PREDICTION OF HEALTH EXPENDITURE IN MALAYSIA USING MACHINE LEARNING

LOCK CHUN HERN

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF PROJECT REPORT

Author's full name	:	LOCK CHUN HERN			
Student's Matric No.	:	MCS241047	Academic Session	:	20242025-02
Date of Birth	:	27 MAY 1997	UTM Email	:	lock@gradute.utm.my
Project Report Title	:	PREDICTION OF HEALTHCARE EXPENDITURE IN MALAYSIA USING MACHINE LEARNING			

I declare that this project report is classified as:

☒ **OPEN ACCESS** — I agree that my report to be published as a hard copy or made available through online open access.

☐ **RESTRICTED** — Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☐ **CONFIDENTIAL** — Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the project report belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this project report for academic exchange.

Signature of Student:

Signature :

Full Name: LOCK CHUN HERN
Date : 27 JUNE 2025

Approved by Supervisor(s)

Signature of Supervisor I:	Signature of Supervisor II
Full Name of Supervisor I ASSOC. PROF. DR. MOHD SHAHIZAN BIN OTHMAN Date : 27 JUNE 2025	Full Name of Supervisor II Date :

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

"I hereby declare that I have read this project report and in my opinion this project report is sufficient in term of scope and quality for the award of the degree of Master in Data Science"

Signature	:	_____
Name of Supervisor I	:	ASSOC. PROF. DR. MOHD SHAHIZAN BIN OTHMAN
Date	:	27 JUNE 2025

Pengesahan Peperiksaan

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar **:**

Nama dan Alamat Pemeriksa Dalam **:**

Nama Penyelia Lain (jika ada) **:**

Disahkan oleh Timbalan Pendaftar di Fakulti:

Tandatangan :

Nama :

Tarikh :

PREDICTION OF HEALTH EXPENDITURE IN MALAYSIA USING MACHINE LEARNING

LOCK CHUN HERN

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master in Data Science

Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2025

DECLARATION

I declare that this project report entitled *"*Prediction of Health Expenditure In Malaysia Using Machine Learning*"* is the result of my own research except as cited in the references. The project report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	.....
Name	:	LOCK CHUN HERN
Date	:	27 JUNE 2025

ACKNOWLEDGEMENT

In this project report, I would like to express my sincere gratitude to Assoc. Prof. Dr Mohd Shahizan bin Othman for his guidance, advices and encouragement. He has shared a wide range of useful resources and provided constructive feedback for my project. Without his continued support, this project would not proceed smoothly.

I also like to thank Universiti Teknologi Malaysia (UTM) and the librarian for maintaining an up-to-date online database for the access to a plethora of online resources. The relevant literatures are extremely helpful in the writing of this project report.

I would also like to express my appreciation to my postgraduate peers and seniors for their assistance and critics. They have provided useful tips to me especially the formatting of this project report. Last but not least, I would like to thanks my family members of their unconditional support.

ABSTRACT

Rising healthcare expenditure is a global challenge. The escalation of healthcare cost, caused by the inflation in medication prices, medical expenses, and the ageing population in Malaysia, necessitates the development of a robust predictive model for health expenditure to aid in the planning of future healthcare budgets. From previous researches, it is suggested that machine learning algorithms have great potential when used in modern health economies. The aim of this research project is to predict health expenditure in Malaysia using machine learning techniques to provide insight for health financing and policy planning. Random Forest and ARIMA are selected and applied to predict health expenditure in Malaysia from 2026 to 2035. The healthcare spending data is sourced from Ministry of Health Malaysia, WHO Global Health Expenditure Database and World Development Indicators Database for the period of 2000 to 2022. Explanatory data analysis shows a gradual increase in health expenditures from 2000 to 2019, and some fluctuations between 2019 to 2022 due to the COVID-19 pandemic. Key determinants affecting health expenditure have been identified by investigating their correlations with the health expenditures. 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. 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. Random Forest performance is poor in the initial result due to insufficient training data and lack of hyperparameter tuning. In future studies, hyperparameter tuning and validation should be prioritised to improve the accuracy and reliability of the models. The methodology of this study can be extended to include other ASEAN countries with similar health economic structures to improve the generalisability of the machine learning models and provide insights for healthcare budget planning based on varying health policies across counties.

ABSTRAK

Kenaikan perbelanjaan kesihatan merupakan cabaran global. Kenaikan kos penjagaan kesihatan disebabkan inflasi kos ubat, kos perubatan dan penuaan populasi menujukkan keperluan untuk perkembangan model ramalan yang berkesan dan dapat membantu dalam perancangan bajet penjagaan kesihatan masa depan. Kajian-kajian sebelum ini telah menunjukkan bahawa algoritma pembelajaran mesin berpotensi tinggi untuk digunakan dalam bidang ekonomi kesihatan moden. Projek kajian ini bertujuan untuk meramalkan perbelanjaan kesihatan di Malaysia dengan menggunakan teknik pembelajaran mesin bagi menyediakan panduan untuk pembelajaan kesihatan dan perancangan dasar kesihatan. Algoritma Random Forest dan ARIMA telah dipilih untuk meramalkan perbelanjaan kesihatan di Malaysia bagi tempoh tahun 2026 hingga 2035. Data perbelanjaan penjagaan Kesihatan diperolehi daripada Kementerian Kesihatan Malaysia, WHO Global Health Expenditure Database, dan World Development Indicators Database bagi tempoh tahun 2000 hingga 2022. Analisis data menunjukkan peningkatan progesif dalam perbelanjaan kesihatan dari tahun 2000 hingga 2019, dan perubahan tidak menentu antara tahun 2019 hingga 2022 akibat pandemik COVID-19. Faktor-faktor utama yang mempengaruhi perbelanjaan kesihatan telah dikenal pasti melalui kajian korelasi. Antara indikator yang dipilih sebagai indikator ramalan termasuk Anggaran Keluaran Dalam Negeri Kasar (KDNK), populasi (dalam ribu), jumlah penduduk berumur 65 tahun ke atas, bilangan doktor untuk setiap 1000 penduduk, jangka hayat ketika lahir dan kadar pertumbuhan penduduk. Daripada hasil keputsan awal, ARIMA berpretasi lebih baik daripada Random Forest dalam peramalan Jumlah Perbelanjaan Kesihatan (TEH) Malaysia bagi tempoh 2019 hingga 2022, dengan keputusan MAE yang rendah iaitu RM2,191 juta, RMSE RM2,731 juta dan nilai R^2 setinggi 0.818, menunjukkan ketepatan yang tinggi. Prestasi Random Forest adalah lemah dalam hasil keputusan awal disebabkan oleh kekurangan data untuk meltihkan model dan kekurangan pelarasan hiperparameter. Untuk kajian masa depan, pelarasan hiperparameter dan pengesahan model perlu diutamakan untuk meningkatkan ketepatan dan kebolehpercayaan model-model ini. Kaedah yang digunakan dalam kajian ini juga boleh dilanjutkan ke negara-negara ASEAN lain yang mempunyai struktur ekonomi kesihatan yang serupa untuk meningkatkan kebolehgeneralisasian model pembelajaran mesin dan mendapatkan pandangan mendalam untuk perancangan bajet kesihatan berdasarkan dasar kesihatan yang berbeza antara negara.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	-	Artificial Intelligence
ACF	-	Autocorrelation Function
ANN	-	Artificial Neural Network
ANFIS	-	Adaptive Neuro-Fuzzy Inference Systems
ARIMA	-	Autoregressive Integrated Moving Average
ASEAN	-	Association of Southeast Asian Nations
BRICS	-	Brazil, Russia, India, China, and South Africa
CHE	-	Current Health Expenditure
COVID19	-	Coronavirus disease 2019
CV	-	Cross Validation
ETS	-	Exponential Smoothing Model
GM	-	Grey Model
GDP	-	Gross Domestic Product
GGHE	-	General Government Health Expenditure
GGHE-D	-	Domestic General Government Health Expenditure
MAE	-	Mean Absolute Error
MHNA	-	Malaysia National Health Account
NHA	-	National Health Account
ML	-	Machine Learning
MOH	-	Ministry of Health Malaysia
MYR	-	Malaysia Ringgit
OECD	-	Organisation for Economic Co-operation and Development
OOP	-	Out-Of-Pocket Health Expenditure
PACF	-	Partial Autocorrelation Function
R^2	-	Coefficient of Determination
RF	-	Random Forest
RMSE	-	Root Mean Squared Error
SVR	-	Support Vector Regression
TEH	-	Total Health Expenditure
UTM	-	Universiti Teknologi Malaysia

WHO - World Health Organization

LIST OF SYMBOLS

y_i	-	Actual value of the dependent variable at observation i
ϕ	-	Autoregressive coefficient
c	-	Constant
y	-	Dependent variable
ε	-	Error term
θ_1	-	Moving average coefficient
$\overline{y_i}$	-	Mean of actual values of the dependent variable at observation i
n	-	Number of observations
$\hat{y}_i$	-	Predicted value of the dependent variable at observation i

CHAPTER 1

INTRODUCTION

1.1 Overview

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 Organization [WHO], 2025a). Health funding sources from the public and private sectors include the government, individuals, private health insurance, and other non-government organizations. Health expenditure can be further classified into 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).

Machine learning approaches have been used to deepen understanding and provide insight into healthcare spending. For example, researchers from Jordan predicted total healthcare expenditure for their country using two neural network strategies: Adaptive Neuro-Fuzzy Inference System and Hybrid Neural Fuzzy Inference System (Saleh et.al., 2023). Support Vector Regression (SVR) and Random Forest (RF) are performed on American healthcare expenditure and enable the prediction of healthcare expenditure as a percentage of Gross Domestic Product (GDP) for 2050, and RF provides comparable results with Autoregressive Integrated Moving Average (ARIMA) model. (Wang et. al., 2024). These researches suggest that machine learning algorithms have great potential when used in modern health economies.

In this study, machine learning techniques will be applied to the health expenditure data from Malaysia to shed light on the future health expenditure. It is

expected to provide valuable insight for policy planning related to healthcare sector in Malaysia.

1.2 Problem Background

In Malaysia, total health expenditure has been increasing from 2011 to 2022, from RM 36.9 billion to RM 78.9 billion, and as % GDP from 3.94% to 4.41% (Ministry of Health Malaysia [MOH], 2024). Recently, the Malaysia government has allocated RM 45.3 billion in Budget 2025 for the Ministry of Health for spending on healthcare alone, which is the second highest after education (Ministry of Finance Malaysia, 2024). The growing spending in healthcare raises concerns as it may cause reduced allocation of the budget for other critical fields. It also means that people living in Malaysia are spending more on healthcare expenses, which might cause a burden to those living under poor economic conditions.

Rising healthcare expenditure is a global challenge. For example, the health spending by 38 countries participating in the Organisation for Economic Co-operation and Development (OECD) is predicted to peak at 11.8% of GDP in 2040. By that time, the increase in healthcare expenses from public sources is estimated to be twice the average growth in government revenues. (Organisation for Economic Co-operation and Development, 2024) With the inflation in medication prices, medical expenses, and the ageing population in Malaysia, it is anticipated that healthcare expenditure will continue to increase. However, limited academic research has been done to predict future health expenditures in Malaysia. Therefore, there is a need for an accurate predictive model to be developed to aid in the planning of future healthcare budgets.

1.3 Problem Statement

Determinants of healthcare expenditure are complex, and their application varies depending on the prediction models (micro, macro level, or component-based). These include demographic factors such as an ageing population, health-related factors

like prevalence of chronic disease, and economic factors, for instance GDP of the country. However, the appropriate determinants of healthcare expenditure have not been established. Identifying the key determinants to be used in this project is important for precise healthcare expenditure prediction.

Furthermore, while most analyses of healthcare expenditure are done in traditional models, they present a trade-off in terms of accuracy and the time frame of prediction. Accurate prediction for each component of health expenditure is required for informed decision-making. Machine learning approaches such as Random Forest and ARIMA can provide new insights into this issue by proposing a better model in estimating future health expenditure, which helps in decision making for policymakers.

In addition, the performance of different machine learning algorithms is inconsistent in existing studies when used to predict health expenditure. This is because machine learning algorithms perform differently depending on the context of the input data. A thorough comparison of performance by RF and ARIMA on the local data is required to determine the best model in predicting the health expenditure of Malaysia.

1.4 Research Question

a) What are the key determinants of health expenditure that contribute to accurate prediction in machine learning algorithms?

b) What is the predicted health expenditure of Malaysia from 2026 to 2035 using machine learning algorithms (Random Forest and ARIMA)?

c) How do different machine learning models perform on Malaysia's health expenditure data, and which model demonstrates the highest accuracy?

1.5 Research Aim

This research aims to predict health expenditure in Malaysia using machine learning techniques to provide insight for health financing and policy planning.

1.6 Research Objectives

The objectives of this research are:

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

1.7 Research Scopes

a) The data will be sourced from the Ministry of Health Malaysia, Department of Statistics Malaysia, World Bank Group, and World Health Organization.

b) The data collected will only involve demographic data and data related to health economics. No individual data that reveals an individual's medical and medication history will be used.

c) The study will use data from 2000 to 2022, providing relevant and up-to-date data for health expenses forecasting

d) This research will apply two machine learning methods: Random Forest and ARIMA.

1.8 Expected Contribution

This project can provide insights for policymakers in the country in planning health expenditures and allow strategic allocation of budgets for health expenses. This helps to ensure the long-term sustainability of funding for Malaysia's healthcare system. Overall, this research is projected to contribute to better health outcomes for the patients and people in Malaysia. The findings of this project are also expected to provide insights for other countries with similar healthcare systems or income levels.

1.9 Thesis Organisation

The following chapters are presented as outlined below:

Chapter 2 covers extensive literature reviews regarding health expenditure. This chapter explores methodology and findings on the health expenditure forecasting from existing research and identifies research gaps.

Chapter 3 dives deep into research methodology. This chapter discusses data collection and data pre-processing steps. This chapter also includes details about proposed steps for exploratory data analysis, feature engineering and selection of machine learning algorithms.

Chapter 4 covers initial findings from this project. Exploratory data analysis is conducted to gain insight from the dataset collected. Feature engineering is conducted to select the appropriate determinants of health expenditure. Initial results from the application of machine learning models, RF and ARIMA, on the total health expenditure of Malaysia are evaluated, compared and discussed in detail.

Chapter 5 ends with a conclusion about this project. Directions for future work will be proposed to extend the research outcomes.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter provides an overview of the significance of health expenditure prediction, machine learning in healthcare and health economics, existing models used for health expenditure predictions, determinants of health expenditure, research gap, and summary. This chapter aims to review existing literature on this study, evaluate the approaches to tackle the issue, strengths and limitations of each model, and analyse the research gap that has not been resolved in the current studies.

2.2 Health Expenditure and Prediction

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). (World Health Organization [WHO], 2025a). In 2018, global spending on health achieved USD 8.3 trillion, which is 10% of global GDP. On average, low-income countries spent 6.4% GDP on health, lower-middle income countries spent 4.8%, while upper-middle income countries spent 6.3%. (WHO, 2020).

Health financing in Malaysia is largely funded by public funding (RM 41,257 million), which consists of 52.3% of total health expenditure, followed by private sources of financing, accounting for RM 37,688 million (47.7%). The TEH of Malaysia is gradually increasing from 2011 to 2022, showing more than a 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). Despite that, public spending on health 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).

Household OOP includes health spending by people for for healthcare service in private, public hospitals and clinics, purchase of pharmaceuticals (over-the-counter and prescription drugs), education and training, medical equipment consumable goods, daycare and alternative medicine. (Ministry of Health Malaysia, 2024). Catastrophic health expenditure, which is defined as OOP exceeding 10% total household consumption, is on a trend of rising based on a study by Sayuti & Sukeri (2022), from an estimated 1.44% in the household expenditure survey 2004/2005 to 2.8% in 2015/2016. The authors conclude that vulnerable groups' access to healthcare services should be improved to avoid the vicious cycle of debt due to rising healthcare costs. The conclusion is supported by the result of a study by Wan Puteh et al. (2023) on the cancer population in Malaysia, which revealed that 54.4% of respondents in their study experienced catastrophic health expenditure. This signifies the importance of well-planned healthcare budget allocation by policymakers.

Prediction models are highly beneficial and practical for determining the sectors where the spending is growing and reveal the driving factors for the health expenditure. Short-term predictions have better accuracy in forecasting future events, but limited action can be taken to change the prediction. Medium to long-term estimation excels in its ability for policy-planning and decision making because they can identify future trends where policy makers can shift the outcomes trajectory (Astolfi et al., 2012).

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.

There are challenges in collecting cost data and the availability of data that need to be addressed, and the authors suggested that data modelling studies can address this issue.

Various attempts have been proposed to improve health expenditure allocation. The prediction of how total health expenditure and its sub-components change contributes significantly to evidence-based policy making. While an increase in health expenditure may translate into better health outcomes for the people, it is crucial for policymakers to closely monitor and address issues of health spending to avoid a significant burden on public resources while ensuring individuals' access to affordable healthcare services.

2.3 Machine learning in healthcare and health economics

Machine learning (ML) is a subset of artificial intelligence (AI), applies computational algorithms to construct and fit statistical models using available real-world data. It aims to estimate outcomes or assign categories to the input data provided based on training (Rubinger et al., 2023). The application of AI and ML in healthcare research is rapidly expanding, from utilising data from electronic medical records for clinical decision-making to estimating the cost of treatment using patient-level factors, owing to the increased accessibility of big health data.

Lee et al. (2022) conducted a systematic literature review on the use of ML in health economics and outcomes research (HEOR). The authors identified that machine learning is useful in predicting disease occurrence, clinical outcomes, health resource consumption and costs. The study concludes that tree-based methods are the most commonly used machine learning techniques, followed by logistic and linear regression, SVR, and neural networks.

However, the applications of AI and ML in healthcare do encounter ethical challenges and privacy-related issues, like patient data protection. Moreover, the lack of high-quality data for training and evaluations may cause performance issues in forecasting and may introduce bias (Wubineh et al., 2024). Therefore, healthcare data

must be handled carefully by complying with the legal and regulatory framework while ensuring the anonymity and security of the data. In addition, data quality must be assessed before applying any machine learning model to ensure its accuracy and reliability.

Nonetheless, there are lots of opportunities in the application of data science techniques, including machine learning in healthcare, for instance, personalised medicines, real-time monitoring, disease prevention, and diagnosis. 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).

2.4 Existing Models used for Health Expenditure Prediction

This subsection compares existing models used for health expenditure projection, which include different machine learning methods and traditional statistical models. Machine learning algorithms included in this subsection are Autoregressive Integrated Moving Average (ARIMA), Artificial Neural Network, and Random Forest (RF). Traditional statistical approaches, Grey Model (GM), and Exponential Smoothing Model (ETS) are included for comparison.

2.4.1 ARIMA

ARIMA model is a time series prediction technique. It is a model that is formed by: autoregressive (AR), integrated (I), and moving average (MA). The Autoregressive (AR) is a regression model that uses past data points (lagged observations) to forecast future values. The integrated (I) part aims to make the time series stationary by performing differencing to eliminate trend and seasonality. Moving average (MA) part

focuses on the dependence between observations and the residual errors. It captures meaningful short-term changes and removes random noise from the time series.

The use of ARIMA model for forecasting health expenditure 5 years forward has been done in China by Zheng et al. (2020), which predicted not only total health expenditure but also its constituent ratios of GGHE, social health expenditure (SHE), and OOP by using time series data from 1978 to 2017. Social health expenditure refers to the basic medical insurance fund collected by various social medical insurance projects. SHE is expected to grow the fastest in China, which will decrease the proportion of GGHE and OOP.

A similar approach was done using ARIMA for the Brazil, Russia, India, China, and South Africa (BRICS) countries to forecast until 2030. The authors projected total health expenses per capita and as % GDP for these countries. Estimation of government, prepaid private, and OOP has been done from 2018 to 2030 (Jakovljevic et al., 2022). However, the authors reported that despite ARIMA providing useful forecasts as a time series model, the results have high uncertainty in values due to a large prediction interval, data quality issues like measurement error and imputation, lack of an obvious pattern and trend in some of the datasets.

Kontopoulou et al. (2023) have reviewed ARIMA versus ML approaches and deep learning methods for time series forecasting in financial, healthcare, and various other sectors. The authors outline that ARIMA has advantages like being more explainable, flexible, and reliable, performs better for limited data or short-term prediction, has low time complexity, and smaller computational requirements. However, ARIMA do come with limitations like difficulty in predicting complex real-world problems due to its univariate modelling approach and is more sensitive to outliers.

2.4.2 Artificial Neural Network

Artificial neural network (ANN) is a ML model that works like networks of biological neurons in the brain. Artificial neurons are arranged in layers and build up a neural network to mimic the human brain. ANN is formed by input layers, output layers, and one or multiple hidden layers in between. Deep learning algorithms refer to the use of neural networks formed by dozens to hundreds of layers to process large and highly complex tasks, for instance, classifying billions of images, powering recommendation systems for e-commerce, and developing strategies in chess and games. (Géron, 2022). The algorithms can perform feature engineering automatically, where the features in the dataset are searched and correlated without human intervention. (Ahmed et al, 2023)

ANFIS and HyFIS are implemented to predict the health spending in Jordan. Both models demonstrated their capability in predicting total health expenditure accurately, with HyFIS outperformed ANFIS. (Saleh et al., 2023) Notably, five variables used in the neural network in this study demonstrate the capability of the neural network to obtain accurate prediction results based on multiple inputs.

Ahmed et al. (2023) suggested that training a deep learning model is time-consuming, computationally expensive, and requires large samples or training data to achieve better accuracy. It also demanded improved optimisation of the parameters to create a more robust model. Deep learning techniques are regarded as “black boxes” because their interpretation is difficult. Their architectures are highly complicated, and the decision-making process is not made clear to the user.

2.4.3 Random Forest

Random Forest (RF) is a ML algorithm that ensembles the output of a combination of decision trees to obtain a prediction outcome for both classification and regression. Each decision tree starts with a root node that branches into several decision nodes, which are made up of a set of questions. The decisions made at those

nodes based on input data will lead to the leaf nodes, which are the terminal node that shows predictions.

Random Forest utilises bootstrap aggregating and feature randomness to create an uncorrelated forest of decision trees. Bootstrap aggregating refers to random sampling with replacement of the original dataset to train each decision tree. Feature randomness means a random subset of features is selected at the split of each tree. This can reduce overfitting, bias, and variance, which provides improved accuracy compared to a decision tree used alone. (IBM, n.d.)

Wang et.al. (2024) compared RF, Support Vector Regression (SVR), and ARIMA in forecasting US healthcare expenditure as % GDP to 2050. The authors revealed that RF and ARIMA yield comparable results (18.8% and 17.9% respectively) when trained with time series data, while SVR struggles to learn effectively with limited data. The authors proposed that healthcare expenditure is affected by multiple factors, therefore, further investigation into the interaction among the factors (e.g. demographics, technology, and political landscapes) can provide a more comprehensive analysis.

Muremyi et. al (2020) study on OOP health expenditure in Rwanda using a tree model using 14 independent variables from household conditions at the micro-level. The authors suggested that while the machine learning approach is criticised compared to traditional statistics due to model assumptions, they do offer better generalisation capability and are effective for selecting predictable features from the datasets.

Another study compared five AI models (RF, ANN, Multiple linear regression (MLR), SVR, Relevance Vector Machine (RVM) to predict healthcare expenditure per capita in Turkey and concludes that the combination of genetic-algorithm feature selection and random forest demonstrated the best prediction performance in the study. (Ceylan & Atalan, 2020). The authors suggested that RF predicts better than single base learners and is less susceptible to overfitting than other ML algorithms. It also works well with large datasets with many input variables and missing values, which contributed to higher performance.

2.4.4 Other Statistical Models

Among the research of health expenditure prediction, various other traditional statistical model other than machine learning were used, which can be contrasted with machine learning for their advantages and limitations. The common approaches from recent literature are the grey model and the exponential smoothing model.

Grey mode (GM) is a time series forecasting equation that predicts using a combination of known and uncertain, previous and current data. GM is often used in economic analysis. The model is represented as GM (M, N), where M is the derivative order and N is the number of independent variables. Li & Zhang (2024) used the univariate GM (1,1) model to predict the trend of total health expenses and the share in GDP in China, which is expected to increase continuously and reach 8.89% of GDP by 2030. A small amount of data is needed for this model, and there is a low data distribution requirement. The study limitation is that the model used is univariate; therefore, other factors affecting the prediction are not considered.

Jia et al. (2021) performed a New Structure of the Multivariate Gray Prediction Model NSGM (1, N) to predict health expenditure in China and compared it against the traditional GM and neural network for evaluation. The authors evaluated 9 driving factors of health spending using grey correlation achieved better prediction accuracy than other models with limited data. However, the authors noted a limitation where the prediction performance of the model might be influenced by the correlation between the variables used.

Sahoo et al. (2023) conducted time series analysis and an exponential smoothing model (ETS) on BRICS countries to predict health expenditure and its components to 2035. ETS forecast by placing more weight on the recent outcomes, and the weights of past observations decrease exponentially. ETS considers trend, error and seasonal components of the time series data and goes through several alternative models before finalising the best-fitting model. The model comes with limitations like limited incorporation of external factors and the assumption of continuity in historical patterns.

2.4.5 Discussion

Strengths and weaknesses of each model described in the literature are tabulated in Table 2.1. For traditional statistical models and ARIMA, the computational requirements are smaller, the model is less complex, thus offering improved explainability. These models are able to provide reliable results with a limited amount of data. However, the model faces limitations in handling large, complex, and multivariate data, which limits their ability to solve real-life complex problems. The prediction terms are usually shorter compared to machine learning techniques.

On the other hand, random forest and artificial neural networks can incorporate a large amount of data, which is especially useful in predicting health expenditure that is affected by multiple driving factors. Non-linearity in the data can be captured as well. Nonetheless, the models also come with weaknesses like large computational requirements and consume more time. Due to the increase in model complexity, the interpretability of machine learning techniques is lower.

Table 2.1 Summary of the strengths and limitations of the model

Model	Strength	Limitation	References
Autoregressive integrated moving average	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)

Table 2.1: Continued

Model	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)

2.5 Determinants of health expenditure

There are two distinct approaches in the prediction of health expenditure observed from the current literature. One approach is by conducting time series analysis using lagged variables (past values) of the dependent variable. Another approach involves the incorporation of determinants of health expenditure, such as GDP, demographics, the number of hospitals and physicians, through econometric analysis. As health expenditure is affected by multiple drivers, the latter has improved forecast accuracy and can project long-term spending. It also allows policymakers to test "what-if" scenarios of new policies and recalculating future expenditure estimation.

Among health spending prediction models discussed above, macro-level models like time-series analysis that focus on total health expenditures are useful for short-term projection when trends are well-defined and uninterrupted. Forecasting models that analyse health spending using the sub-components in health expenditures or other determinants of health expenditures are more flexible but require more training data for forecasting. The determinants of health used in the studies are summarised in Table 3.2 below. It can be seen that determinants used as independent variables in the research vary depending on the researchers and their country of study, but some common elements can be found among them, which include GDP, the number of physicians and hospitals, and population over 65 years old.

In the context of Malaysia, Khan et al (2016) propose that GDP per capita, population growth, population structure, and technology have a positive influence on healthcare expenditure by applying Autoregressive Distributed Lag (ARDL) approach, which is an econometric model to analyse the relationship between time series data. Yap & Selvaratnam (2018) use a similar approach and suggested that GDP per capita, healthcare cost index, population aged >65 years, and the mortality rate of infants are important determinants for public health expenditure in Malaysia.

Table 2.2 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

2.6 Research Gap

Several research gaps were discovered in the literature review process. Firstly, there is limited academic research on Malaysia's health expenditure forecasting despite rising needs. In contrast, many countries have research on health expenditure using advanced forecasting techniques to support health financing decisions. The application of a predictive model by using open data sources in Malaysia to forecast future health expenditure could address the gap. By leveraging time series models and machine learning algorithms, this study can support policy decisions and improve financial planning in healthcare.

From the literature, it is shown that health expenditure is affected by multiple factors, and traditional statistical models struggle to capture complex and non-linear relationships. Machine learning techniques can offer improved multivariate forecasting accuracy in health expenditure. Comparison of performance metrics between the models used can be made to determine the best model.

Lastly, different determinants or independent variables of health expenditure are used in the studies from different countries, and there is no general consensus between studies on which determinants to include. This can be addressed by comparing determinants of health expenditure in global and local economic studies. The selected determinants' correlation with health expenditure in Malaysia will need to be analysed before applying them to the prediction model.

2.7 Summary

This chapter discusses the challenges of health spending in Malaysia and highlights the research gap where publications on prediction models for health expenditure are limited. Several machine learning approaches to forecast health spending are reviewed, with their respective strengths and limitations outlined. The determinants of health expenditures used in the previous studies are compared and

contrasted with local Malaysia research to propose relevant factors for the prediction model.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter includes a research framework describing the research workflow. The framework starts with the problem formulation, which provides a context for the problem identified from current research through literature reviews. Relevant data is collected based on the problem defined from reliable sources. Data pre-processing and exploratory data analysis are proposed for the dataset collected to discover patterns in the dataset and prepare the data for prediction. Two machine learning models are proposed for the prediction of health expenditure. The chapter ends with a summary after discussing result evaluation and hyperparameter tuning.

3.2 Research Framework

The research framework of this study involves 6 phases. The details of the research framework are listed in Figure 3.1.

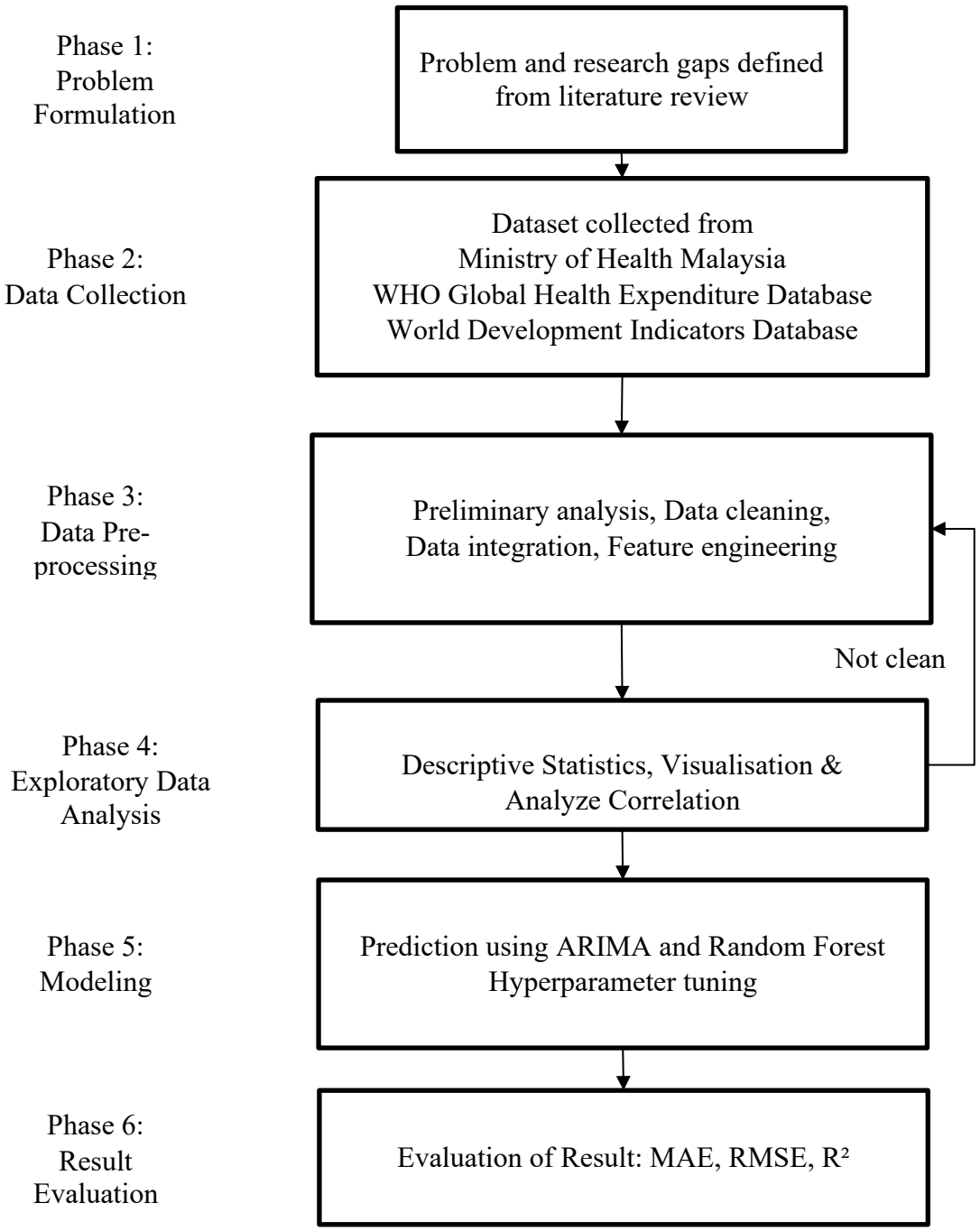

Figure 3.1 Research Framework of Health Expenditure Prediction

3.3 Problem Formulation

This study aims to predict the health expenditure of Malaysia to 2035, to provide insight for policymakers in health financing planning. From the literature review, several research gaps have been identified. Limited academic research has been done on Malaysia's Health expenditure forecasting, and there is no general consensus between studies on which determinants of health expenditure to be used for forecasting.

Current problems to be solved:

1. Data collection needs to be done for health expenditure-related data in Malaysia from year 2000-2022. Issues with data need to be resolved through data pre-processing, which includes data cleaning and data integration.

2. The correlation between variables is required to be analysed through exploratory data analysis. This will support the feature selections for the prediction model.

3. ARIMA and Random Forest will be used for the prediction of health expenditure. Evaluation and comparison of models need to be done to select the best model.

3.4 Data Collection

Data were collected from Ministry of Health Malaysia website, WHO Global Health Expenditure Database and World Development Indicators Database. Table 3.1 contains dataset details and the sources.

MHNA_2022.csv and MHNA_2017.csv datasets were scraped from the table in MHNA 2022.pdf and MHNA 2017.pdf to provide accurate data on total health

expenditure. The table scraping was done by using importing tabula library in Python. The scraped table was saved in the form of a CSV file.

For the dataset NHA indicators.xlsx, the data explorer of the WHO Global Health Expenditure Database was assessed on 26 May 2025. The respective indicators (as shown in Table 3.2) were selected, with the country chosen as Malaysia with the period set as 2000 to 2022. Currency is set as in million current National Currency Units (NCU). Excel file was generated and downloaded.

The last dataset is downloaded from World Development Indicators Database. The indicators were selected accordingly, with the year set as 2000 to 2022 and the country set as Malaysia. A Zip file was generated and downloaded, with 2 Excel files containing data and metadata of the dataset. Only the dataset P_Data_Extract_From_World_Development_Indicators.csv was used for the subsequent steps.

According to literature reviews, these datasets provide data required for predictive modelling. Table 3.2 describes the variables present in the datasets. The dataset contained time series data of health expenditures from 2000 to 2022, including Total Health Expenditure, Current Health Expenditure, Domestic General Government Health Expenditure and Out-of-pocket as a percentage of Current Health Expenditures.

The determinants of health proposed were included for further analysis, which include 8 distinct features, which are population size, GDP, number of physicians, number of hospital beds, population aged 65 years old and above, infant mortality rate, annual population growth and life expectancy at birth. These features were analysed in the exploratory analysis sections before being selected as predictors of health expenditures.

Table 3.1 Dataset details and sources

No	Dataset	Year	File size	Columns	Rows	Sources
1	MHNA_2022.csv	2011 - 2022	1KB	3	13	(MOH, 2024)
2	MHNA_2017.csv	1997 - 2017	1KB	3	22	(MOH, 2019)
3	NHA indicators.xlsx	2000-2022	8KB	26	8	(World Health Organization, 2025b)
4	P_Data_Extract_From_World_Development_Indicators.csv	2000-2022	2KB	27	12	(The World Bank, World Development Indicators, 2025)

Table 3.2 Dataset and variables

Dataset	Variables
MHNA_2017.csv, MHNA_2022.csv	Total Health Expenditure **(TEH)** in million (MYR)
NHA indicators.xlsx	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)
P_Data_Extract_From_World_ Development_Indicators.csv	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 birth)
	Population Growth (annual %)
	Life Expectancy at birth, total (years)

3.5 Data Pre-processing

The data pre-processing for this project can be divided into the following subsections: preliminary analysis, data cleaning, data integration and feature engineering. Each step is crucial in ensuring the quality of the data for analysis and modelling. This step was done iteratively with exploratory data analysis to prepare a cleaned dataset before modelling.

3.5.1 Preliminary analysis

Preliminary analysis is the process of inspection of raw data to understand the dataset. It is done in the first part of data-preprocessing because it determines what subsequent steps should be done to the dataset to achieve the goals.

First, several Python libraries were imported into Jupyter notebook, which runs on a web browser. The libraries imported include numpy, pandas, and matplotlib. Next, the datasets were imported into the notebook. The dataset's information was accessed through pandas method df.info() and df.head() to check for the first few rows of data, how many rows and columns are present, data types of each column. Missing values, or null values, and duplicated values in the datasets are checked. The issues with the dataset were identified and noted to be resolved in the next step.

3.5.2 Data Cleaning

All the collected dataset needs to be cleaned before applying machine learning models to ensure accurate prediction is obtained and prevent errors from occurring due to missing values or inappropriate formats. The data cleaning process was done on each dataset individually. The data cleaning of the dataset starts by dropping unused columns and rows after exploring the dataset. The structure of the dataset was fixed by making sure the features are aligned across the columns. This step involves transposing the DataFrame.

```
#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
```

Figure 3.2 Dropping Unused Columns and Rows

```
#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
```

Figure 3.3 Fixing the Structure by Transposing the Dataframe

The datasets contain time series data. If the data changes gradually, missing values are filled by linear interpolation, which is estimated from the data points close to the missing values. However, if the missing values cannot be interpolated, for example, located at the start or end, backfill and forward fill were applied accordingly. The data points were imputed instead of being dropped because dropping data might disrupt the continuity of the time series data. Duplicated rows are identified and dropped. As the period used is between 2000 to 2022, the data outside the intended time frame is dropped as well.

```
# 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)']

wdi_data
```

Figure 3.4 Filling Missing Values

The format of each dataset is reorganised by ensuring the year is the first column, in ascending order. If the dataset has year as its columns, it is transposed to make sure that all datasets are aligned the same. The index will need to be converted into columns if it contains important information, like the year. The index will be reset when necessary. This will prepare for the process of merging the dataset. The data types are changed into suitable data types for the columns, for example, text and integer. If there are commas between the integers or unintended brackets or parentheses, they will need to be removed. The columns are also renamed to reflect their features. It will make the dataset cleaner and easier for interpretation.

```
#change datatype of whole dataset except year to float
columns_to_convert = wdi_data.columns.difference(['Year'])
wdi_data[columns_to_convert]= wdi_data[columns_to_convert].astype(float)
```

Figure 3.5 Correcting Data Types of the Columns

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

Figure 3.6 Correcting Data Types of the 'Year' Column and Set as Index

3.5.3 Data Integration

Data integration refers to the compilation of datasets from various sources into a unified dataset. This allows data to be compared easily and is ready for the machine learning algorithm. 4 datasets are being used in this project. The datasets are merged into a single dataset after the cleaning process through concatenation and merging. Pandas' merge method will be used for this function, with an inner join chosen and merge on the 'Year' column. This will align variables across the dataset by using year as the key to consolidate a comprehensive dataset for analysis.

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

Figure 3.7 Concatenation

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

Figure 3.8 Merging the Datasets

3.5.4 Feature Engineering

Feature refers to a variable in the dataset. In this step, the new feature was calculated from the existing features. For example, the Out-of-pocket health expenditure is represented as a percentage of Current Health Expenditure (CHE). Calculation of actual out-of-pocket health expenditure was conducted by multiplying the columns of CHE by the column with the percentage.

```
# 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)']
)
```

Figure 3.9 Feature Engineering

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. This is based on the analysis done in the exploratory data analysis.

3.6 Exploratory Data Analysis

Exploratory Data Analysis was carried out to understand the dataset and identify hidden patterns in the dataset. This process can provide understanding of the data distribution, trends, relationship between features, and detect outliers and

anomalies. EDA and data pre-processing are iterative processes where the anomalies that are detected can be cleaned by repeating the data cleaning process. Descriptive statistics of the dataset will be computed and presented to provide an overview of the dataset. This will be done by df.describe() method.

Python libraries imported for this step include Matplotlib and Seaborn. Time series analysis was done by plotting a line graph for the health expenditures to analyse the data. Visualization helps to determine trend and seasonality in the time series data. It will determine whether the data is stationary and provide insight into the need for differencing in the next step.

```
# 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()
```

Figure 3.10 Plotting Line Graph for Total Health Expenditure

Detailed correlation analysis was done by calculating the correlation between the features and target variable. Correlation between the variables were visualised by a Correlation heatmap. Renaming columns is done for easier view in visualization This step supports the feature selection step in data pre-processing.

```
#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
```

Figure 3.11 Renaming Columns and Compute Correlation

```
# 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)
```

Figure 3.12 Generate Heatmap from the Correlation Computed

3.7 Modeling

Two different approaches were used for data modelling. The models were trained on training data, evaluated with performance metrics according to test data, and then used to predict health expenditure to 2035. ARIMA and Random Forest were applied to the pre-processed dataset to predict the future health expenditure in Malaysia. For ARIMA, input data will consist of past data of total health expenditure, GHE and OOP as a time series analysis to predict to future value of health expenditure. For Random Forest, as the machine learning model can handle more variables than the ARIMA model, multiple features were selected for data modelling.

3.7.1 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). It integrated autoregressive modelling and moving average modelling, which are distinct approaches for predicting time series data. 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. It is represented by the equation

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \quad (3.1)$$

where y_t is the dependent variable, and the lagged observations of y_t are used as predictors. c is the constant term, ε_t is the white noise or error term, while ϕ are the autoregressive coefficients, and p is the number of lagged components. The changes in ϕ varies the patterns while ε_t varies the scale of the time series (Hyndman & Athanasopoulos, 2025).

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. MA can be presented in the equation

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (3.2)$$

where y_t is the dependent variable, ε_t is the random noise, ε_{t-1} is the previous noise, θ is the moving average coefficient, c is the constant or long-term mean of the process, and q is the number of lagged error components. It is combined with AR to improve attention for recent incidents than the pure AR process (Siegel, 2016).

The integrated (I) part aims to turn the time series stationary by performing differencing to eliminate trend and seasonality. Trend is the long-term increase or decrease in the time series, for example, rising prices in GDP. Seasonality is the regular patterns that occur. Differencing can stabilise the mean of the time series by eliminating the changes in the level of a time series, thus removing or reducing trend and seasonality (Hyndman & Athanasopoulos, 2025). To determine the requirement for differencing, a unit root test needs to be done, and the Augmented Dickey-Fuller (ADF) test is chosen in this project

The integrated part is done so that non-stationary time series can be used for ARIMA process because both AR and MA assume stationarity. By combining all three components, the ARIMA model smooths out the changes while maintaining a general pattern of the time series (Siegel, 2016). Forecasting with ARIMA is useful for the prediction of health expenditure, where the values do not return to the long-term mean value.

Python is used to model ARIMA in this project. The required packages are imported from the statsmodels library as described in Figure 3.13.

```
# 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
```

Figure 3.13 Packages imported for ARIMA model

ADFuller was imported, which conducts Augmented Dickey-Fuller test to confirm stationarity of the data. Differencing has to be done if the data is not stationary. The order of differencing required is the minimum number of differences needed to produce a stationary series.

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

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

Figure 3.14 ADFuller

Before running the ARIMA (*p, d, q*) model, *p, d, q* terms need to be determined. ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) are two statistical tools used to determine the terms. The number of lagged values that ACF cuts off is set as q, and the same applies for p using PACF. A suitable value of d is

selected by ensuring that after the differencing all trends and seasonality have been eliminated.

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

plt.show()
```

Figure 3.15 Plotting ACF and PACF plot

The forecast can be done after the *p, d, q* terms are determined. Chronological splitting is used, where data is split in 80% training and 20% testing. This means data from 2000 to 2018 is used as training data, while 2019 to 2022 is used as testing data. The training data was fit to the ARIMA model by setting the parameters accordingly. The forecast result and actual result were visualised by using line plot after reversing the differencing done to the forecasted result. The model performance will also be evaluated by using AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) model statistics to decide the best combination of terms. The lower the value, the better the model fits. Other statistical metrics, as mentioned in Section 3.8, are computed and compared with other models. After the model is tuned, estimation of health expenditure to 2035 will be done.

```
# 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
```

Figure 3.16 Train Test Split and Fitting Data to ARIMA model

3.7.2 Random Forest

Random Forest is a supervised machine learning method that is chosen to model the dataset. Tree-based methods can be applied to regression problems, therefore, suitable for the prediction of healthcare expenditures. 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, each having moderate predicting capability, to achieve higher forecasting accuracy at the cost of some interpretability.

In bagging or bootstrap aggregation, several decision trees are created based on bootstrapped training samples. This is to overcome the main weakness of decision trees: high variance, which means each decision tree can produce very different results from the training data if we split the data at random and feed it to the decision trees. Bootstrapping refers to the process of resampling the training dataset with replacement to create many simulated samples. This allows multiple decision trees to be trained on the bootstrapped training dataset, and the results from each decision tree are combined and averaged to reduce variance.

Random Forest improves the bagging procedure by decorrelating the trees. This algorithm limits only several features, randomly selected from the total number of features in the training set, to be considered during each split in the tree. Commonly, m predictors are used at each split, calculated by the equation $m = \sqrt{p}$, where m, the number of predictors (features or independent variables), equals to square root of p, the full set of predictors. The algorithm aims to resolve the issue with bagging where the strong predictor in the training data will always be used at the top split in the tree, which results in the similarity of decision tree results when other less determining predictors are used at the split downwards (James et al., 2023).

Python is applied for this research and scikit learn library is imported for the purpose of training random forest on the cleaned data. There are several classes that are imported from different modules as shown in Figure 3.17.

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import cross_val_score, KFold
```

Figure 3.17 Classes Imported for the Random Forest Model

Next, the feature columns (X) and the target column (y) to predict will be set. The feature columns include all the features selected in the feature engineering step, while the target column to predict is the total health expenditure. Then, the data is split into a training and a testing set. As the data is a time series, chronological splitting is used. Data from 2000 to 2018 is used as training data, while 2019 to 2022 is used as testing data. This will result in 4 sets, which are X_train and X_ test for the independent variable, y_train, y_test for the dependent variable.

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

# 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)']
```

Figure 3.18 Train Test Split and Feature Selection

The random forest regressor is initiated. This model builds several decision trees and combines the predictions. Random state is set at 20 to ensure the result. There are multiple parameters in this model, including where number of decision trees, the maximum depth of the tree, the minimum samples to split an internal node and a leaf node, maximum features and the bootstrap option.

```
#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 3.19 Instantiate the Random Forest and Predict the Result

The regressor is fit to X_train and y_train, respectively, to train the model. After the model is trained, predictions can be made using the X_test as input and predict() method on the model. The prediction can be visualised by using a residual plot and a line plot to compare with the y_test result. The performance of the model was then evaluated with the metrics as described in the next section. All parameters in this model will be tuned as described in Section 3.9, Hyperparameter tuning, to achieve the best performance. The parameter setting with the best performance will be selected for final comparison with ARIMA model. After the model is fine-tuned, the prediction of health expenditure for 2035 will be done.

3.8 Evaluation

The models will be evaluated for their performance. This step is crucial as it not only validates the performance of the model, but also enables comparison of performance between different models. There are 3 key evaluation metrics applicable to regression models that will be used in this project, which will be discussed in detail in the following subsection.

```
: 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))
```

Figure 3.20 Evaluation metrics imported from Sklearn.metrics

3.8.1 Mean Absolute Error (MAE)

Mean absolute error calculates the mean of the errors by their absolute value. The absolute difference between the actual value and predicted value is calculated to avoid the effects of negative values cancelling out the positive value during summation. The absolute values of error are summed up and divided by the number of observations to get the average error. The equation of MAE is

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.3}$$

where n is the number of observations, y_i is the actual value and $\hat{y}_i$ is the predicted value. The metric is included because it is a simple metric for interpretation.

3.8.2 Root Mean Squared Error (RMSE)

Root Mean Squared Error is calculated by summing up squares of all the errors, calculating the mean, then square-rooting the result. The metric is square-rooted, so it is easier to compare to other metrics, like with mean absolute error. The equation of RMSE is

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \tag{3.4}$$

where n is the number of observations, y_i is the actual value and $\hat{y}_i$ is the predicted value. This metric measures the average magnitude of prediction error and is suitable for this project, where larger prediction errors need to be penalised more heavily to ensure the reliability of expenditure forecasting.

3.8.3 Coefficient of Determination (R^2 score)

The coefficient of determination, or R^2 is indicates the goodness of fit of a model. It reflects the model's performance in forecasting the outcomes. R^2 score equals to 1 indicates all the actual value lies perfectly on the prediction model, while $R^2 = 0$ indicates the model does not fit any actual value. The equation of R^2 is

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \tag{3.5}$$

where n is number of observations, y_i is the actual value, $\hat{y}_i$ is the predicted value and $\overline{y_i}$ is the mean of actual value. Higher R^2 indicates better fit to the model.

3.9 Hyperparameter tuning

For the Random Forest model, hyperparameter tuning will be conducted using GridSearch Cross Validation to optimise the key parameters. Grid Search is chosen over the randomised search method because the dataset is small, and the Grid Search CV will result in better performance.

6 main parameters in this model are aimed to be tuned:

i) The number of trees in the forest.

ii) The maximum depth of the tree.

iii) The minimum samples to split an internal node.

iv) The minimum number of samples required to be at the leaf node.

v) The maximum features when finding the best split.

vi) Bootstrap. This will be set between True and False to determine whether the samples are bootstrapped when building the tree.

The parameters are set in a parameter grid, which is a dictionary mapping each parameter to the range of values that are intended for grid search. GridSearchCV is imported from sklearn.model_selection. The number of cross-validation (cv) is set to 5-fold. The RandomForestRegression, parameter grid, and cv are set as the parameters for GridSearchCV, then fit to the training data. The best parameter and best estimator are printed out to show the best combination of hyperparameters from the search. Lastly, the model is updated by using the parameters from the result above.

3.10 Summary

This chapter discusses the research methodology, from problem formulation, data collection, data cleaning, data pre-processing, exploratory data analysis, modelling and result evaluation. The ARIMA model and Random Forest model are discussed in depth, from the introduction, equations, strengths and weaknesses, parameters, statistical tools for setting up the model, hyperparameter tuning and the proposed application of the model by using Python. Three statistical metrics and proposed for the evaluation of the results between models. Initial results of the project will be presented in the next chapter.

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)
```

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	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

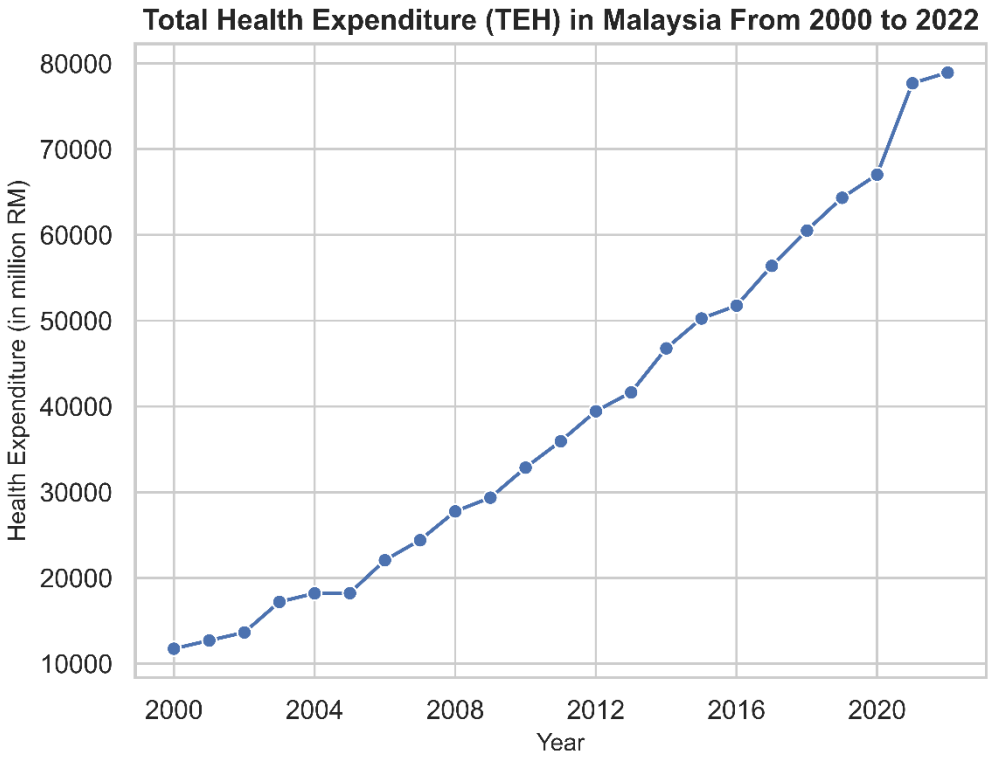

Figure 4.3 Total Health Expenditure in Malaysia over Year

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.

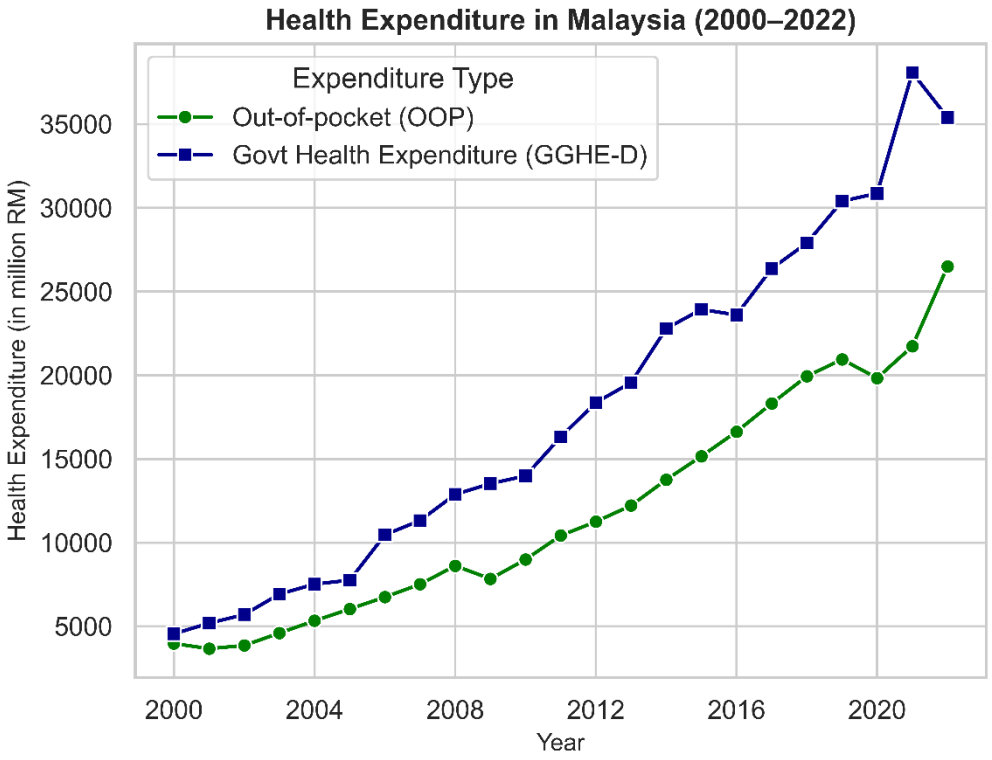

Figure 4.4 Health Expenditure by Type in Malaysia over Year

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

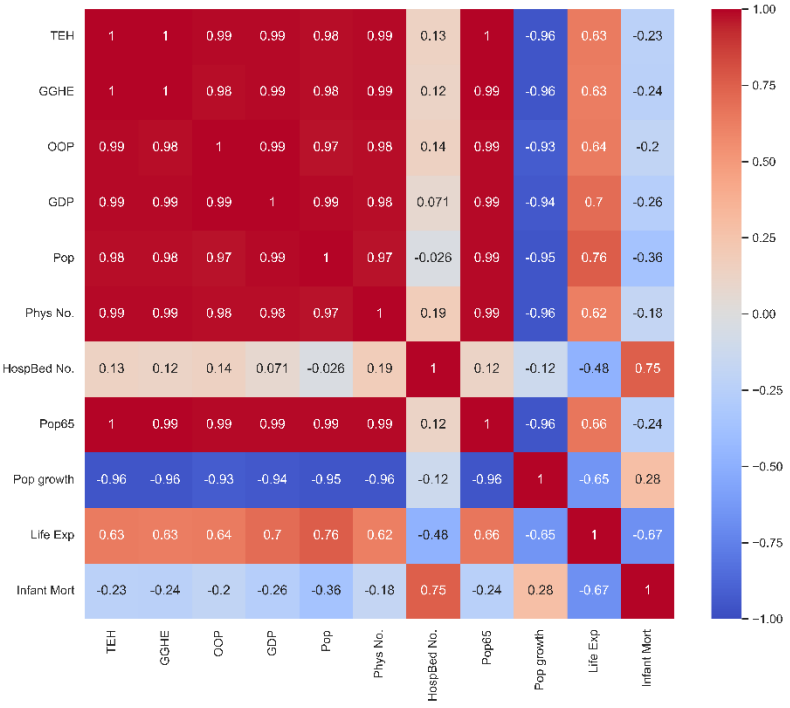

Figure 4.5 Correlation Heatmap

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

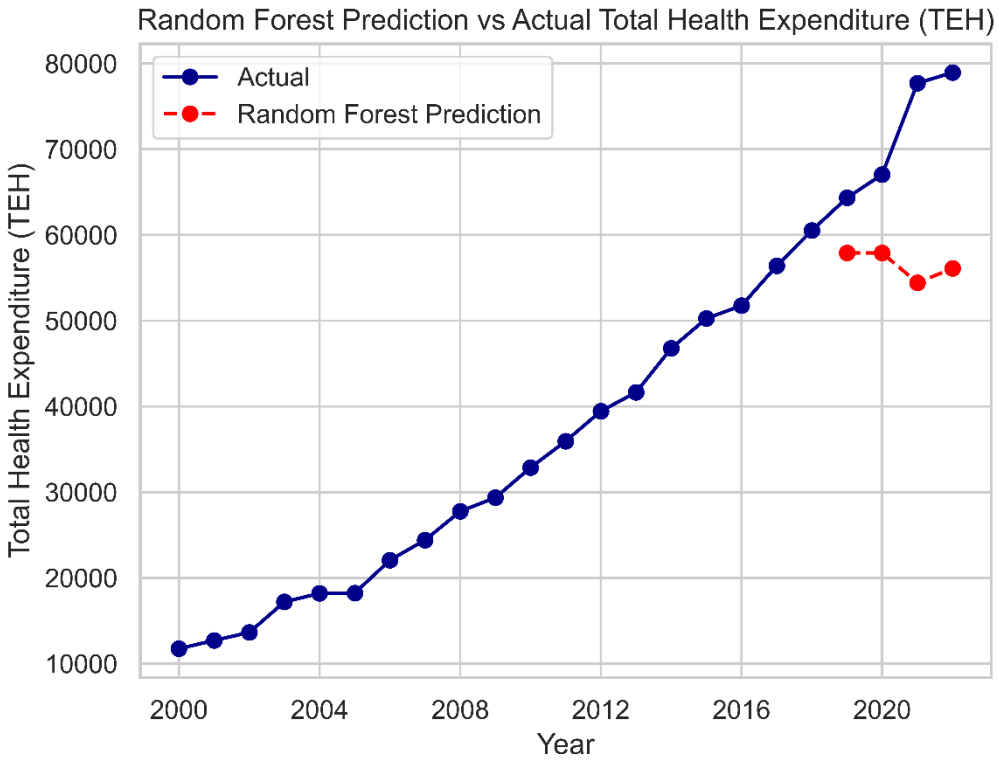

Figure 4.8 Random Forest Prediction versus Actual Total Health Expenditure (2000 to 2022)

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.

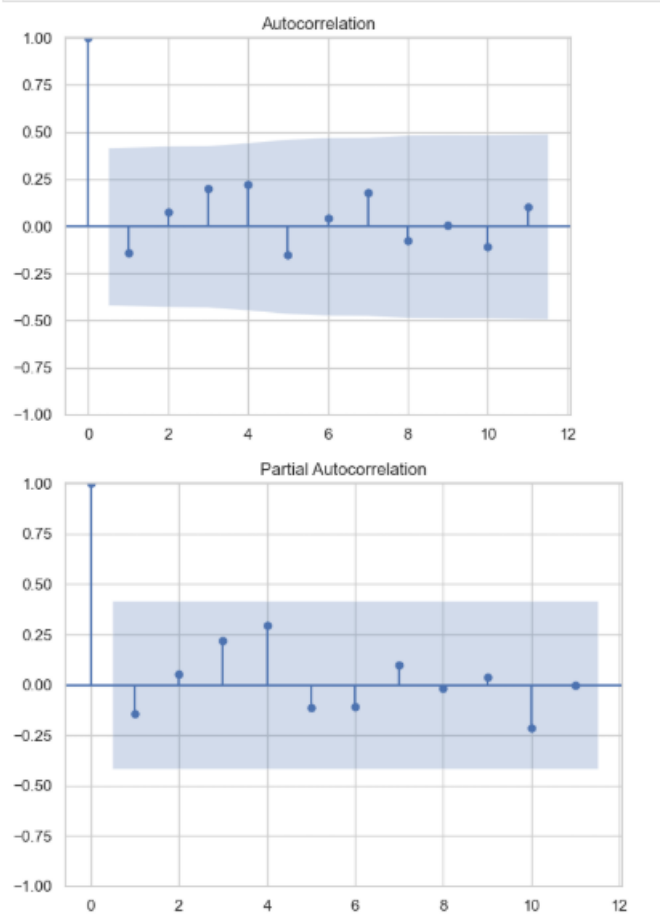

Figure 4.10 ACF and PACF Plot Result

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 - 01-01-2018	HQIC	326.277
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

```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
```

Figure 4.12 Reverse First-order Differencing and Compute Evaluation Metrics

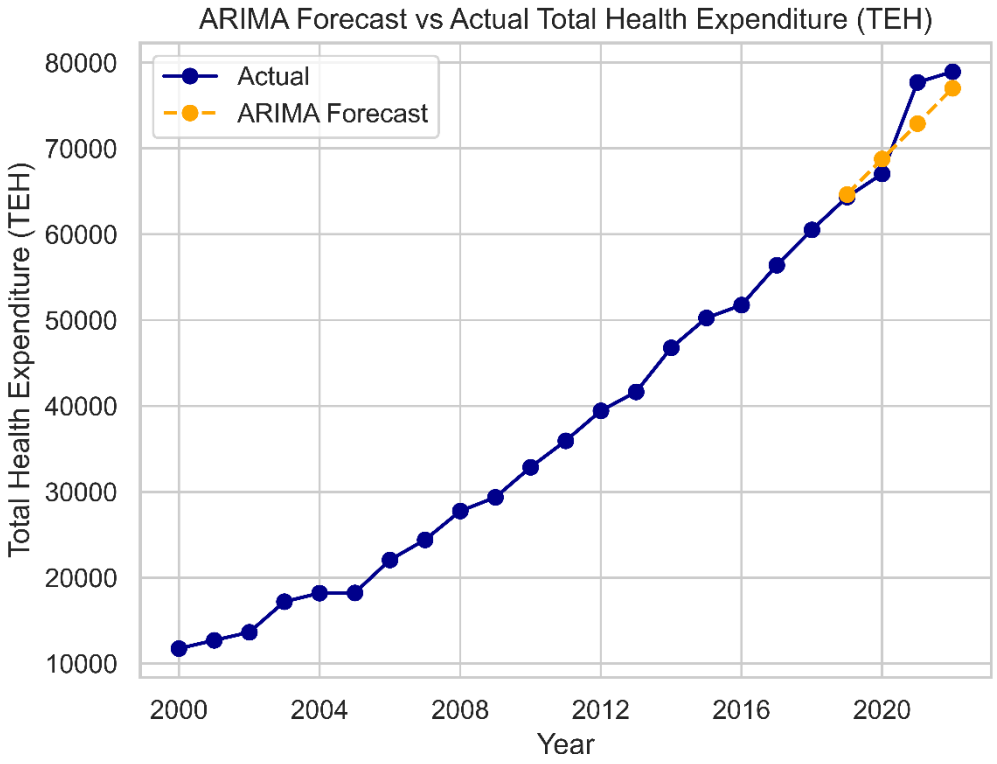

Figure 4.13 ARIMA Forecast versus Actual Total Health Expenditure (2000 to 2022)

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.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

5.1 Conclusion

This project has analysed and predicted the health expenditure in Malaysia. The healthcare spending data is collected from Ministry of Health Malaysia, WHO Global Health Expenditure Database and World Development Indicators Database for the period of 2000 to 2022. The collected datasets are pre-processed and merged for analysis and forecasting. Explanatory data analysis shows a gradual increase in health expenditures from 2000 to 2019, and some fluctuations between 2019 to 2022 due to the COVID-19 pandemic. Key determinants affecting health expenditure have been identified by investigating their correlations. Feature engineering has been done to select features to train the machine learning model.

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. However, it is notable that in 2021, the ARIMA struggled to predict accurately due to an unexpected surge in health spending caused by the COVID-19 pandemic. 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 and suggest that a complex model like Random Forest requires additional data and further optimisation to perform effectively in the forecasting task. In 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.

5.2 Future works

In this project, the health expenditures of Malaysia are predicted using machine learning models. Hyperparameter tuning can be conducted for the model to further improve the results. Currently, the random forest does not predict with good accuracy and requires further tuning. Also, the lagged features of health expenditures can be calculated and used as input for the random forest model, simulating how ARIMA model works to enhance its accuracy. Cross-validation should be done to ensure the reliability of the forecast by the models.

This study is conducted based on overall health expenditure data and macroeconomic data only. The health expenses prediction can be used for smaller components in the healthcare spending, for instance, outpatient and inpatient services, pharmaceutical expenditures, education and training, provided that the accessibility to detailed data is granted. In addition, individual factors like patients' age, gender, medical conditions, current medications, income level and family history of illness can be considered for individual healthcare cost prediction. This will provide a better understanding to the end-users of Malaysia's healthcare system and promote improved health outcomes for the public.

Moreover, the methodology of this study can be extended to ASEAN countries with similar health economic structures, for instance Thailand, Indonesia and Philippines. By applying the machine learning models across multiple countries, researchers can compare the differences in health expenditure trends, key determinants of health spending and most importantly, forecasting accuracy in different nations. This may improve the generalisability of the machine learning models and provide insight for improving healthcare budget planning based on varying health policies across counties.

REFERENCES

Ahmed, S.F., Alam, M.S.B., Hassan, M., Rozbu M.R., Ishtiak T., Rafa N., Mofijur M., Shawkat Ali A.B.M. & Gandomi Amir H. (2023). Deep learning modelling techniques: current progress, applications, advantages, and challenges. Artif Intell Rev 56, 13521–13617.https://doi.org/10.1007/s10462-023-10466-8

Astolfi, R., L. Lorenzoni & J. Oderkirk (2012), "A Comparative Analysis of Health Forecasting Methods", OECD Health Working Papers, No. 59, OECD Publishing, Paris, https://doi.org/10.1787/5k912j389bf0-en.

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

Devi, P. & Bansal, K.L. (2024). Data science in healthcare: techniques, challenges and opportunities. Health Technol. 14, 623–634. https://doi.org/10.1007/s12553-024-00861-8

Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (3rd ed.). O'Reilly Media.

IBM. (n.d.). What is random forest? IBM. Retrieved May 19, 2025, from https://www.ibm.com/think/topics/random-forest

Hyndman, R.J., & Athanasopoulos, G. (2025). Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts. Retrieved June 6, 2025, from https://otexts.com/fpp2

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An introduction to statistical learning: With applications in Python. Springer. https://doi.org/10.1007/978-3-031-38747-0

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

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

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

Khan, H. N., Razali, R. B., & Shafie, A. B. (2016). Modeling Determinants of Health Expenditures in Malaysia: Evidence from Time Series Analysis. Frontiers in pharmacology, 7, 69. https://doi.org/10.3389/fphar.2016.00069

Khor, K.S., Chua, E.P.W. & Fried, C. (2024). Sustainability and Resilience in the Malaysian Health System. Center for Asia-Pacific Resilience and Innovation (CAPRI). https://www3.weforum.org/docs/WEF_PHSSR_CAPRI_Malaysia_2024.pdf

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., & Matsopoulos, G. K. (2023). A Review of ARIMA vs. Machine Learning Approaches for Time Series Forecasting in Data Driven Networks. Future Internet, 15(8), 255. https://doi.org/10.3390/fi15080255

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

Lee, W., Schwartz, N., Bansal, A., Khor, S., Hammarlund, N., Basu, A., & Devine, B. (2022). A scoping review of the use of machine learning in health economics and outcomes research: Part 2—Data from nonwearables. Value in Health, 25(12), 2053–2061. https://doi.org/10.1016/j.jval.2022.07.011

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

Lorenzoni, L., Marino, A., Morgan, D., & James, C. (2019), Health Spending Projections to 2030: New results based on a revised OECD methodology, OECD Health Working Papers, No. 110, OECD Publishing, Paris, https://doi.org/10.1787/5667f23d-en.

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

Muremyi, R., Haughton, D., Kabano, I., & Niragire, F. (2020). Prediction of out-of-pocket health expenditures in Rwanda using machine learning techniques. The Pan African medical journal, 37, 357. https://doi.org/10.11604/pamj.2020.37.357.27287

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

Organisation for Economic Co-operation and Development. (2024). Fiscal sustainability of health systems: How to finance more resilient health systems when money is tight? https://doi.org/10.1787/880f3195-en

Rubinger, L., Gazendam, A., Ekhtiari, S., & Bhandari, M. (2023). Machine learning and artificial intelligence in research and healthcare. Injury, 54(Supplement 3), S69–S73. https://doi.org/10.1016/j.injury.2022.01.046

Sayuti, M., & Sukeri, S. (2022). Assessing progress towards Sustainable Development Goal 3.8.2 and determinants of catastrophic health expenditures in Malaysia. PloS one, 17(2), e0264422. https://doi.org/10.1371/journal.pone.0264422

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

Siegel, A. F. (2016). Time series: Understanding changes over time. In A. F. Siegel (Ed.), Practical business statistics (7th ed., pp. 431–466). Academic Press. https://doi.org/10.1016/B978-0-12-804250-2.00014-6

The World Bank, World Development Indicators (2025). P_Data_Extract_From_World_Development_Indicators [Data file]. Retrieved from https://databank.worldbank.org/source/world-development-indicators#

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

Wan Puteh, S. E., Abdullah, Y. R., & Aizuddin, A. N. (2023). Catastrophic Health Expenditure (CHE) among Cancer Population in a Middle Income Country with Universal Healthcare Financing. Asian Pacific journal of cancer prevention : APJCP, 24(6), 1897–1904. https://doi.org/10.31557/APJCP.2023.24.6.1897

World Health Organization. (2020). Global spending on health 2020: weathering the storm. World Health Organization. https://iris.who.int/handle/10665/337859.

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

World Health Organization. (2025b). NHA indicators. [Dataset] Retrieved from https://apps.who.int/nha/database/Select/Indicators/en

Wubineh, B.Z., Deriba, F.G., & Woldeyohannis, M.M. (2024). Exploring the opportunities and challenges of implementing artificial intelligence in healthcare: A systematic literature review. Urologic Oncology: Seminars and Original Investigations, 42(3), 48–56. https://doi.org/10.1016/j.urolonc.2023.11.019

Yap, K. W., & Selvaratnam, D. P. (2018). Empirical analysis of factors influencing the public health expenditure in Malaysia. Journal of Emerging Economies and Islamic Research, 6(3), 1-14.

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.