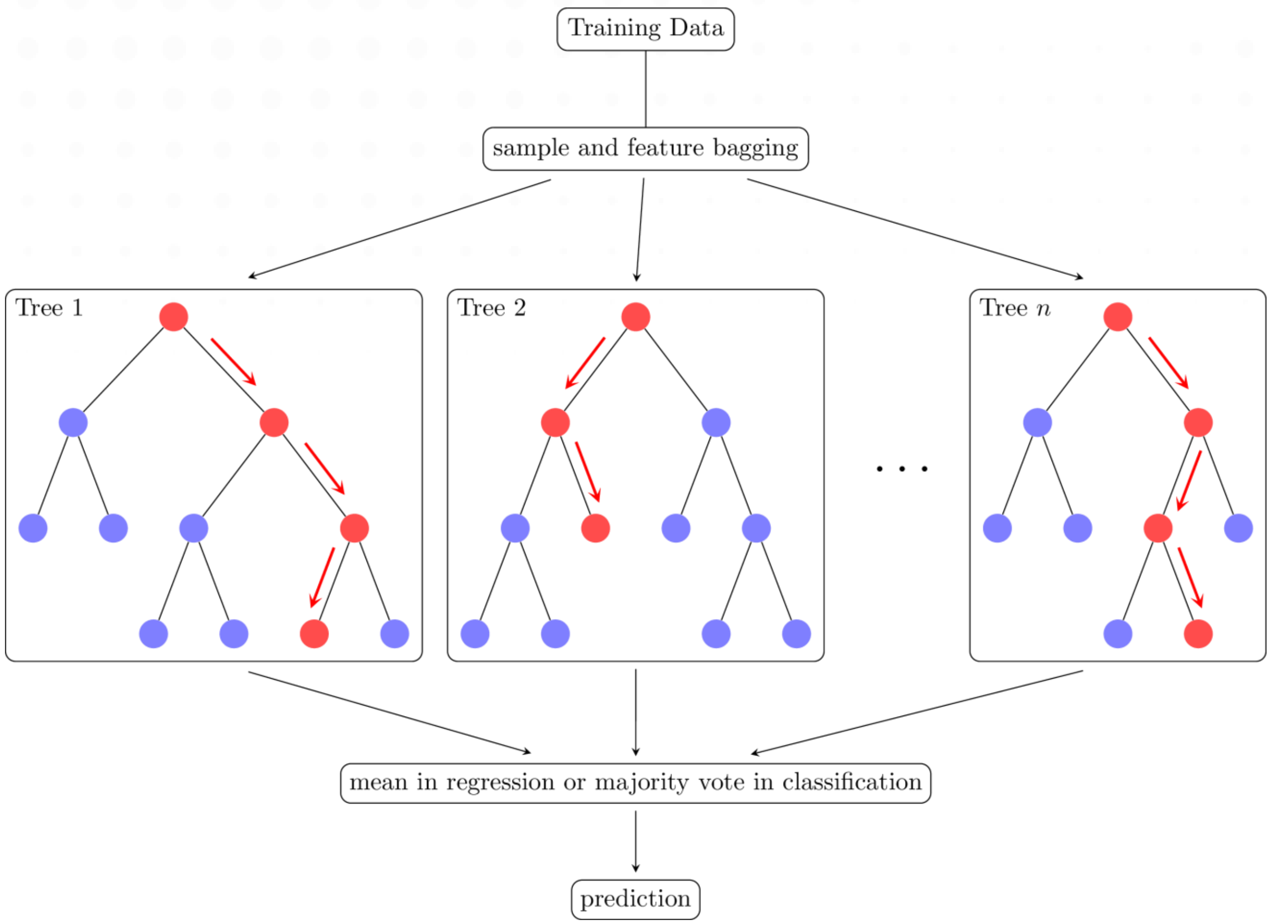
-Lower risk of overfitting and bias

-Lower overall variance

-Better generalization capability

Random Forest

- Random Forest is a supervised machine learning method
- Tree-based methods that can be applied to regression problems
- The individual decision tree is easy to interpret, however, it is not as accurate as other supervised learning approaches.
- Random forest **ensembles** multiple decision trees, to achieve higher forecasting accuracy at the cost of some interpretability.



Source: Riebesell, J. (2022). Random forest [Diagram]. TikZ.net. https://tikz.net/random-forest/

Random Forest in python

01

```python
from sklearn.ensemble import RandomForestRegressor
from sklearn import metrics
```

02

```python
# Split train and test
train = df.iloc[:-int(len(df) * 0.2)]
test = df.iloc[-int(len(df) * 0.2):]
```

03

```python
# dropping health expenditure for X and use total health expenditure for y
X_train = train.drop(['Total Health Expenditure (TEH)','Domestic General Government Health Expenditure (GGHE-D)',
                      'Out-of-pocket Health Expenditure(OOP)'],axis =1)
y_train = train['Total Health Expenditure (TEH)']

# dropping health expenditure for X and use total health expenditure for y
X_test = test.drop(['Total Health Expenditure (TEH)','Domestic General Government Health Expenditure (GGHE-D)',
                    'Out-of-pocket Health Expenditure(OOP)'],axis =1)
y_test = test['Total Health Expenditure (TEH)']
```

04

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

05

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

ARIMA

- ARIMA model is a time series predicting model that can be broken down into three parts: autoregressive (AR), integrated (I), and moving average (MA).
- ARIMA is represented as ARIMA (p, d, q) model, where p is the order of the autoregressive component, d is the degree of differencing involved, and q is the order of the moving average part, which corresponds to the three components above.
- The Autoregressive (AR) part of the ARIMA model represents a combination of past data points to forecast future values.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

- The integrated (I) part aims to turn the time series stationary by performing differencing to eliminate trend and seasonality. Augmented Dickey-Fuller test is use to determine the requirement for differencing.

ARIMA

- The moving average (MA) part focuses on the relationship between observations and the residual errors.
- It predicts using past forecast errors in a regression.
- MA model can capture meaningful short-term changes and remove random noise from the time series.
- It is combined with AR to improve attention for recent incidents than the pure AR process

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

- The integrated part is done so that non-stationary time series can be used for ARIMA process because both AR and MA assume stationarity.

ARIMA in python

UTM UNIVERSITI TEKNOLOGI MALAYSIA

01

```
# Import packages required
from statsmodels.tsa.stattools import adfuller
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from statsmodels.tsa.arima.model import ARIMA
```

02

```
result = adfuller(df['Total Health Expenditure (TEH)'])

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

03

```
# ACF plot
plot_acf(df['TEH_diff1'].dropna())
# PACF plot
plot_pacf(df['TEH_diff1'].dropna())

plt.show()
```

04

```
# train test split
train = df.iloc[:-int(len(df) * 0.2)]
test = df.iloc[-int(len(df) * 0.2):]


#fitting the time series data to the ARIMA model
model = ARIMA(train['TEH_diff1'], order=(0, 1, 0)).fit()
print(model.summary())
```

05

```
# forecast result
forecasts = model.forecast(len(test))
forecasts
```

Evaluation Metrics

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

01 Mean Absolute Error

- Calculate mean of the errors by their absolute value
- Simple metric for interpretation.

02 Root Mean Square Error

- Measures the average magnitude of prediction error
- Larger prediction errors will be penalised more heavily
- The metric is square-rooted, easier to compare to other metrics

03 Coefficient of Determination

- Indicates the goodness of fit of a model.
- R^2 score = 1 indicates all the actual value lies perfectly on the prediction model
- $R^2 = 0$ indicates the model does not fit any actual value

Initial Findings

Pre-processed Dataset

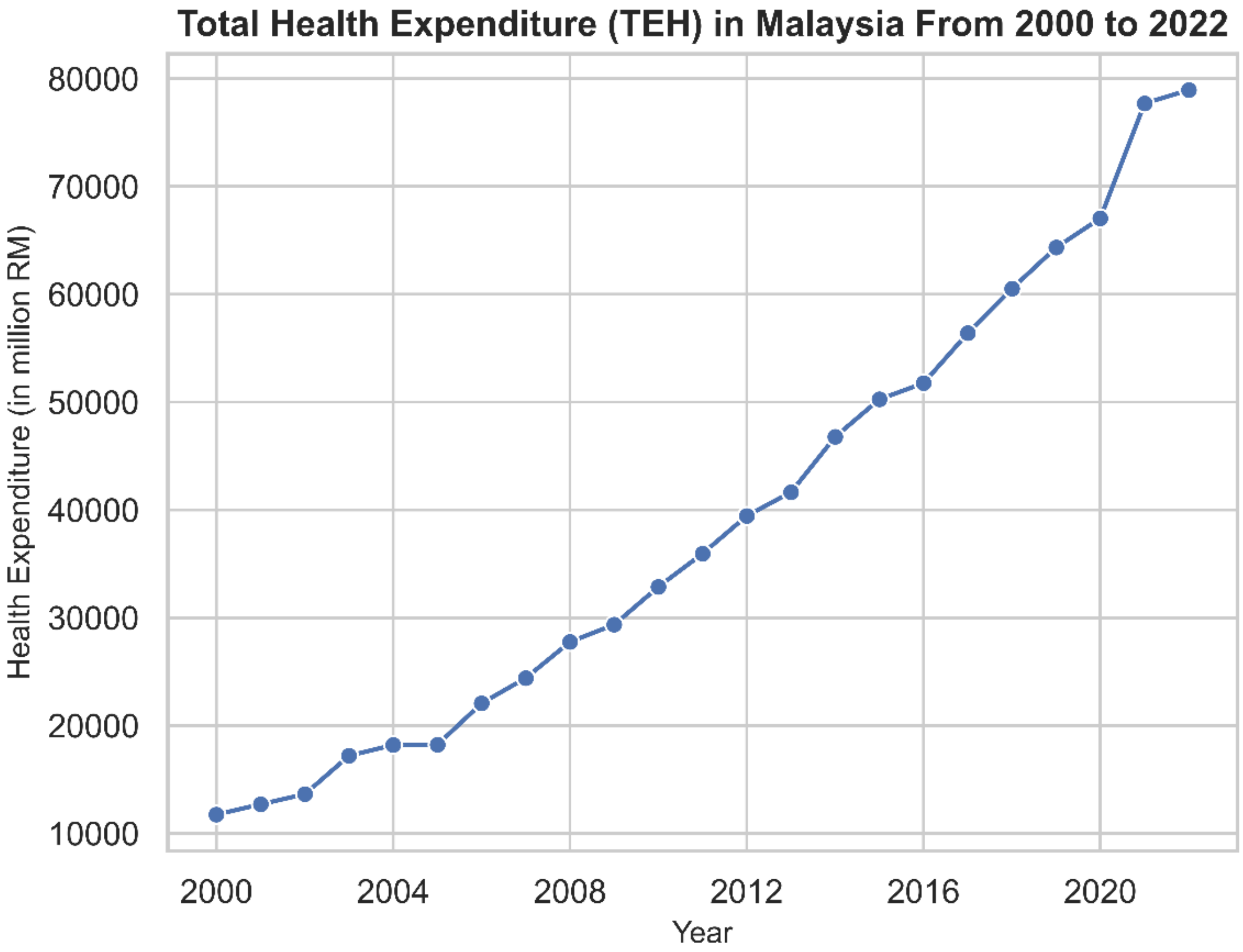
UTM UNIVERSITI TEKNOLOGI MALAYSIA

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

Year	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)
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							
2002-01-01	13640	5704.470433	417367	24102.4765							
2003-01-01	17203	6927.368331	456095	24679.6020							
2004-01-01	18200	7521.882374	516302	25256.7725							
2005-01-01	18231	7759.413210	569371	25836.0715							
2006-01-01	22072	10469.676324	625100	26417.9090							
2007-01-01	24414	11323.238597	696910	26998.3885							
2008-01-01	27758	12881.971119	806480	27570.0590							
2009-01-01	29365	13527.291955	746679	28124.7775							

```
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
```

Exploratory Data Analysis

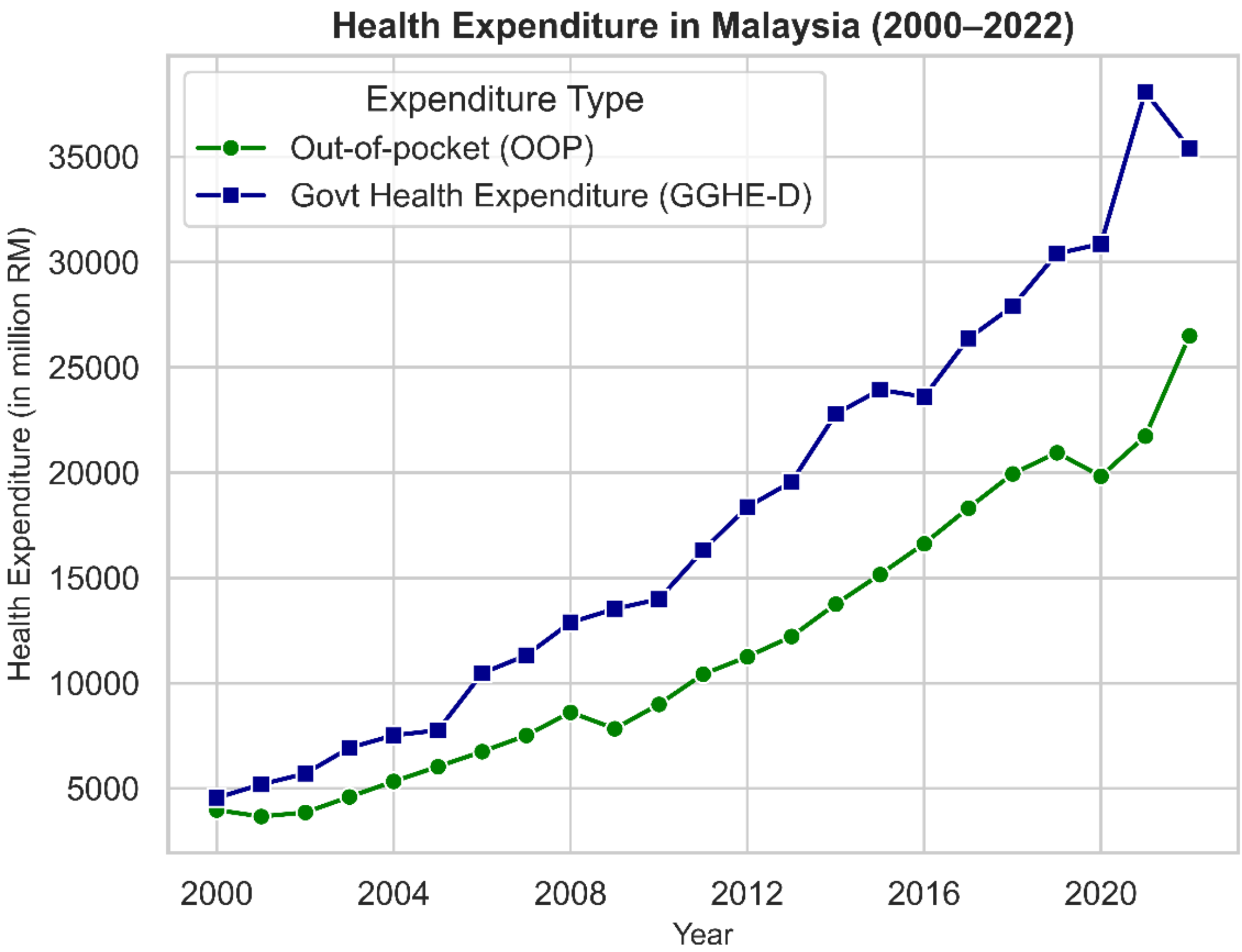


- The line chart shows that there is a gradual increase in health expenditure over 23 years, except that there is a steep increase from 2020 to 2021 (RM 67 million to RM 77 million).

- 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)

- Strong positive trend observed from the linechart, the time series is not stationary. Therefore, differencing has to be applied to stabilise the mean when carrying out ARIMA modelling

Exploratory Data Analysis



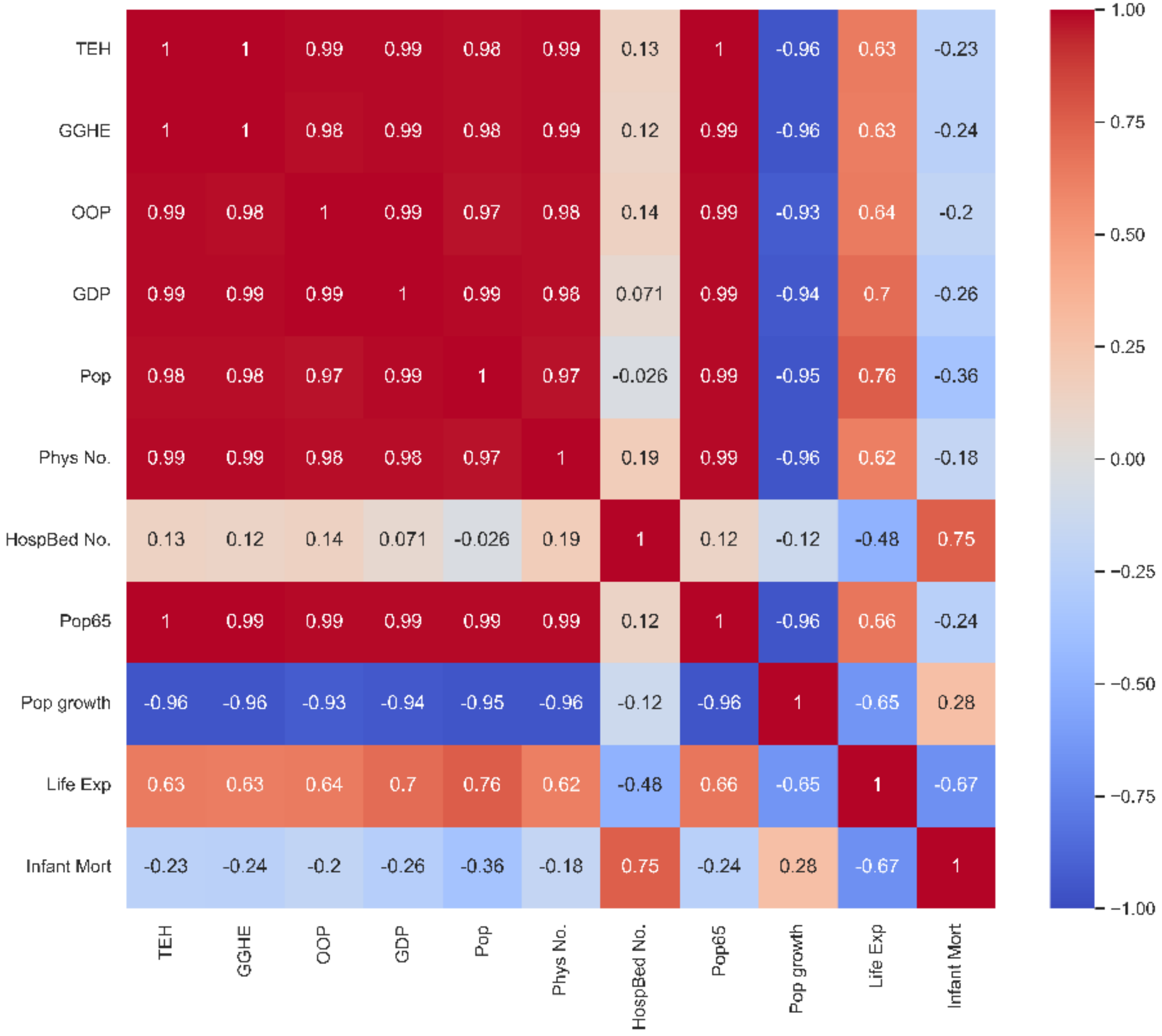
- Domestic General Government Health Expenditure shows a steeper upward trend versus Out-of-pocket Health Expenditure, despite both beginning at a similar starting point at 2000
- There is a steady growth in both expenditure types from 2000 to 2018
- GGHE-D shows a sharp rise from 2020 to 2021, acting 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 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).

Correlation Heatmap

- Strong positive correlation between total health expenditure (TEH), domestic general government health expenditure (GGHE-D) and out-of-pocket health expenditure (OOP)

Against all three of the health expenditures:

- **Strong Positive** Correlation**:** Gross domestic product (GDP), population in thousands (Pop), total population aged 65 years old (Pop 65) and number of physicians per 1000 people
- **Moderate positive** correlation: Life expectancy at birth (Life Exp)
- **Strong negative** correlation: Population growth in annual % (Pop growth)
- No of hospital beds has a **weak positive** correlation
- **Weak negative** correlation for infant mortality rate



Feature Selection

- Number of hospital beds and infant mortality rate **weakly correlate** with health expenditures
- These two columns are not selected as the features in the machine learning models for the prediction of health expenditure
- To ensure the accuracy of the prediction
- Features reduction can reduce complexity of the model and reduce computational and time resources

Other features are chosen as the predictive indicators for health expenditures, supported by the literatures and exploratory data analysis done on these features.

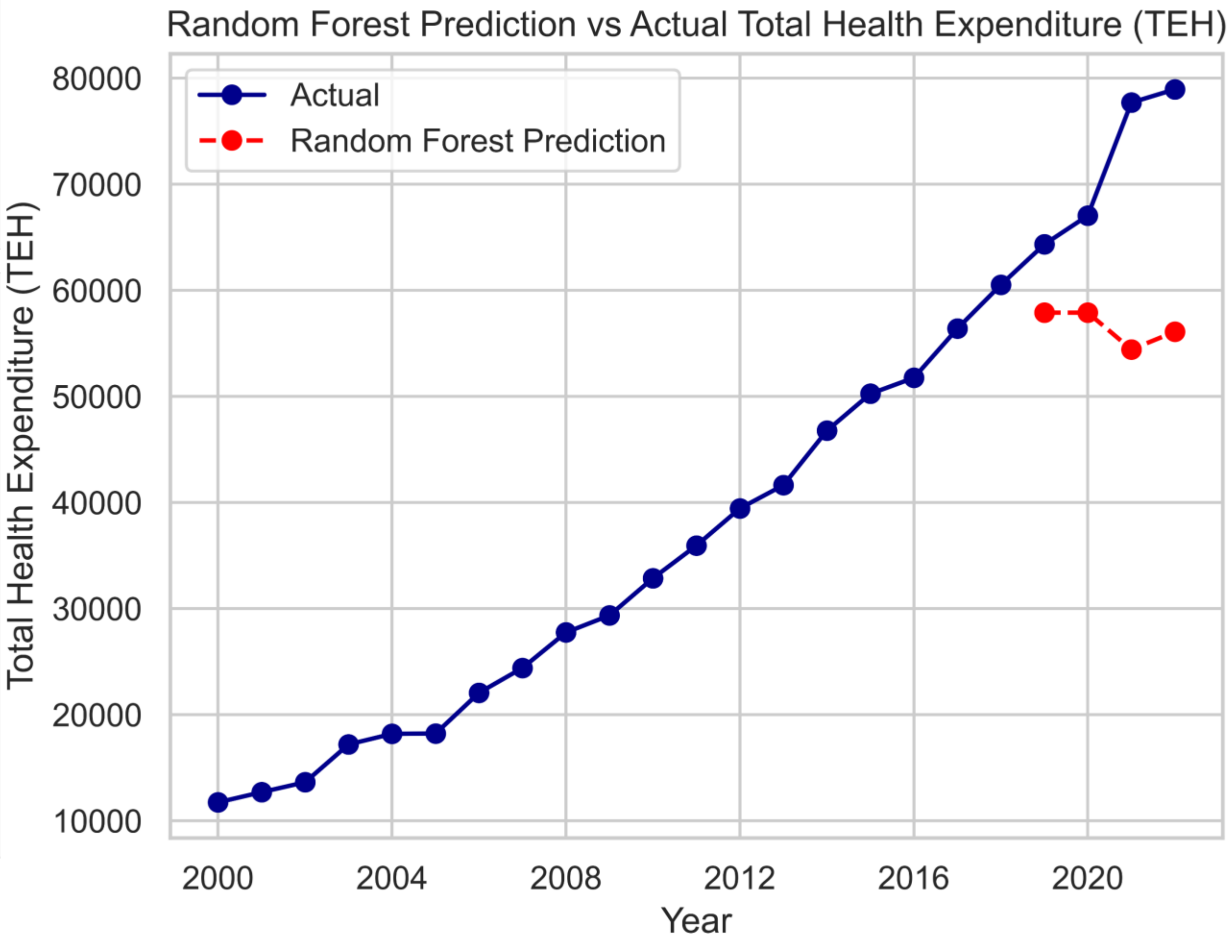
```
# drop number of hospital bed and infant mortality due to weak correlation
df = df.drop(['Hospital beds (per 1,000 people)','Mortality rate, infant (per 1,000 live births)'], axis= 1 )
```

Random Forest

```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)
```

```python
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score

# 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
```

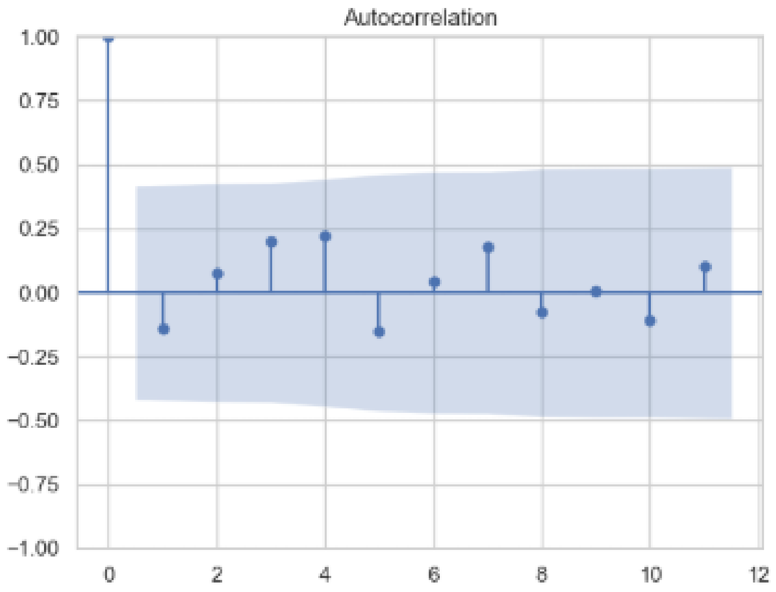


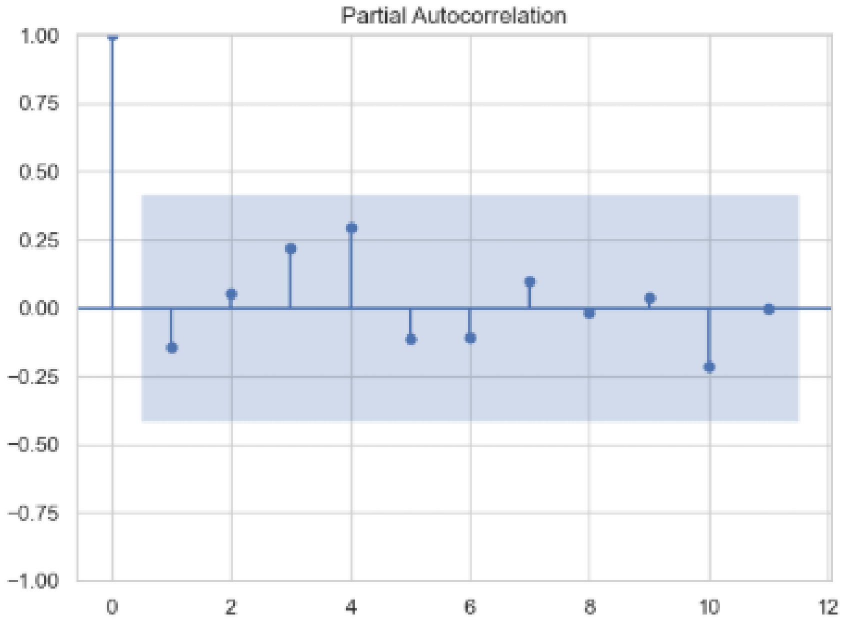
ARIMA

```python
# 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
```

- ADF test result show differencing needs to be done. The time series data is differenced once and saved as 'TEH_diff1'.
- From the ACF and PACF plot result it can be seen that the cut-off point is at 0. Therefore, p and q are set as 0 for the ARIMA model.





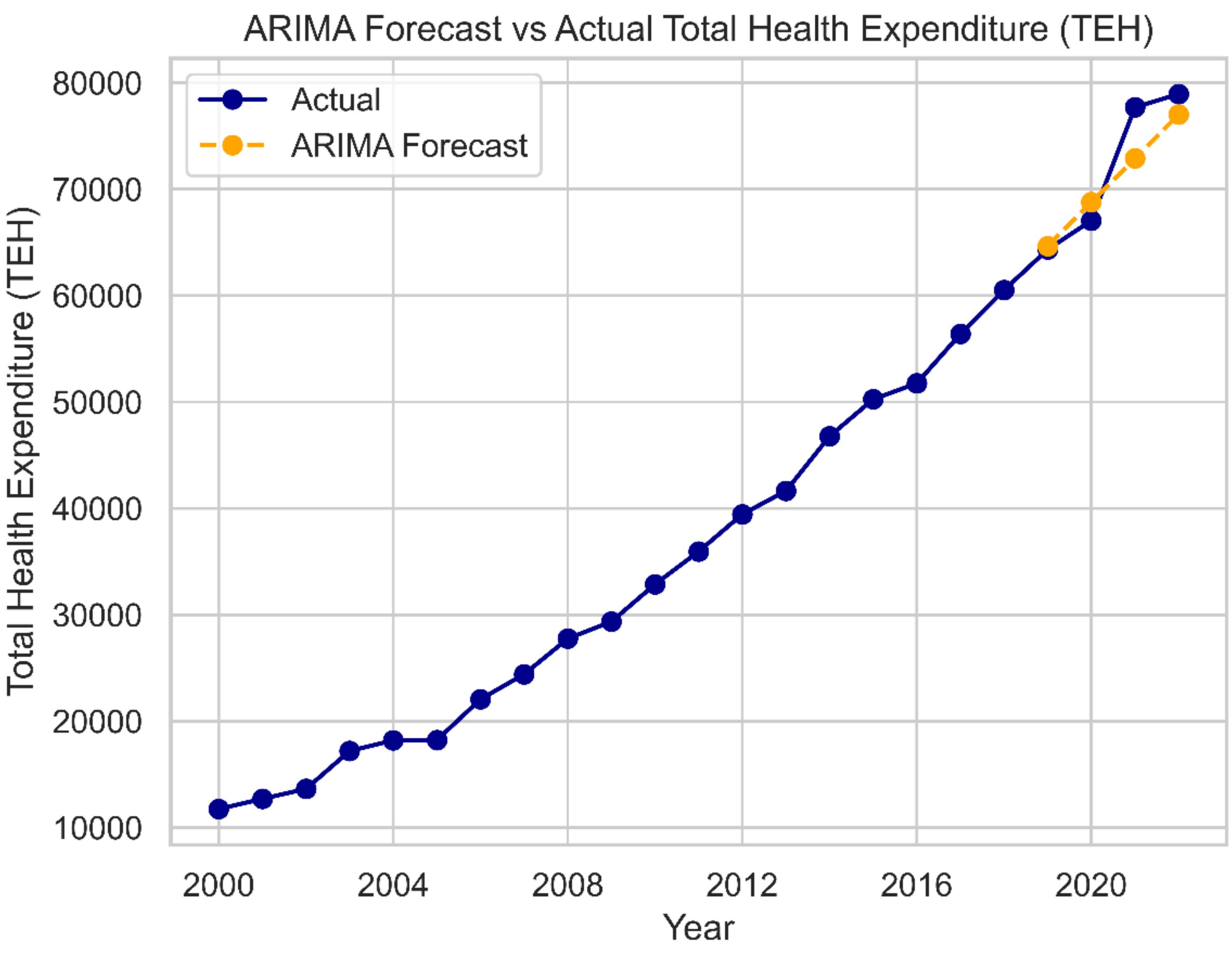
ARIMA

```python
from statsmodels.tsa.arima.model import ARIMA
# train test split
train = df.iloc[:-int(len(df) * 0.2)]
test = df.iloc[-int(len(df) * 0.2):]


#fitting the time series data to the ARIMA model
model = ARIMA(train['TEH_diff1'], order=(0, 1, 0)).fit()
print(model.summary())

# forecast result
forecasts = model.forecast(len(test))
forecasts
```

```python
# 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)
```

```python
# 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
```



Comparison of Results

Models	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	Coefficient of Determination (R_2)
Random Forest	15314	16951	-6
ARIMA	2191	2731	0.818

- From the initial findings, it can be concluded that ARIMA outperform Random Forest in forecasting Malaysia's Total Health Expenditure (TEH) from 2019 to 2022, achieving a low MAE of RM 2,191 million, a low RMSE of RM 2,731 million and a high R^2 of 0.818, indicating strong predictive accuracy.
- 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.

Conclusion & Future Works

Conclusion

- **ARIMA outperform Random Forest** in forecasting Malaysia's Total Health Expenditure (TEH) from 2019 to 2022, achieving a low MAE of RM 2,191 million, a low RMSE of RM 2,731 million and a high R^2 of 0.818.
- In contrast, Random Forest performance is poor due to the **absence of lagged values, insufficient training data and lack of hyperparameter tuning.**
- The results conclude that a time-series model like ARIMA is suitable for health expenditure forecasting when **small datasets** are provided
- The results also conclude that a complex model like Random Forest **requires additional data and further optimisation** to perform effectively in the forecasting task.
- For the master's project, hyperparameter tuning and validation should be prioritised to improve the accuracy and reliability of the models before forecasting Malaysia's health expenditure up to 2035.

Future Works

ASEAN Countries

The methodology of this study can be extended to ASEAN countries with similar health economic structures, for instance Thailand, Indonesia and Philippines.

Smaller components of healthcare expenses

Prediction can be applied smaller components in the healthcare spending, for instance, outpatient and inpatient services, pharmaceutical expenditures, education and training

Individual healthcare cost prediction

Individual healthcare cost prediction using patient's factors like patients' age, gender, medical conditions, current medications, income level and family history of illness.

References

Ceylan, Z., & Atalan, A. (2021). Estimation of healthcare expenditure per capita of Turkey using artificial intelligence techniques with genetic algorithm-based feature selection. Journal of Forecasting, 40(2), 279–290. https://doi.org/10.1002/for.2747

Healthcare Asia. (2024, April 16). *Malaysia's health expenditure to grow at 8.3% CAGR until 2028* [Screenshot]. Healthcare Asia. https://healthcareasiamagazine.com/healthcare/in-focus/malaysias-health-expenditure-grow-83-cagr-until-2028

Hassandarvish, M. (2024, December 10). *Malaysia's soaring medical inflation: How rising costs are straining public and private healthcare* [Screenshot]. Malay Mail. https://www.malaymail.com/news/malaysia/2024/12/10/7am-omla-malaysias-soaring-medical-inflation-how-rising-costs-are- straining-public-and-private-healthcare/159355

Jia, H., Jiang, H., Yu, J., Zhang, J., Cao, P., & Yu, X. (2021). Total Health Expenditure and Its Driving Factors in China: A Gray Theory Analysis. *Healthcare*, *9*(2), 207. https://doi.org/10.3390/healthcare9020207

Jakovljevic, M., Lamnisos, D., Westerman, R., Chattu V. K. & Cerda A. (2022). Future health spending forecast in leading emerging BRICS markets in 2030: health policy implications. Health Res Policy Sys 20, 23. https://doi.org/10.1186/s12961-022-00822-5

Kazemian, M., Abdi, Z., Meskarpour-Amiri, M. (2022) Forecasting Iran national health expenditures: General model and conceptual framework. Journal of Education and Health Promotion 11(1):p 87, | DOI: 10.4103/jehp.jehp_362_21

Klazoglou, P. & Dritsakis, N. (2018). Modeling and Forecasting of US Health Expenditures Using ARIMA Models. 10.1007/978-3-319-70055-7_36.

Ku Abd Rahim, K. N., Kamaruzaman, H. F., Dahlui, M., & Wan Puteh, S. E. (2020). From Evidence to Policy: Economic Evaluations of Healthcare in Malaysia: A Systematic Review. Value in health regional issues, 21, 91–99. https://doi.org/10.1016/j.vhri.2019.09.002

Li, H.Y., & Zhang, R.X. (2024). Analysis of the structure and trend prediction of China's total health expenditure. Frontiers in Public Health, 12, 1425716. https://doi.org/10.3389/fpubh.2024.1425716

Ministry of Finance Malaysia. (2024, October 11). *The 2025 Budget speech* [PDF]. Ministry of Finance Malaysia. https://belanjawan.mof.gov.my/pdf/belanjawan2025/ucapan/ub25-en.pdf

References

Ministry of Health Malaysia (2019). Malaysia National Health Accounts: Health Expenditure Report 1997–2017. Ministry of Health Malaysia. Retrieved from https://www.moh.gov.my/moh/resources/Penerbitan/Penerbitan%20Utama/MNHA/Laporan_MNHA_Health_Expenditure_Report_1997-2017_03122019.pdf

Ministry of Health Malaysia. (2024). *Malaysia National Health Accounts (MNHA) 2011–2022* [PDF]. Ministry of Health Malaysia. https://www.moh.gov.my/moh/resources/Penerbitan/Penerbitan%20Utama/MNHA/MNHA_2011-2022.pdf

Odnoletkova, I., Chalon, P. X., Devriese, S., & Cleemput, I. (2025). Projections of public spending on pharmaceuticals: A review of methods. *PharmacoEconomics, 43*, 375–388. https://doi.org/10.1007/s40273-024-01465-w

Özcan, T., Tüysüz, F. (2018). Healthcare Expenditure Prediction in Turkey by Using Genetic Algorithm Based Grey Forecasting Models. In: Kahraman, C., Topcu, Y. (eds) Operations Research Applications in Health Care Management. International Series in Operations Research & Management Science, vol 262. Springer, Cham. https://doi-org.ezproxy.utm.my/10.1007/978-3-319-65455-3_7

Parzi, M. N. (2024, October 16). *Growing medical costs present major challenges for healthcare sector* [Screenshot]. New Straits Times. https://www.nst.com.my/news/nation/2024/10/1120615/growing-medical-costs-present-major-challenges-healthcare-sector

Riebesell, J. (2022). *Random forest* [Diagram]. TikZ.net. https://tikz.net/random-forest/

Sahoo, P.M., Rout, H.S. & Jakovljevic, M. (2023). Future health expenditure in the BRICS countries: a forecasting analysis for 2035. *Global Health* **19**, 49. https://doi.org/10.1186/s12992-023-00947-4

Saleh, M. H., Alkhawaldeh, R. S., & Jaber, J. J. (2023). A predictive modeling for health expenditure using neural networks strategies. *Journal of Open Innovation: Technology, Market, and Complexity, 9*(3), 100132. https://doi.org/10.1016/j.joitmc.2023.100132

Wang, J., Qin, Z., Hsu, J., & Zhou, B. (2024). A fusion of machine learning algorithms and traditional statistical forecasting models for analyzing American healthcare expenditure. *Healthcare Analytics, 5*, 100312. https://doi.org/10.1016/j.health.2024.100312

World Health Organization. (2025). *Health expenditure*. World Health Organization. Retrieved April 17, 2025, from https://www.who.int/data/nutrition/nlis/info/health-expenditure

Zheng, A., Fang, Q., Zhu, Y., Jiang, C., Jin, F., & Wang, X. (2020). An application of ARIMA model for predicting total health expenditure in China from 1978-2022. Journal of Global Health. 10. 10.7189/jogh.10.010803.

THANK YOU

univteknologimalaysia utm.my utmofficial

VIDEO LINK : https://youtu.be/RfOSxvLuB18