GOVERNMENT REVENUE PREDICTION USING FEED FORWARD NEURAL NETWORK

NADZIRA NOOR[1], ALIZA SARLAN[2], NORSHAKIRAH AZIZ[3]

[1]Department of Computer and Information Sciences, Universiti Teknologi Petronas (UTP), Seri Iskandar, 32610, Malaysia.
[2]Department of Computer and Information Sciences, Universiti Teknologi Petronas (UTP), Seri Iskandar, 32610, Malaysia.
[3]Department of Computer and Information Sciences, Universiti Teknologi Petronas (UTP), Seri Iskandar, 32610, Malaysia.

E-mail: [1]nadzira.nka@gmail.com, [2]aliza_sarlan@utp.edu.my, [3]norshakirah.aziz@utp.edu.my

ABSTRACT

A country's federal government receives revenue from several sources. Example in Malaysia the sources are direct tax, indirect tax and non-tax revenue. The federal government will then use the revenue for operations and developments in the country. There are currently limited methods to predict federal government revenue for upcoming years. Having different and better method can help to better plan the collection activities and managing the resources. For now, Malaysia federal government can only forecast or estimate the revenue. Business intelligence on the other hand is currently booming in the business world as it helps to improve and provides relevant information for decision making process. One of the branches of business intelligence is predictive analytics, where it can be used to predict future outcomes provided past data are available. Patterns can be identified to predict the upcoming trend. From the observation, predictive analytics can be applied in any financial prediction which includes federal government revenue. Numerous machine learning methods exist such as linear regression, polynomial regression, various types of neural network, decision tree, random forest, multiple linear regression and so on. Based on the literature review done, feed forward neural network is highly used and thus selected for this study. Hyperparameter tuning is conducted to determine the ideal parameters for feed forward neural network to be applied for federal government revenue prediction. From the result, it is found out that using Softsign activation function and Adam optimizer can give better accuracy. Completing the study, it contributes to provide another way to accurately predict the federal government's revenue and subsequently be advantageous to the federal government.

Keywords*: Predictive Analytics, Machine Learning, Revenue Prediction, Feed Forward Neural Network, CRISP-DM*

1. INTRODUCTION

1.1 Background

Every country has its own government. In Malaysia, there are 3 tiers of government. They are federal, state and local government [1]. Each government has its own functionality and financial account. Based on the government activities, it will have its own revenue and expenditure. The revenue comes from tax revenue, non-tax revenue and non-revenue receipts. Afterward, the revenue will be used for covering development and operation costs of the government. The operating expenditure includes emoluments, debt service charges, supplies and services, retirement charges, grants and transfers to state governments, subsidies and social assistance, and others. As for development expenditure, it includes economic, social, security and general administration expenses. It is important for a government to have higher revenue than the expenditure to avoid borrowing from other entities. When the government is stable and can generate higher revenue, more investors will be interested to invest and do business in the country [2].

Analysis of financial statements is crucial to determining an organization capacity to generate revenue. It helps the top management make wise decisions about how much money to spend on a specific category or purpose [3]. For now, federal government in Malaysia is doing revenue estimation yearly. It is then documented in the Fiscal Outlook and Federal Government Revenue Estimates report. Estimation or forecasting uses statistical method [4]. It might not be very accurate.

In the information technology world, business intelligence (BI) is getting more attention from all sorts of industry or domain. This is because it has many benefits such as improving business process, saving cost, increasing revenue and reducing risk [5], [6]. Data analytics is a branch of BI. It further categorized into descriptive, predictive and prescriptive analytics [7]. Descriptive analytics is to identify current situation of a business or organization, predictive analytics is to know what might happen in the future based on previous data, and prescriptive analytics is to suggest what can be done to achieve better result based on the prediction [8].

Predictive analytics are using machine learning while estimation or forecasting is using statistical method. Predictive analytics are getting more attention over forecasting as it can get higher accuracy. Although a lot of domains are applying it, financial revenue prediction has lesser attention towards this technology. Since Malaysia government revenue is using estimation, it is a good opportunity to explore financial revenue using predictive analytics.

1.2 Problem Statement

Federal government revenue in Malaysia is currently using estimation or forecasting method [9]. In previous years, the forecasting also had a large error up to -40.18%. Data analytics on the other hand are becoming more famous nowadays because of its accuracy. Besides descriptive and prescriptive, predictive analytics are being used in a lot of domains. Predictive analytics is using machine learning methods. It replaces forecasting that uses statistical methods [4]. Although a lot of domains have applied predictive analytics, less studies for government revenue prediction are to be found [10]. Besides that, sometimes organizations own datasets but are not being fully utilized [11].The organization also might not have knowledge and skills in current advance technologies that can be used.

One might use it for descriptive analytics but not for predictive analytics. Predictive analytics nowadays are proven to be more accurate, unfortunately the study on predictive analytics for government revenue is scarce. It shows there is a need to explore predictive analytics for federal government revenue using machine learning methods. Therefore, the problem statement for this study is low accuracy for revenue forecasting on federal, state and country may lead to budget management issue especially when large forecast error happened [12]–[15].

1.3 Research Questions and Aims

There are 2 research questions for this study:

1. What is the suitable machine learning method that can be used to predict federal government revenue with high accuracy?
2. What are the optimal settings for the machine learning method chosen?

From the research question, there are 2 aims as following:

1. To identify the suitable machine learning method that can achieve high accuracy for federal government revenue prediction.
2. To find out the optimal setting for the machine learning method chosen by applying hyperparameter tuning.

2. LITERATURE REVIEW

The literature review consists on background of Malaysia government revenue, business intelligence that focuses on data analytics, and related works on financial prediction. These items are essential as each of them contributes to the foundation of this study. There are few criteria applied for literature selection such as year, database and domain. For the year, 7 recent years are selected which are from 2016 to 2022, while the database used for literature searching includes Scopus, IEEE and Google Scholar.

The domain is divided into 2 parts. First is for the government revenue background. Focus will be on Malaysia country. The background of Malaysia government, financial and revenue are studied. The second part is on technical which is the method. Research related to financial prediction that uses machine learning are selected. Filtering is also done to focus on research using regression method instead of classification method.

2.1 Malaysia Government Revenue

This subsection explains on Malaysia government, financial division in Malaysia, federal government revenue and its categories, the importance and current situation of Malaysia’s federal government revenue.

2.1.1 Malaysia government

Malaysia is located in Asia. The country has population of 32.6 million people. Within the 330,548 square kilometers, it has 13 states and 3 federal territories. The states are Perlis, Kedah, Pulau Pinang, Perak, Selangor, Negeri Sembilan, Melaka, Kelantan, Terengganu, Pahang, Johor, Sabah and Sarawak. The federal territories in Malaysia includes Kuala Lumpur, Putrajaya and Labuan. To ensure the country and its people is in good condition, the government is responsible to do so. Figure 1 displays the tiers of Malaysia government. They are federal, state and local government.

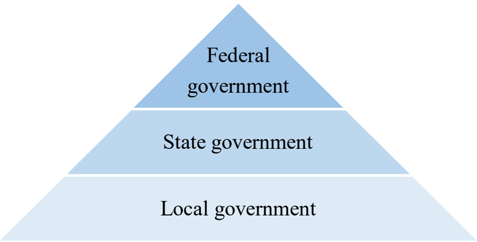


Figure 1: Tiers of Government in Malaysia

Each government has their own functions, responsibility and financial account. The financial account consists of revenue, operating expenditure and development expenditure. Federal government is the top most government. It handles all national and administration aspect. Some of it includes safety and justice of the country, education and health of the people, communication and transportation facilities, and national financial matter [16]. The revenue of federal government is mainly used in the national level.

Next is state government where the jurisdiction is within the state itself. Some items handled by state government are Islamic law, land, agriculture, forestation, state holidays and local governments [16]. For state government, revenue collected is mainly used for the wellbeing of the people. This includes maintaining the state, disposing rubbish, cleaning drains, paying electric bills of public areas and beautifying the landscape [17]. Every state government has a number of local governments under it based on the cities and districts. The state government also needs to take care of all local governments under the same state. The function of having local government is to be the intermediary between state government and the people, ensuring policies are being carried out fairly [17].

2.1.2 Financial division in Malaysia

In Malaysia, the financial division starts from consolidated public sector (CPS). CPS comprises of general government units and non-financial public corporations (NFPC). CPS acts as investor in insufficient investment of private sectors such as fishery and agriculture, and strategic areas such as public transportation. Next is general government, which functions as a unit that conducts main economic task of the government. It is further divided into federal government, state government, local government, and federal statutory bodies. Lastly, for NFPC, it is the public sector unit that is responsible for sale of commercial and industrial goods and services [18]. Some examples of NFPCs are PETROLIAM Nasional Bhd (PETRONAS), Telekom Malaysia Bhd, Tenaga Nasional Bhd, Prasarana Malaysia Bhd and TH Plantation Bhd. Figure 2 visualizes the financial division in Malaysia.

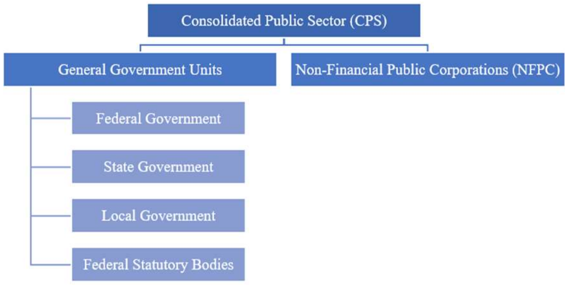


Figure 2: Financial Division in Malaysia

Every unit has their own financial account. It generally consists of consolidated revenue, consolidated operating expenditure and consolidated development expenditure, and financial position or the overall balance that can be surplus or deficit. For consolidated operating expenditure and consolidated development expenditure, when add up both it will become total expenditure. Every division needs to report its financial position. Example for CPS, the reason to report is so that the public sector size, impact of CPS activities on the economy and root of fiscal risks can be identified [9]. Same as general government, by identifying the financial position, performance and impact on economy operations can be estimated.

As for financial risk management, it is contained in the general government level. This is done by executing laws such that local governments can only borrow from and with the approval from its own state government. On the other hand, state government and federal statutory bodies borrow from and with approval of federal government [9]. Currently Malaysia government is able to do estimation for financial position. This can be seen, for the year 2023, federal government financial position has been estimated. The estimated revenue is RM272,570 billion, with overall balance of RM99,070 deficit, which equivalent to -5.5% gross domestic product [19].

2.1.3 Federal government revenue and its categories

Revenue in general is the income obtained from activities or business operation of an organization and it is also known as sales, service income, interests, dividends and royalties [11]. Federal government revenue is the income for the federal government gained from various economic activities which is then used for development and operation purposes. There are 3 main categories of federal government revenue in Malaysia. They are tax revenue, non-tax revenue and non-revenue receipts. Tax revenue is divided into direct tax and indirect tax. Direct tax is further divided into income tax and other direct tax. The main and subcategories are as in Figure 3.

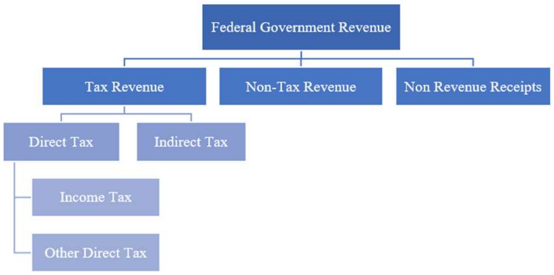


Figure 3: Federal Government Revenue Main and Subcategories

For income tax it includes individual, companies, petroleum and merchants, while others categories under direct tax includes assets and heritage duty, stamp duty, real property gain tax and others. Indirect tax consists of export duty, import duty and cigarette tax, excise duty, sales tax, service tax, goods and service tax, tourism tax, and others. The second main category which is non-tax revenue are made up of PETRONAS Dividend, petroleum and gas royalty, motor vehicle license and road tax, Bank Negara dividend, and other non-tax revenue. The third and last main category is non-revenue receipts consists of 2 subcategories which are non-revenue receipts and federal territory receipts.

2.1.4 Importance of federal government revenue

Government relies on taxes to fund public spending and raise citizen social standards [20]. Another importance of government revenue is to be used for development activities [21]. Government can also decide how much tax should be collected if it has a proper understanding of the taxes earned in prior years. Revenue is always important for any government. It is the source for operating and development expenditure. Whether it is a country, federal, state or a district, revenue is very important to ensure the welfare of the people and improvement of the area.

In Malaysia, the operating and development expenditure will be covering sectors such as economic, social, security and general administration as shown in Figure 4 [1]. The economic expenditure includes for agriculture, energy and public utilities, trade and industry, transportation, communication and environment. For social sector the expenditure is on education and training, health and housing. Lastly for security sector, the expenditure covers defense and internal security.

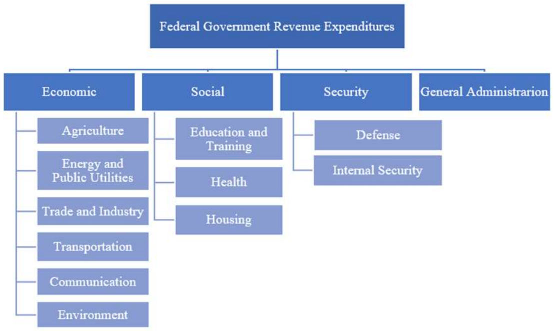


Figure 4: Federal Government Revenue Expenditures

2.1.5 Current situation of federal government revenue in Malaysia

Malaysia government publish yearly Fiscal Outlook and Federal Government Revenue Estimates. The report presents on the fiscal policy overview, federal government revenue, federal government expenditure, debt management, fiscal risk and liability, CPS and the financial account. It also includes the summary and detail of federal government finance revenue.

For the latest year 2021, total financial revenue is RM233,752 million. The direct tax is RM130,116 million (55.7%), indirect tax is RM43.588 million

(18.6%) and non-tax revenue is RM60,048 million (25.7%). For all 3 categories are having decrement for 2 years as the effect of COVID-19 to the health and economy worldwide. Currently geopolitical tension, global supply chain challenges and inflation issue are affecting the world. Despite that, as the COVID-19 is moving towards endemic phase, the report estimated that in the upcoming years 2022 and 2023, the federal government revenue will have increment compared to year 2019.

2.2 Business Intelligence and Data Analytics

BI is a technique, idea, or strategy for leveraging data at hand to enhance decision-making. [22]. It is getting more interests from the academia, business and management since 1989 [23]. The main objective of BI is to assist companies and organisations in operating more profitably. Dashboards and data visualisation, reporting, data mining, extract, transform, and load (ETL), and online analytical processing are some of the key features of BI technologies. [24]. Additionally, there are readily accessible technologies like Power BI, SpagoBI, Qliksense, Tableau and Jaspersoft on the market.

BI is widely employed across a variety of sectors. It helps industries including finance, retail, hotels, insurance, healthcare, education, and even government operate more effectively. In finance for example, prediction of loan approval can be conducted by applying machine learning algorithm [25]. Another example is the use of machine learning in banks to predict loan default that applies eXtreme Gradient Boosting (XGBoost), random forest, AdaBoost, k-nearest neighbor (k-NN) and multilayer perceptrons [26].

Data is owned by companies and organizations, but it is usually not fully utilized. Data can be turned into useful information with the help of BI. This problem can be solved through data vizualizations. Due to their visual orientation, it facilitates faster information processing in humans. Color-coded tables, graphs, and charts can analyze and summarize raw data more quickly. For higher management of a company or organization, these data can be created on a dashboard so they can get a quick, clear picture of the present situation before deciding what to do next or how to proceed with a business strategy. Applying data analytics in BI can have a significant impact on enterprises and organizations going forward.

Under the umbrella of BI is data analytics. Data analytics is the act of identifying trends and important information, making conclusions, and assisting in decision making by doing some steps such as inspect, clean, transform and model specific dataset related to an organization or business [27]. It helps an organization or business to convert information to useful knowledge with the intention to help on making effective decision [28]. This is highly important to improve the organization in the future be it the revenue or operational wise. Figure 5 shows the types of data analytics. Predictive, descriptive, and prescriptive analytics are the 3 types.

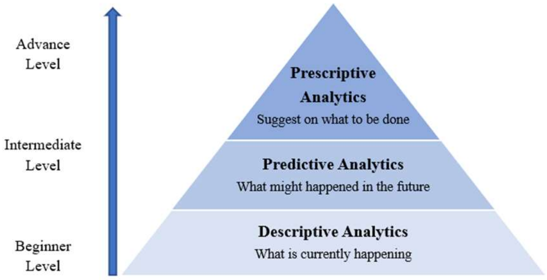


Figure 5: Types of Data Analytics

The first or beginner level is descriptive analytics. It interprets what is the current situation. The second or intermediate level is predictive analytics. It identifies what can happen in the future based on available historical data. The third or advance level is prescriptive analytics. It is to suggestions for actions that can improve the process or increase revenue [8]. Following is the further explanation of each data analytics.

2.2.1 Descriptive analytics

Descriptive model uses data aggregation to determine what had occurred, such as identifying the relationship between the data and describing past events [21]. By utilizing query and reporting tools, historical performance data can be gathered, classified, and categorized in a structural way [28]. This has been applied in many businesses and organizations. They have benefited from descriptive analytics. Some examples of the tools are Microsoft Power BI, Tableau, Jaspersoft and Qliksense [24]. It is mainly used for the top management to visualize, analyze and conduct strategic planning. A sample of Malaysia government dashboard that applies descriptive analytics is as in Figure 6, Malaysia Economic Recovery Dashboard.

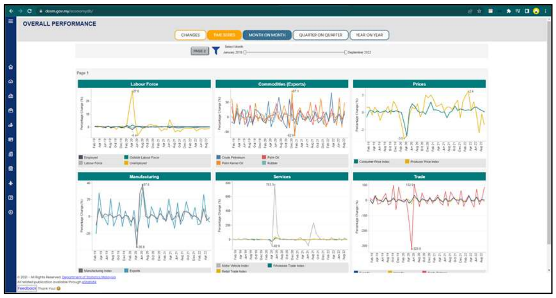


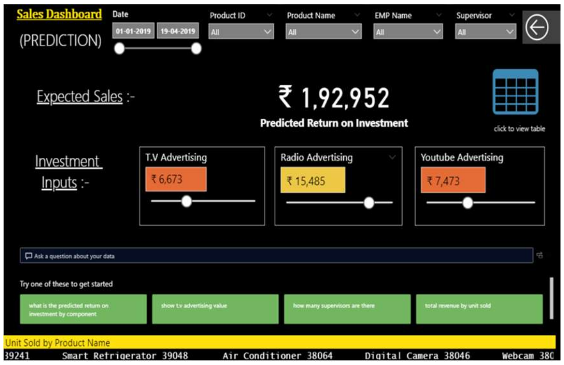
Figure 6: Malaysia Economic Recovery Dashboard [29]

Based on the 9 key indicators, it summarizes the current economic situation in Malaysia [29]. The key indicators include gross domestic product, labour force, commodities, prices, manufacturing, services, trade, balance of payments and others. By having the time series dataset, Malaysia Economic Recovery Dashboard able to show data by monthly, quarterly and annually. The next level to utilize the data is to do predictive analytics.

2.2.2 Predictive analytics

Predictive analytics can be done by using statistical and machine learning algorithms based on historical data to identify possible future, such as what may occur [21], [30]. It is helps businesses and organizations to solve complicated issues and come upon new opportunities. Predictive models are further subdivided into classification and regression models. Classification models determine categorical labels, which are definite and textual. On the other hand, regression models predict numerical values, which are empirical.

Predictive analytics has been applied in many areas such as medical, agriculture, human resource, and financial. Some examples are prediction of COVID-19 cases, agricultural demand such as rice and potatoes, employment trends, tax avoidance, loan approval and restaurant revenue. Despite the fact that some areas of this technology have yet to be explored, one of it is federal government revenue. This study is looking into this domain, which focuses on federal government revenue prediction using machine learning method. Example of sales prediction dashboard is shown in Figure 7.


Sales Dashboard (PREDICTION)

Date: 01-01-2019 – 19-04-2019

Product ID: All

Product Name: All

EMP Name: All

Supervisor: All

Expected Sales :- ₹ 1,92,952

Predicted Return on Investment

click to view table

Investment Inputs :-

T.V Advertising: ₹ 6,673

Radio Advertising: ₹ 15,485

Youtube Advertising: ₹ 7,473

Ask a question about your data

Try one of these to get started

what is the predicted return on investment by component

show t.v advertising value

how many supervisors are there

total revenue by unit sold

Unit Sold by Product Name

39241 Smart Refrigerator 39048 Air Conditioner 38064 Digital Camera 38046 Webcam 380

Figure 7: Sales Prediction Dashboard [31]

The dashboard is developed using Microsoft Power BI [31]. It applies multiple linear regression as the machine learning method. The programming language used is Python. By inserting the investment for tv, radio and YouTube advertisement, the predicted sales will be calculated and displayed. The dashboard also allows user to select the specific date and product to be predicted.

2.2.3 Prescriptive analytics

Prescriptive analytics utilizes optimization methods to suggest businesses or organization directions and their consequences [21]. It suggests what has to be done. Example as a doctor in a clinic, when a patient come due to any sickness, the doctor will examine patient. To ensure the patient will recover, the doctor will prescribe specific medication or treatment, for certain number of periods. Prescriptive analytics is the same, only that system or machine learning will do the job as a doctor to identify and analyze current situation of a business or organization, what might happen in the future and what needs to be done to get the best results.

2.3 Related Works on Financial Prediction

Revenue is defined as income derived from an organization's operation such sales and service incomes. For non-tax revenue, it is revenue obtained from other sources than tax. In this case study, prediction is done by [11] using back propagation feed forward neural network (FFNN), which is a type of artificial neural network (ANN). Levenberg – Marquardt and sigmoid bipolar are used as the training method and activation method respectively. For the dataset, it is in numeric and time series format. It consists of 8 years of dataset is in yearly frequency, 7 years are for training while the last year for testing. To ensure a smooth research, Cross Industry Standard Process for Data Mining (CRISP-

DM) is being implemented. This includes business understanding, data understanding, data preparation, modeling, evaluation and deployment. There are 2 things being experimented. Firstly, is for the data input to be partition or non-partition. Secondly, is the number of hidden layer neurons. Based on previous studies, number of neurons can be double of input neurons, 2n, two-thirds of input and output neurons, 2/3 (n+o), number of neurons plus or minus one, and so on. The best results obtained in the study is when the data is partitioned and number of hidden layer neurons is two-thirds of input and output neurons, 2/3 (n+o). It totals up to 7 hidden layer neurons. For Mean Squared Error (MSE) is used and the best result achieved is 0.00002059 of MSE.

Another research involving ANN is done to predict total assets for an Indonesian bank [4]. The ANN is applied to Auto-Regressive (AR) and Multi Input Single Output AR with external input (MISO-ARX). The time series dataset consists of total asset, net income and return on assets for 12 years. It is then normalized before being used. ANN used is also known as FFNN that applies back propagation. It adjusts the input and hidden layer weight by adding small random number to reduce the error. The settings of model include 10 hidden neurons and applying log-sigmoid for its activation function. Difference between AR and MISO-ARX is that AR only takes 1 variable while MISO-ARX takes multiple variables for prediction. Result shows that ANN MISO-ARX achieved 99.37% accuracy while ANN AR achieved 92.8% accuracy.

On study concluded that the most effective individual model overall was FFNN [6]. It demonstrated a strong directional analysis while comparing with autoregressive integrated moving average (ARIMA) and exponential smoothing model. Unfortunately FFNN does not always achieve higher accuracy. A house price prediction was done using few machine learning methods [32]. By comparing fuzzy logic, FFNN and k-NN, it shows that fuzzy logic able to get higher accuracy followed by FFNN and k-NN.

A stock market prediction was done using 5 approaches [33]. They are are statistical, pattern recognition, machine learning, sentiment analysis and hybrid approach. Looking into machine learning, there were few interesting findings. Firsly, for time series prediction, the Long Short-Term Memory (LSTM) network has shown a lot of potential. It also outperforms Recurrent Neural Network (RNN) and Gated Recurrent Unit (GRU). Secondly, multilayer FFNN able achieve good results by applying back propagation and Adam optimization algorithm. Thirdly, random forest beats several current prediction techniques in terms of accuracy and return on trade, and it is resilient to market volatility. However, here random forest is used in classification instead of regression prediction. Fourthly, to identify the direction of stock market, they were 2 methods mention which are the XGBoost and ANN. It is also commented that ANN even performs better than ARIMA and other statistical models. Lastly, when comparing deep learning with traditional machine leaning methods, there were one research mention that deep learning model able to get very low MSE which results in higher accuracy. But there were also another research reported that traditional machine learning method is better at evaluating the direction compared to deep learning.

RNN was chosen to predict fuel for next day as it is able to analyze the hidden relationship and non-linearity between variables [34]. It consists of 2 main steps that are forward propagation and backward propagation. During forward propagation, nonlinear activation function applied to internal layers while linear activation function applied to output layer. As for backward propagation, MSE is selected for loss function in RNN. Clipping is also implemented as a workaround to prevent gradient vanishing during backward propagation. Besides the barrel price, other inputs are international crude oil, foreign exchange rates and value added tax (VAT). Not all things are taken into account, for example, political and demand supply situations. By using 5 years of dataset, the RNN model able to predict 90% accuracy of next day fuel price and 80% accuracy of the next week fuel price.

For the LSTM model, it consists of 1 input layer, 3 hidden layers and 1 output layer. Adam optimization is used with random weigh initiation, linear activation function and epoch of 100. For each layer there will be hidden neurons. Here hyperparameter tuning is done by comparing 2, 4, 6 and 8 number of neurons. After calculating the root mean squared error (RMSE), it is found out the most optimal is hidden layer with 4 neurons for all revenue and commodity price datasets, coal, gold and oil. As for comparison with ARIMA, the model settings are (5,1,0) that stands for 0.1 coefficients, standard error below 1 and p value below 0. It is proven LSTM achieve better accuracy, as the average RMSE is 0.10 for both revenue and commodity price, while ARIMA achieve average RMSE of 0.34 and 0.42 for revenue and commodity price respectively.

Demand and revenue prediction was conducted for transportation and logistics in India [35]. Few supervised methods used and also a statistical method. They are linear regression, decision tree, random forest, bagging, boosting, mutivariate regression and ARIMA. Overall predicting revenue and demand, multivariate regression and random forest able to get the lowest RMSE which equals to higher accuracy.

Similarly Wang & Chen (2019) did similar research on different domain that is semiconductor suply-chain company. Both machine learning regression methods and statistical methods are compared. Multivariate adaptive regression splines, random forest, and support vector regression (SVR) are selected under the regression methods while self ARIMA, dynamic ARIMA and vector auto regression (VAR) are selected for statistical methods. Though the result is a bit different from previous research. For chip-design firm dataset, VAR and SVR performs better, while for chip P&T firms dataset, VAR and random forest performs better compared to the others. It shows that both machine learning regression and statistical method performs similarly.

A comparison of few machine learning methods involving decision tree, support vector machine (SVM), random forest, improved random forest was done [2]. The main objective is to do prediction on enterprise return on net assets. From the comparison in the sense of accuracy, improved random forest achieved the best accuracy followed by random forest, decision tree and SVM, with the MSE of 0.003, 0.014, 0.062 and 0.091 respectively. This shows base method that is random forest is able to get high accuracy compared to the others.

Predicting tax evasion for goods and service tax is carried out by applying linear regression as part of the process [37]. There are basically 3 steps. First is clustering analysis to identify dealers based on the correlation parameters. Then, Benford's analysis is done to identify the dealer is genuine or suspicious. Linear regression is then build based on genuine dealer. Lastly then same model is applied on suspicious dealer to predict their actual amount of tax they need to pay. Specifically for the linear regression model, the RMSE is 0.000411 which is a good result that shows no overfitting. The R-squared value is 0.937 which shows the model is not underfitting. This proves that linear regression is a good method to be used in this dataset or domain.

To fight tax evasion, Moroccan tax is predicted [20]. By comparing 2 methods that are linear regression and polynomial regression, the normalized root mean squared error (NRMSE) is the same, 25.81. But for coefficient of determination (COD), linear regression achieved better with 0.82 while polynomial achieved -4.07. While the approach in mass appraisal for South Korea residential property based on ordinary least squares (OLS) linear regression, the model's stability and accuracy are still in doubt [38]. It is then suggested to use random forest to do the prediction. Results also shown that random forest obtained higher accuracy, 94.5% compared to linear regression, 80%.

An extension to linear regression is multiple linear regression. Organic potato yield is predicted by comparing multiple linear regression and ANN model [5]. Tillage and soil characteristics were discovered to have a considerable impact on potato yield. More accurately than the ANN model, the multiple linear regression model predicted crop yield better. Through, the ability of the ANN model to predict the link between potato yield, tillage, and soil characteristics was more promising.

Though, best financial prediction can be done by using ANN, which shows better performance compared to regression analysis [39]. This was done in a study where few machine learning methods was compared for rice yield prediction. ANN overperformed regression, regression tree and ensemble method.

3. METHODOLOGY

This study applies CRISP-DM. It is selected because it has been used and recommended by other researchers as well. It is structured and organized, thus able to help researches to better conduct the experiment. CRISP-DM consists of 5 steps that are business understanding, data understanding, modelling, data processing and evaluation. To have a better understanding, Figure 8 visualizes the modelling activities or CRISP-DM flow.

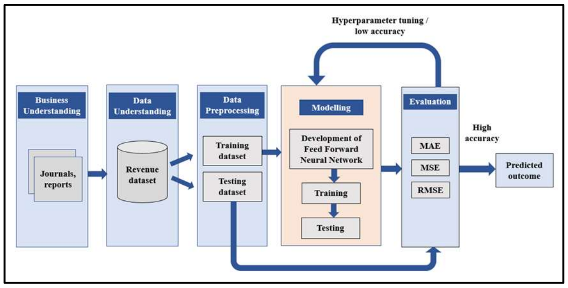


Figure 8: Modelling Activities / CRISP-DM Flow

3.1 Step 1: Business Understanding

Step 1 is business understanding. The background in connection with federal government revenue is studied, problem identified, and literature review is also conducted. The reason to study the structure of the Malaysia financial division, sources of federal government revenue and the usage is to better understand the importance and needs of the predictive model.

For literature review on the technical part, more than 40 research has been studied related to predictive analytics and mostly related to financial domain. The goal is to identify the commonly used method for financial or regression prediction, that able to produces higher accuracy. There are numerous machine learning methods that can be applied. In this study more than 15 methods are identified. This includes random forest, LSTM, linear regression and multiple linear regression. One of the highly suggested method is FFNN.

3.2 Step 2: Data Understanding

Next, the second step is data understanding. This is an important step in CRISP-DM as it gives high level overview of the dataset and the relationship between it [40]. By conducting data understanding, more information can be obtained about the dataset and appropriate action can be taken if needed. First dataset is acquired. Then the attributes, datatype and connectivity are identified. Data hierarchy is also constructed. This can give more understanding of the relationship and type of data that are being dealt with throughout the study. It can also give more idea on how to develop, train and test the model. Figure 9 visualizes the data hierarchy for Malaysia federal government revenue.

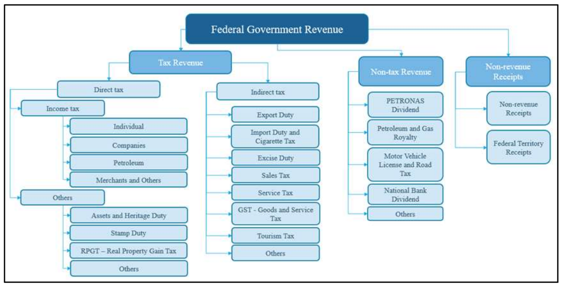


Figure 9: Malaysia Federal Government Revenue Data Hierarchy

The Malaysia federal government revenue dataset has 3 main categories which are tax revenue, non-tax revenue and non-revenue receipts. Under tax revenue, there are direct tax and indirect tax. Direct tax consist of income tax and others. Under each of the categories mention earlier are the subcategories.

The datasets of Malaysia federal government revenue are obtained from 2 sources. First is the Ministry of Finance Malaysia website and second is Department of Statistic Malaysia website. The first dataset consists of 47 years from 1970 until 2016. The second dataset consists of 11 years from 2008 until 2019. Merging of both datasets are done manually by comparing and matching the main and subcategories. After both are merged, the total datasets are 50 years, from 1970 until 2019. It is in yearly frequency. The dataset consists of 32 columns which includes year, main categories, subcategories, and total federal government revenue. There are 50 rows which is equivalent to 50 years of yearly dataset. The dataset is time-series based and contains numerical values.

3.3 Step 3: Data Preprocessing

In the third step, data preprocessing is done. The importance of data preprocessing is to ensure data analysis results are valid and reliable, which can be done for example by removing outliers and filling in missing values [41]. Basically, data cleaning, conversion and selection are carried out. After acquiring the dataset, it is then preprocessed by ensuring no null values and invalid dataset. It is also to select the right dataset for training and testing later on. The preprocessing is done manually in *comma-separated values* (*CSV*) file. This is because the dataset is small and easier to be done this way rather than automated using Python. Figure 10 shows the substeps involved in data preprocessing for this study.

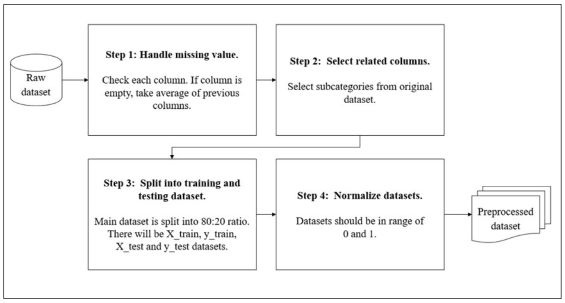

Raw dataset
Step 1: Handle missing value.
Check each column. If column is empty, take average of previous columns.
Step 2: Select related columns.
Select subcategories from original dataset.
Step 3: Split into training and testing dataset.
Main dataset is split into 80:20 ratio. There will be X_train, y_train, X_test and y_test datasets.
Step 4: Normalize datasets.
Datasets should be in range of 0 and 1.
Preprocessed dataset

Figure 10: Steps for Data Preprocessing

Firstly, missing values are handled. For null columns, the average for previous years is taken for that specific column. Secondly, related features or columns are selected. Here subcategories are selected for final dataset. This is also known as column-wise data variable reduction by using domain knowledge [41]. Thirdly, splitting is done to divide dataset into X and Y training dataset, and X and Y testing dataset. X is the independent variable while Y is the dependent variable. For the training and testing ratio used are 80:20. Shuffle option is set to false. Lastly, the datasets are normalized so that the dataset will be in range between 0 and 1.

For the dataset columns, it started with 32 columns which are year, main categories, subcategories, and total federal government revenue. To do the column selection or step 3 in data preprocessing, the main categories are removed. Only subcategories and total federal government revenue remains. Figure 11 shows the subcategories selected for federal government revenue dataset.

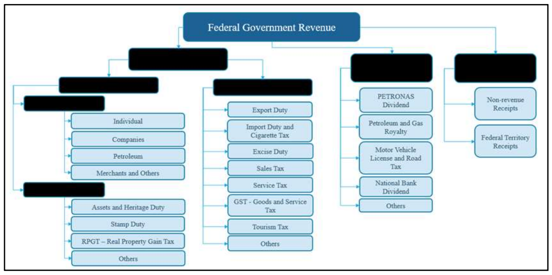


Figure 11: Subcategories Selected for Federal Government Revenue Dataset

3.4 Step 4: Modelling

The fourth step is modelling. It is a process of creating the predictive model to identify the pattern by using the acquired dataset [40]. Model development, training, and testing are carried out. For the programming language, Python is used in this study. This is because there are many readily available libraries can be used for data processing, development of predictive models and result analysis. Example of the libraries are NumPy, Pandas, Scikit Learn and Matplotlib. Besides that, execution can be faster.

FFNN is chosen as the machine learning method to be developed for the proposed model. This has been decided based on the commonly and highly used method in predictive analytics, which has been identified in step 1, business understanding. The FFNN model structure consists of 1 input layer, 1 hidden layer, and 1 output layer. As example FFNN structure is as shown in Figure 12.

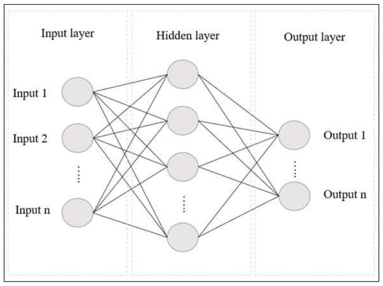


Figure 12: FFNN Structure

A normal flow will start from input layer to hidden layer and lastly output. Each node at input layer represents the input variable. The output layer will have node based on the output. In the case of this study, input layer will consist of the variables from the dataset of the federal government revenue. As for the output layer, there is only 1 which is the value of the federal government revenue.

For this study, the model applies hyperparameter tuning while having some fixed setting for the model. The fixed setting includes the model structure. There will be 1 input layer, 1 hidden layer and 1 output layer. The input layer will have 23 neurons which represents 23 input variables from dataset. The output layer will only have 1 neuron, which is the predicted federal government revenue. As the loss function, mean squared error is selected.

Table 1: Parameter Used for Hyperparameter Tuning

Optimizer	adam, rmsprop, sgd, adadelta, adagrad, adamax, nadam, ftrl
Activation function	softmax, softplus, softsign, relu, tanh, sigmoid, hard_sigmoid, linear
No of neurons (for hidden layer)	10, 20, 30, 40, 50, 100
No of Epochs	50, 100, 200, 300, 400, 500

Table 1 shows the parameter used for FFNN model hyperparameter tuning. This includes optimizer, activation function, number of neurons for hidden layer and number of epochs. For optimizers, there are Adam, RMSprop, Stochastic Gradient Descent (SGD), AdaDelta, AdaGrad, AdaMax, NAdam and FTRL. For activation function, Softmax, Softplus, ReLU, Tanh, Sigmoid, Hard Sigmoid and Linear are experimented. Number of neurons for activation function are in the range of 1 to 100. The selected ones are 10, 20, 30, 40, 50 and 100. Lastly the number of epochs chosen are 50, 100, 200, 300, 400 and 500.

The model is trained after it has been developed. It uses a training dataset. The model then picks up on the data pattern or correlation between the variables. The model is then tested using testing dataset. The model is used to do prediction and the predicted data is obtained. The evaluation process then compares it to the testing dataset.

3.5 Step 5: Evaluation

After training and testing are completed, evaluation is carried out. This step is to ensure the predictive model is in good balance, meaning it will not be overfitting or underfitting [40]. The evaluation methods used are mean absolute error (MAE), MSE and RMSE. Y testing dataset and predicted dataset are compared.

MAE is the average difference between predicted and actual results [42]. In other words, it is to identify how close the values of predicted and actual data [43]. The lower the MAE, the higher accuracy the predictive model is, while the higher the value of MAE, the more inaccurate the predictive model. The formula for MAE is as following:

$$\text{MAE} = \frac{\sum_{i=1}^{N}|\text{Actual}-\text{Predicte}\ \ |}{\text{N}} \quad (1)$$

For MSE, the difference values between actual and predicted are squared to remove the negative values. To evaluate, the higher the score the inaccurate the prediction is and vice versa. Formula of MSE is as below:

$$\text{MSE} = \frac{\sum_{i=1}^{N}(\text{Actual}-\text{Predic}\ \ \ )^2}{\text{N}} \quad (2)$$

As for RMSE, it is the square root of MSE. This will get back the actual values. Same as the other evaluation formulas, the lower the value the better the predictive model is [44]. RMSE formula is as following:

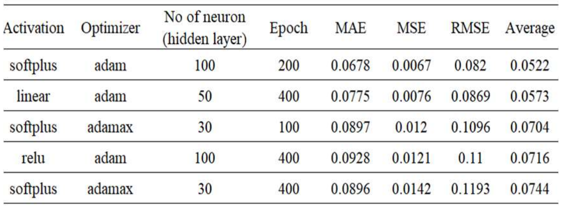
$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\text{Actual}-\text{Predic}\ \ \ )^2}{\text{N}}} \quad (3)$$

The evaluation is done for each hyperparameter tuning. This is to find the best settings for FFNN model to be used for financial data prediction specifically federal government revenue. Training, testing and evaluation is repeated based on the hyperparameter settings.

4. RESULTS AND DISCUSSION

A series of experiment done to find the optimal settings. It is also known as hyperparameter tuning. Activation functions, optimizers, number of neurons and number of epochs are the parameters involved. There are 8 activation functions, 8 optimizers, 6 numbers of neurons and 6 numbers of epochs used. In total there are 2304 parameter combinations for the experiment.

Table 2: Top 5 Results for Hyperparameter Tuning



Activation	Optimizer	No of neuron (hidden layer)	Epoch	MAE	MSE	RMSE	Average
softplus	adam	100	200	0.0678	0.0067	0.082	0.0522
linear	adam	50	400	0.0775	0.0076	0.0869	0.0573
softplus	adamax	30	100	0.0897	0.012	0.1096	0.0704
relu	adam	100	400	0.0928	0.0121	0.11	0.0716
softplus	adamax	30	400	0.0896	0.0142	0.1193	0.0744

After doing the hyperparameter tuning, the higher accuracy on in this case lowest average is identified. Table 2 is showing the top 5 optimal settings for FFNN that achieved higher accuracy using Malaysia's federal government revenue dataset. The first setting is using Softplus activation, Adam optimizer, 100 neurons for hidden layer and 200 epochs. It achieved 0.0678 MAE, 0.0067 MSE and 0.082 RMSE. On average it achieved 0.0522 for overall accuracy.

Second setting adapts linear activation function, Adam optimizer, 50 neurons and 400 epochs. Third setting applies Softplus activation function, Adamax optimizers, 30 neurons and 100 epochs. Fourth setting uses ReLU activation function, Adam optimizer with 100 hidden neurons and 400 epochs. Lastly, Softplus activation function is set along with Adamax optimizer, 30 hidden neurons and 400 epochs. Second, third, fourth and fifth settings achieved 0.0573, 0.0704, 0.0716 and 0.0744 respectively of overall accuracy. It can be seen that Softplus optimizers is occurring more for the activation function. Adam and Adamax also seems to be better choice for optimizer.

Table 3: Parameters for FFNN with Optimal Setting

Parameters	Values
Train Test	80:20
Dataset Description	X - all sub categories Y - total federal government revenue
Layers	input, 1 hidden layer, output
Activation Function	Softplus
Optimizer	Adam
No of neurons (hidden layer)	100
No of Epoch	200

The best settings are as shown in Table 3. For the dataset, it is split to 80:20 ratio for training and testing. The shuffle option is set to false. X consist of subcategories of Malaysia's federal government revenue while Y is the total of Malaysia's federal government revenue. General setting is it consists of 1 input layer, 1 hidden layer and 1 output layer. The input layer is having 23 neurons as this is the dependent variables or X values. The output layer is having 1 neuron as the independent variable or Y value. As for the specific settings from hyperparameter tuning, the model is using Softplus activation function, Adam optimizer, 100 number of neurons for hidden layer and 200 epochs for the model.

Figure 13 visualizes the actual values and predicted values of federal government revenue using the best settings. Blue line represents the actual values while orange line represents the predicted values. The years predicted are the 10 latest years from the dataset which are from 2010 to 2019.

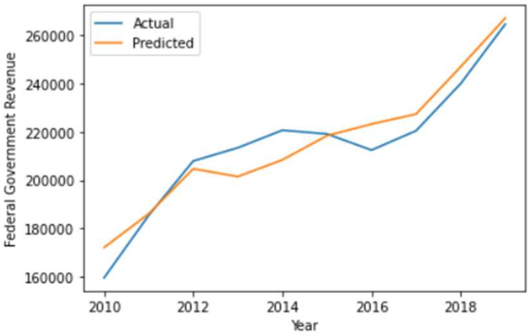


Figure 13: Actual vs Predicted Federal Government Revenue

Figure 14 shows the training and validation loss compared to number of epochs. As the number of epochs increases, the loss also decreases for both training and validation. Since validation loss did not increase towards 200 epochs, this means the model is not overfitting or underfitting.

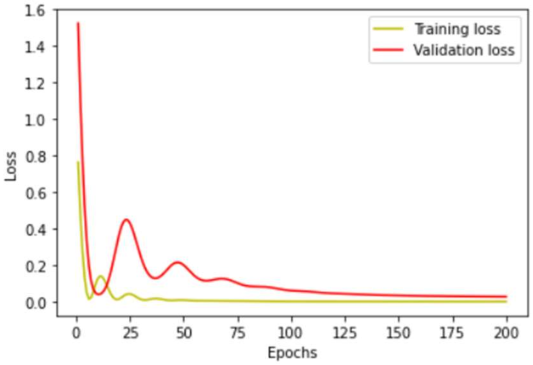


Figure 14: Loss vs Epoch Graph

Overall FFNN shows good results to predict federal government revenue. Though, the results of the parameters may vary during each run time of the model and also on using different datasets. Based on this study, it is suggested to use Adam for the optimization and Softplus for the activation function.

5. CONCLUSION

Predictive analytics using machine learning has benefited a lot of domains over forecasting. Though some domains have yet to explore it. This includes federal government revenue specifically in Malaysia that is currently using forecasting. Predictive analytics uses machine learning while forecasting uses statistical methods. Since there are studies that shows predictive analytics can achieve higher accuracy compared to forecasting, it is a good opportunity to explore predictive analytics for federal government revenue.

The first objective of this study is to identify the suitable predictive analytics methods that can achieve high accuracy for federal government revenue. Related works on financial prediction has been studied. There are over 15 methods identified. One of the highly suggested is FFNN.

The second objective is to identify optimal settings of the selected method. To do this, hyperparameter tuning is carried out. A series of training, testing and evaluation are done. To calculate the accuracy, MAE, MSE and RMSE are used as well as the average. Result shows that by using Adam optimizer and Softplus activation function with 100 neurons and 200 epochs can produce higher accuracy for the Malaysian federal government revenue prediction with the average of 0.0522 error. Overall, both objectives have been fulfilled.

For future work, other ANN machine learning method can be experimented such as RNN, LSTM and others. It is also recommended to use more datasets from other states or countries that have higher frequency such as quarterly instead of yearly. More datapoints will be better for machine learning training and testing process. Different dataset also results differently although using the same machine learning method and parameters.

Another recommendation is to consider external elements that may influence federal government revenue. For example, the interest rate, currency exchange, and economic situation locally and globally. Health situations such as COVID-19 can be included as well. By studying the external elements, specific elements that affects the federal government revenue can be identified. From there better plans can be conducted to increase the federal government revenue.

ACKNOWLEDGMENT

This research is supported by Yayasan Universiti Teknologi Petronas (YUTP) grant of cost center 015LC0-289.

REFERENCES:

[1] Ministry of Finance Malaysia, "2022 Fiscal Outlook and Federal Government Revenue Estimates," *Minist. Financ. Malaysia*, pp. 101–257, 2022, [Online]. Available: https://budget.mof.gov.my/2022/fiscal-outlook/

[2] Y. Cai, Q. Yin, Q. Su, X. Huang, Y. Zhang, and T. Liu, "Prediction Method of Enterprise Return on Net Assets Based on Improved Random Forest Algorithm," *J. Phys. Conf. Ser.*, vol. 1682, no. 1, 2020, doi: 10.1088/1742-6596/1682/1/012083.

[3] E. Mathew and S. Abdulla, "The lstm technique for demand forecasting of e-procurement in the hospitality industry in the uae," *IAES Int. J. Artif. Intell.*, vol. 9, no. 4, pp. 584–590, 2020, doi: 10.11591/ijai.v9.i4.pp757-765.

[4] Fariyanti, Iskandar, R. Malani, and B. Suprapty, "Total asset prediction of the large Indonesian bank using adaptive artificial neural network back-propagation," *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 75–79, 2018, doi: 10.14419/ijet.v7i2.2.12737.

[5] K. Abrougui, K. Gabsi, B. Mercatoris, C. Khemis, R. Amami, and S. Chehaibi, "Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR)," *Soil Tillage Res.*, vol. 190, pp. 202–208, Jul. 2019, doi: 10.1016/j.still.2019.01.011.

[6] J. Creighton and F. H. Zulkernine, "Towards building a hybrid model for predicting stock indexes," *Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017*, vol. 2018-Janua, pp. 4128–4133, 2017, doi: 10.1109/BigData.2017.8258433.

[7] T. D. Oesterreich, E. Anton, and F. Teuteberg, "What translates big data into business value? A meta-analysis of the impacts of business analytics on firm performance," *Inf. Manag.*, vol. 59, no. 6, p. 103685, 2022, doi: 10.1016/j.im.2022.103685.

[8] E. Ahmed *et al.*, "The role of big data analytics in Internet of Things," *Comput. Networks*, vol. 129, pp. 459–471, 2017, doi: 10.1016/j.comnet.2017.06.013.

[9] Ministry of Finance Malaysia, "Fiscal Outlook and Federal Government Revenue Estimates 2021," 2020, [Online]. Available: https://www.treasury.gov.my/index.php/en/fiscal-economy/fiscal-outlook-2021.html

[10] V. A. Ahnaf and Suharjito, "Financial revenue prediction model in mining industry using deep learning," *ICIC Express Lett.*, vol. 13, no. 11, pp. 987–993, 2019, doi: 10.24507/icicel.13.11.987.

[11] F. A. Lubis and Albarda, "Data Partition and Hidden Neuron Value Formulation Combination in Neural Network Prediction Model," 2018.

[12] Chimilila and Cyril, "Forecasting Tax Revenue and its Volatility in Tanzania," *African J. Econ. Rev.*, vol. 5, no. 1, pp. 84–109, 2017.

[13] B. Zhao, "Forecasting the New England States' Tax Revenues in the Time of the COVID-19 Pandemic," 2020, [Online]. Available: https://ssrn.com/abstract=3648649

[14] L. Dadayan, "State Revenue Forecasts Before Covid-19 and Directions Forward," pp. 1–9, 2020, [Online]. Available: https://www.urban.org/policy-centers/cross-center-initiatives/state-and-local-finance-

[15] H. Shahnazarian, M. Solberger, and E. Spånberg, "Forecasting and Analysing Corporate Tax Revenues in Sweden Using Bayesian VAR Models," *Finnish Econ. Pap.*, vol. 28, no. 1, pp. 50–74, 2017.

[16] M. Samsudin, A. M. Dali, R. Hasan, and I. Saidoo, *Sejarah Tingkatan 5 - BTBP*, 1st ed. Dewan Bahasa dan Pustaka, 2020.

[17] M. Department of Statistics, "Laporan Sosioekonomi Negeri Perak 2019," 2020.

[18] Ministry of Finance Malaysia, "Economic Report 2017/18," *Financ. Stat.*, 2018, doi: 10.1057/fs.2010.78.

[19] M. of F. Malaysia, "Economic Outlook 2023," 2022, [Online]. Available: https://budget.mof.gov.my/pdf/2023/economy/economy-2023.pdf

[20] H. Jihal, M. A. Talhaoui, A. Daif, and M. Azzouazi, "Predictive Analytics as A Service on Moroccan Tax Evasion," *Artic. Int. J. Eng. Technol.*, vol. 7, no. April, pp. 90–92, 2018, [Online]. Available: https://www.researchgate.net/publication/332353664

[21] S. K. Babu and S. Vasavi, "Predictive

analytics as a service on tax evasion using Gaussian Regression process," *Helix*, vol. 7, no. 5, pp. 1988–1993, 2017, doi: 10.29042/2017-1988-1993.

[22] V. H. Trieu, "Getting value from Business Intelligence systems: A review and research agenda," *Decis. Support Syst.*, vol. 93, pp. 111–124, 2017, doi: 10.1016/j.dss.2016.09.019.

[23] Z. Sun, L. Sun, and K. Strang, "Big Data Analytics Services for Enhancing Business Intelligence," *J. Comput. Inf. Syst.*, vol. 58, no. 2, pp. 162–169, 2018, doi: 10.1080/08874417.2016.1220239.

[24] K. Gowthami and M. R. P. Kumar, "Study on Business Intelligence Tools for Enterprise Dashboard Development," *Int. Res. J. Eng. Technol.*, vol. 4, no. 4, pp. 2987–2992, 2017, [Online]. Available: https://www.irjet.net/archives/V4/i4/IRJET-V4I4721.pdf

[25] M. A. Sheikh, A. K. Goel, and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," in *2020 International Conference on Electronics and Sustainable Communication Systems, ICESC 2020*, 2020, pp. 490–494. doi: 10.1109/ICESC48915.2020.9155614.

[26] L. Lai, "Loan Default Prediction with Machine Learning Techniques," in *2020 International Conference on Computer Communication and Network Security, CCNS 2020*, 2020, pp. 5–9. doi: 10.1109/CCNS50731.2020.00009.

[27] A. Aboud and B. Robinson, "Fraudulent financial reporting and data analytics: an explanatory study from Ireland," *Account. Res. J.*, vol. 35, no. 1, pp. 21–36, 2022, doi: 10.1108/ARJ-04-2020-0079.

[28] W. Raghupathi and V. Raghupathi, "Contemporary business analytics: An overview," *Data*, vol. 6, no. 8, pp. 1–11, 2021, doi: 10.3390/data6080086.

[29] Department of Statistics Malaysia Official Portal, "Malaysia Economic Recovery Dashboard." https://www.dosm.gov.my/economydb/ (accessed Nov. 12, 2022).

[30] N. Radhakrishnan, M. Awasthi, and P. Mahalakshmi, "A survey on predictive analysis in employment trends," *Int. J. Eng. Technol.*, vol. 7, no. 2, pp. 358–360, 2018, doi: 10.14419/ijet.v7i2.24.12082.

[31] M. Raje, P. Jain, and V. Chole, "Sales Analysis and Prediction Dashboard Using Power Bi," no. 06, pp. 522–529, 2021, [Online]. Available: www.irjmets.com

[32] M. F. Mukhlishin, R. Saputra, and A. Wibowo, "Predicting house sale price using fuzzy logic, Artificial Neural Network and K-Nearest Neighbor," *Proc. - 2017 1st Int. Conf. Informatics Comput. Sci. ICICoS 2017*, vol. 2018-Janua, pp. 171–176, 2017, doi: 10.1109/ICICOS.2017.8276357.

[33] D. Shah, H. Isah, and F. Zulkernine, "Stock market analysis: A review and taxonomy of prediction techniques," *Int. J. Financ. Stud.*, vol. 7, no. 2, 2019, doi: 10.3390/ijfs7020026.

[34] L. M. Chaitanya, D. H. Ravi, and R. Bharathi, "Fuel Price Prediction Using RNN," *2018 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2018*, pp. 1510–1514, 2018, doi: 10.1109/ICACCI.2018.8554642.

[35] N. Darapaneni, S. Muthuraj, K. Prabakar, and M. Sridhar, "Demand and revenue forecasting through machine learning," *Proc. 2019 IEEE Int. Conf. Commun. Signal Process. ICCSP 2019*, pp. 328–331, 2019, doi: 10.1109/ICCSP.2019.8698011.

[36] C. H. Wang and J. Y. Chen, "Demand forecasting and financial estimation considering the interactive dynamics of semiconductor supply-chain companies," *Comput. Ind. Eng.*, vol. 138, no. September, p. 106104, 2019, doi: 10.1016/j.cie.2019.106104.

[37] J. Mathews, P. Mehta, S. Kuchibhotla, D. Bisht, S. B. Chintapalli, and S. V. K. Visweswara Rao, "Regression Analysis towards Estimating Tax Evasion in Goods and Services Tax," *Proc. - 2018 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2018*, pp. 758–761, 2019, doi: 10.1109/WI.2018.00011.

[38] J. Hong, H. Choi, and W. S. Kim, "A house price valuation based on the random forest approach: The mass appraisal of residential property in south korea," *Int. J. Strateg. Prop. Manag.*, vol. 24, no. 3, pp. 140–152, 2020, doi: 10.3846/ijspm.2020.11544.

[39] U. Inyaem, "Construction Model Using Machine Learning Techniques for the Prediction of Rice Produce for Farmers," *2018 3rd IEEE Int. Conf. Image, Vis. Comput. ICIVC 2018*, pp. 870–874, 2018, doi: 10.1109/ICIVC.2018.8492883.

[40] V. Kotu and B. Deshpande, "Data Science:

Concepts and Practice, Second Edition," 2019.

[41] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Frontiers in Energy Research*, vol. 9. Frontiers Media S.A., Mar. 29, 2021. doi: 10.3389/fenrg.2021.652801.

[42] J. Y. Kuo, H. C. Lin, and C. H. Liu, "Building graduate salary grading prediction model based on deep learning," *Intell. Autom. Soft Comput.*, vol. 27, no. 1, pp. 53–68, 2021, doi: 10.32604/iasc.2021.014437.

[43] A. Kibekbaev and E. Duman, "Benchmarking regression algorithms for income prediction modeling," *Inf. Syst.*, vol. 61, pp. 40–52, 2016, doi: 10.1016/j.is.2016.05.001.

[44] M. Kaur, M. Dhalaria, P. K. Sharma, and J. H. Park, "Supervised machine-learning predictive analytics for national quality of life scoring," *Appl. Sci.*, vol. 9, no. 8, 2019, doi: 10.3390/app9081613.