College Dublin
Computing • IT • Business

Assessment Cover Page

Student Full Name	CHAN CHEE XIANG
Student Number	2024463
Module Title	Strategic Thinking
Assessment Title	CA3 - Analysis of Malaysia New Born Baby Dataset
Assessment Due Date	2025-05-18
Date of Submission	2025-05-12

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Analysis of Malaysia New Born Baby Dataset Report

Table of Contents

Video Presentation:
https://drive.google.com/file/d/1AfmDRb9ZIAsOxgLgrGNc1_ySa3I1_0PE/view?usp=sharing

Analysis of Malaysia New Born Baby Dataset Report

Introduction

Motivation

Tracking newborn birth trends is widely important for country planning, resource arrangement, and policy-making in several sector like healthcare, education, and social services. Birth rates also affected by many factors, including economic conditions, population changes, government policies, and cultural beliefs. Predicting birth numbers may help governments and some organisations to prepare future needs, such as building hospitals, training doctors, and designing support programs.

Additionally, Malaysia is a diverse and growing country, studying birth trends statistic could also help government to understand the population changes and regional differences. Traditional statistical methods may not fully capture the trend patterns in the data. However, Machine learning can help by analyse large datasets to find hidden gems to improve predictions and forecasting. This report is aim to examines how machine learning can be used to predict birth data by using the dataset from Malaysia's Department of Statistics (DOSM), and given different insights from each sector for decision-making by the government

Research Problem

Predicting birth numbers are difficult because many factors would affect them, such the economy, government policies, cultural practices, and environmental changes. Traditional statistical methods like linear regression prediction are most rely on the simple mathematic rules. However, these methods are often failed to predict because the birth trends are complicated and it change over the time.

The Malaysia newborn baby dataset includes the data lie state, date (in yearly), sex, and ethnicity of the baby. While this data is helpful, but linear regression prediction is struggle to analyse properly. In the other way, advance machine learning methods like Random Forest Regression, Support Vector Regression (SVM or SVR) and ARIMA could improve accuracy by automatically to find complex patterns in the data. However, it was unclear how well those methods would work for prediction and whether the results would be easy to understanding.

The challenge of this research is finding methods that not only to predict future birth numbers accurately but also have to explain past trends clearly. Solving this would help governments planners for future needs

Analysis of Malaysia New Born Baby Dataset Report

Research Questions

This project aims to investigate below questions about predict birth count numbers in the Malaysia by using machine learning:

1. How well machine learning methods predict birth numbers by using Malaysia's historical data organized by state, date, sex, and ethnicity?
2. Which of these models works the best while measure the accuracy by using standard tests in Mean Squared Error (MSE), R-squared (R^2), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE)?
3. What are the main advantages and possible problems by using advanced machine learning algorithms instead of traditional statistical methods for birth predictions with Malaysia's newborn data?

Research Objective

The objective of this research is to conduct the analysis by using the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework. This involves several goals:

1. **Exploratory Data Analysis (EDA):** Examine the dataset to understand the structure, contents, feature correlation and some basic statistics. This includes identifying trends over time, differences across states, sexes and ethnic groups.
2. **Data Processing:** The method of handle the data and preprocess for modelling. This includes handle missing values, column transforme, encode categorical variables, and create new features to facilitate modelling.
3. **Model Development:** Apply multiple machine model techniques to the data for predict the number of birth (abs). These include:
    a) The regression model - Linear Regression.
    b) The ensemble tree-based model - Random Forest Regression to capture non-linear relationships.
    c) Support Vector Machine (SVM) for regression to evaluate performance with a different algorithmic approach.
    d) ARIMA model for time-series forecasting to predict future birth counts.

    Each model will be train and run hyperparameter tuning via cross-validation and optimize to ensure fair performance comparison.
4. **Model Evaluation and Comparison:** Evaluate models by using appropriate metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 to compare their performance. This objective is to identify which model are the best fits for historical data by using a train-test split function. This will directly test the hypothesis on whether which model performs best on unseen data

5. **Documentation and Ethical Considerations:** Document the entire process following CRISP-DM methodology and discuss some ethical considerations. Additionally, outline how the research was structured and how it could contribute to the broader knowledge base or policy-making efforts.

To achieve these objectives, the thesis is structured as follows as provides some background info for the algorithms and related studies in this domain. Describes the methodology in detail, by following the CRISP-DM process to presents the results of the exploratory analysis and the performance on each model, discussion and interpret of research findings. Concludes the report by summarizing answers from the research question, discussing ethical considerations and the project timeline.

Literature Review and Related Work

Technology and Algorithms

This capstone project is performing machine learning methods guided by the CRISP-DM framework. The CRISP-DM defines six-phase approach to any data mining project, that is : Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. The process is repetitive and flexible rather than moving strictly step-by-step because analysts frequently look back on the previous steps and discovered in later stages.

By using CRISP-DM method, the structured approach begins by understanding the project goals, then organizing the data, followed by creating models, and finally checking if the outcomes meet the original objectives. This framework is now one of the most popular in data mining methodology because it works with every industry.

Several algorithms shall apply into this capstone project and corresponding to the modelling are:

1. **Linear Regression:** Linear regression is a basic statistical method for understanding how a single outcome relates to one or more variables. It works by assuming a straight line connect between the input factors with the result. Because it is simple and easy to explain, when it is often used as a starting point for analysis (James et al., 2013). It can also work well for many real-world problems, it may can fail to capture patterns if the data has non-linear trends or complicated relationships between variables.

2. **Random Forest Regression:** Random Forests is a model that combine many decision trees during training and make predictions by average results from all these trees. It is popular because it is very accurate and able to avoids overfitting by using randomness in selecting data and features, it can also handle complex patterns in the data without adjust the data scale. It can identify which factors are the most important for predictions, helping users understand what influences outcomes (Breiman, 2001).

3. **Support Vector Machine (SVM):** Support Vector Regression (SVR) is a version of Support Vector Machines (SVM). When it try to create a prediction line that stay close to the actual data points it keeps the line as simple as possible. SVR works well with data that

has many features and can handle complex patterns using special mathematic functions named "kernels". However, its success often relies on choosing the right settings for values like C and the margin size (epsilon) (Drucker et al., 1997).

4. **ARIMA (AutoRegressive Integrated Moving Average):** ARIMA is a classic method for predicting future data that changes over time. It combines three key elements, that is : autoregression (AR/p) are using past value to predict future data, differencing (I/d) are removing trend by looking at the past changes, and moving average (MA/q) are using past errors to improve predictions.

   By mixing all these parts, ARIMA can manage trends and unstable patterns in the data. It creating a simple formula that links with current predictions to both past data and past mistakes.

5. In IBM, ARIMA Model Researchers have successfully applied ARIMA to population and birth-rate data since the 1970s, especially when studying a single series with clear trends or seasons. In our study, I will use ARIMA to model for year-to-year birth prediction, which complements our machine-learning models. Unlike other models, ARIMA explicitly for time-series patterns such as autocorrelation. Additionally, I will manually select the best p, d, and q settings to forecast future births value.

Analysis of Malaysia New Born Baby Dataset Report

Related Work

A recent study have also used Malaysia's public birth data from 2000 to 2023 to apply machine learning to predict birth trends. Ramli (2025) tested different models, including Linear Regression, Random Forest, Prophet, and XGBoost. The results showed that XGBoost can handle more complex model, worked better than the simple linear regression for predicting birth count.

The study also tried to predict a baby's sex using only year and ethnicity data, but the models was failed in the report, it prove that without more relevant factors, guessing a baby's gender is no better than a 50 - 50 chance.

Vital statistic 2023 reports from Malaysia's Department of Statistics (DOSM) confirmed that birth rates are falling. This shows why predictive models are useful. They can track how fast the decline is happening and predict if it will become stable in the future.

Additionally, research on China's birth rates tested via traditional methods like ARIMA and linear regression. These methods work well for short-term forecasts because they follow trends over time. However, the method usually study only one overall trend at a time.

In other point of view, machine learning models can use multiple factors like location or ethnicity to analyse many trends at once. This could lead to better predictions since it use more data. For instance, a past study might found some adding demographic details to machine learning models, it might can improve accuracy compared to just looking at national totals.

Our research follows a similar approach to compare different models to see which works best for predicting birth trends.

In summary, the literature suggests that:

1. There is a known downward trend in Malaysian birth counts in the decades.
2. ARIMA and other time-series models are commonly used for forecasting for birth or population.
3. Machine learning models are a newer approach showing in prediction accuracy for similar datasets.
4. Classification in this domain such as predicting an attribute like sex is might be unsuccessful without others related features.

Analysis of Malaysia New Born Baby Dataset Report

Methodology

The methodology outline of the project is following the CRISP-DM framework. Firstly, to give an overview of the project aim and solution approach, then delve into the specific dataset, data preprocessing, modelling techniques, and evaluation criteria.

The aim is to analyse and model Malaysia's annual newborn births data (2000–2023) to identify key patterns and build predictive models for birth counts. To approach this by following a structured firstly, understanding and preparing the data, then applying several machine learning as well as a time-series, and finally evaluating which approach are the best performance. The overall architecture of the analysis can be visualized as follows.

This project uses an iterative cycle of business and data understanding, data preprocessing and analysis, model training, model evaluation, and deployment, and it will repeat these phases multiple times to refine our models and improve from findings.

Data Understanding

```
In [3]: # Understand the dataset, discovered missing value (abs)
        df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7854 entries, 0 to 7853
Data columns (total 5 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   state      7854 non-null   object
 1   date       7854 non-null   object
 2   sex        7854 non-null   object
 3   ethnicity  7854 non-null   object
 4   abs        7790 non-null   float64
dtypes: float64(1), object(4)
memory usage: 306.9+ KB
```

```
In [149]: df.head(10)

Out[149]:
```

date	state	sex	ethnicity	abs
2000-01-01	Johor	both	overall	67183.0
2000-01-01	W.P. Labuan	female	bumi_malay	299.0
2000-01-01	Melaka	female	other_noncitizen	20.0
2000-01-01	Johor	both	bumi_other	867.0
2000-01-01	Kedah	both	bumi_other	79.0
2000-01-01	Kelantan	both	bumi_other	539.0
2000-01-01	Melaka	both	bumi_other	185.0
2000-01-01	W.P. Labuan	male	bumi_malay	350.0
2000-01-01	Melaka	male	other_noncitizen	30.0
2000-01-01	Pahang	both	bumi_other	1987.0

Figure 1 : Data Understanding by df.info() and df.head()

The dataset provides yearly record of live births across Malaysia from 2000 to 2023, disaggregated by state, sex, and ethnicity.

Each entry includes the following details:

1. state: Births are categorized by location, cover 13 states and 3 Federal in Malaysia
2. date: Represent in *YYYY-MM-DD* format
3. sex: Gender recorded as male, female, or "both" .
4. ethnicity: Classified of the ethnic into seven groups by Malay Bumiputera (bumi_malay), Other Bumiputera (bumi_other), Chinese, Indian, other citizens, and non-citizens (other_noncitizen), along with an "overall" category for combined totals.
5. Live Birth Count (abs): The absolute number of births for each demographic combination.

From Figure 1 clearly to shows that, the overall file size is 306.9+ KB, in the file contains 7854 entries, and 5 features which is 4 features are objects datatype and 1 feature is float datatype. The entries of "abs" features are only 7790, meaning that 64 entries were missing.

Analysis of Malaysia New Born Baby Dataset Report

Data Preprocessing and Analysing

Before building any models, the data preprocessing was performed to clean and transform from the raw data into a form suitable for analysis. The steps are outlined below:

1. Data Cleaning

```
In [7]: # After I perform df[].unique() for sex and unique, I clearly confirmed that sex = both and ethnicity = overall are duplicate dat
        # Hence I decide to drop dataset where contain sex="both" or ethnicity="overall"
        # Clean the dataset by dropping rows where sex="both" or ethnicity="overall"
        df_cleaned = df[(df['sex'] != 'both') & (df['ethnicity'] != 'overall')].copy()

        # Check the shape after cleaning and shows the comparison result
        print("Original dataset shape:", df.shape)
        print("Cleaned dataset shape:", df_cleaned.shape)
        print(f"Toral Removed {df.shape[0] - df_cleaned.shape[0]} rows")

Original dataset shape: (7854, 5)
Cleaned dataset shape: (4488, 5)
Toral Removed 3366 rows
```

Figure 2: Data Cleaning where drop sex contains both and ethnicity contains overall

Figure 2 shows, the code to dropped aggregate entries (sex = both and ethnicity = overall) to avoid duplication. This cleaning has removed 3,366 redundant rows. Hence, the dataset went from 7,854 to 4,488 rows.

2. Filling Missing Value

```
In [9]: # Group by state, sex, and ethnicity to calculate median values (I perfrom group median it will more precise )
        grouped_medians = df_cleaned.groupby(['state', 'sex', 'ethnicity'])['abs'].transform('median')
        # Fill missing values with group medians
        df_cleaned['abs'] = df_cleaned['abs'].fillna(grouped_medians)
```

Figure 3: Filling "abs" missing value by grouped of state, sex and ethnicity

The dataset included 64 missing entries (out of 7,854 records), rather than removing these records or excluding the dataset, which would risk losing valuable information from other categories. Therefore, the missing values were addressed through grouped median imputation. This reason to perform this method due to ensure the data remained usable for analysis while preserving consistency with historical patterns. By imputing rather than deleting, the integrity of the dataset was maintained, and no records were discarded unnecessarily.

Analysis of Malaysia New Born Baby Dataset Report

3. Encoding and Transformation

```
In [28]: # Identify categorical into numerical by OnehotEncoder and filling missing number from the most frequent appears categorical
         # because it is the fastest way and would not make the graph skew to one side.
         # I create pipeline to run imputer and one hot encoder (from class 29/4/2025)
         #https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html
         categorical_cols = ['state', 'sex', 'ethnicity']

         # Build Pipeline preprocessor
         preprocessor = ColumnTransformer(transformers=[
             ('cat', Pipeline([
                 ('imputer', SimpleImputer(strategy='most_frequent')),
                 ('onehot', OneHotEncoder(handle_unknown='ignore')) # Convert back to Numeric
             ]), categorical_cols)
         ])

         preprocessor
```

Figure 4: Create Pipeline to filling categorical missing value in most frequent and convert to numeric by One-Hot Encoder

For the machine learning models from scikit-learn, it needed to turn categorical variables into numeric form. So, set up a preprocessing pipeline using scikit-learn's ColumnTransformer, then imputed missing with most frequent just for a safety step, and then applied One-Hot Encoding. One-Hot Encoding can convert each categorical feature into multiple binary (0 or 1) columns, representing the presence or absence of each category. The (handle_unknown='ignore') parameter ensures that if unknown categories appear during transformation, they are ignored rather than causing an error.

4. Data Splitting and Target Transformation

```
In [26]: # Feature engineering: extract year, month, quarter, I think of extract it first maybe later could present a nice visualization r
         df_cleaned['year'] = df_cleaned['date'].dt.year
         df_cleaned['month'] = df_cleaned['date'].dt.month
         df_cleaned['quarter'] = df_cleaned['date'].dt.quarter
         df_cleaned['log_abs'] = np.log1p(df_cleaned['abs'])

         # Difrenciate features and targets
         X = df_cleaned.drop(['abs', 'log_abs', 'date'], axis=1)
         y = df_cleaned['abs']
         y_log = df_cleaned['log_abs']

         # Train-test split into Train 80% and Test 20%
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42) # from Original abs
         _, _, y_log_train, y_log_test = train_test_split(X, y_log, test_size=0.2, random_state=42) # from Log-Transformed abs

         print(f"\nTraining shape: {X_train.shape}")
         print(f"Testing shape: {X_test.shape}")

         print(f"y_log Training shape: {y_log_train.shape}")
         print(f"y_log Testing shape: {y_log_test.shape}")

         Training shape: (3590, 6)
         Testing shape: (898, 6)
         y_log Training shape: (3590,)
         y_log Testing shape: (898,)
```

Figure 5: Create log-transformation data frame and split into train/test

Feature Engineering: Extracted month and quarter, though in this dataset all births are aggregated yearly. These might not add any informational value for modelling since there is no within-year variation given, but they were created for a potential use.

Train-Test Split: Split the data into a training set and testing set to allow for evaluation of model generalization. I perform a standard 80/20 split (80% training, 20% testing), stratified randomly to split with a fixed random seed for reproducibility. The reason is wanted to evaluate the models' ability to predict unseen combinations in general, not just for the future.

Log Transformation: Due to the right-skewed distribution on the target variable 'abs', a log transformation was applied to make it more normalize distributed. Created log_abs which is the log (specifically in natural log of abs+1) of the birth counts to makes the distribution more

normal and can help some models fit better. Models were trained on both original and log-transformed scales.

5. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed to understand the distribution and relationships within the data. Visualizations and findings are present as below:

i) Total births by year to verify the overall birth trend

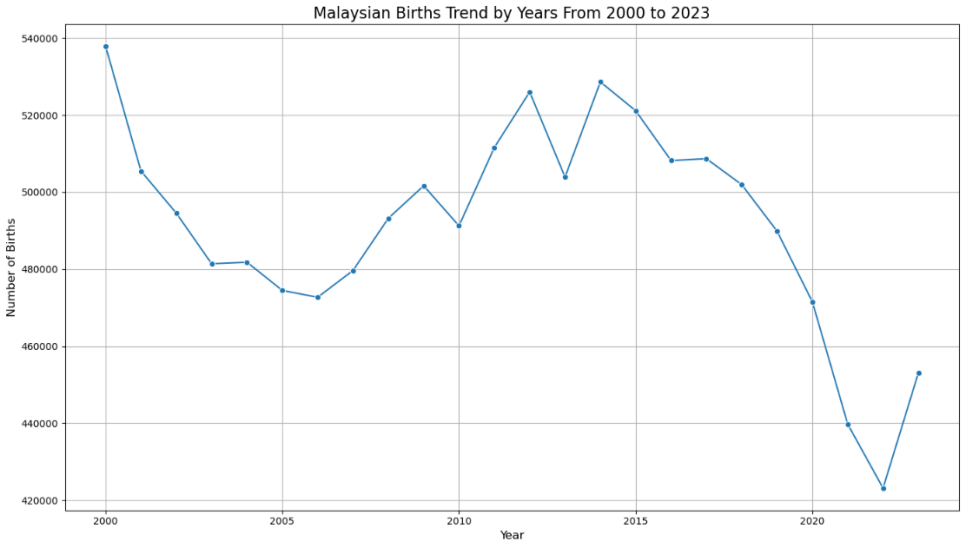


Figure 6: Malaysian Births Trend by Years From 2000 to 2023

The total number of live births per year in Malaysia has been steadily declining over the period. Figure 6 shows the total births by year, summed across all states with both sexes and all ethnicities. The trend starts around the year 2000 with the highest values and shows a continuous downward slope towards 2023。

ii) Total births by state to see which states contributed the most births

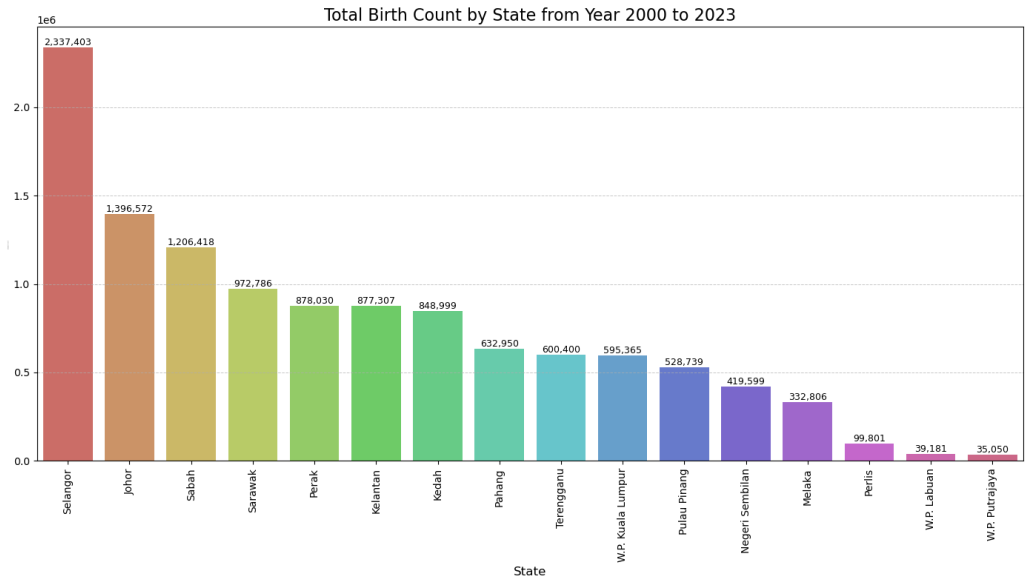


Figure 7: Total Birth Count by State from Year 2000 to 2023

Figure 7 shows the total births by state revealed that Selangor had the highest number of births in overall, followed by other highly populated states like Johor, Sabah, Sarawak, and Perak. Selangor is the highest number expected since it has the most populous state. Other states like W.P. Putrajaya and W.P. Labuan had the lowest totals because they are small territories.

iii) Total births by ethnicity and by sex to observe the demographic composition

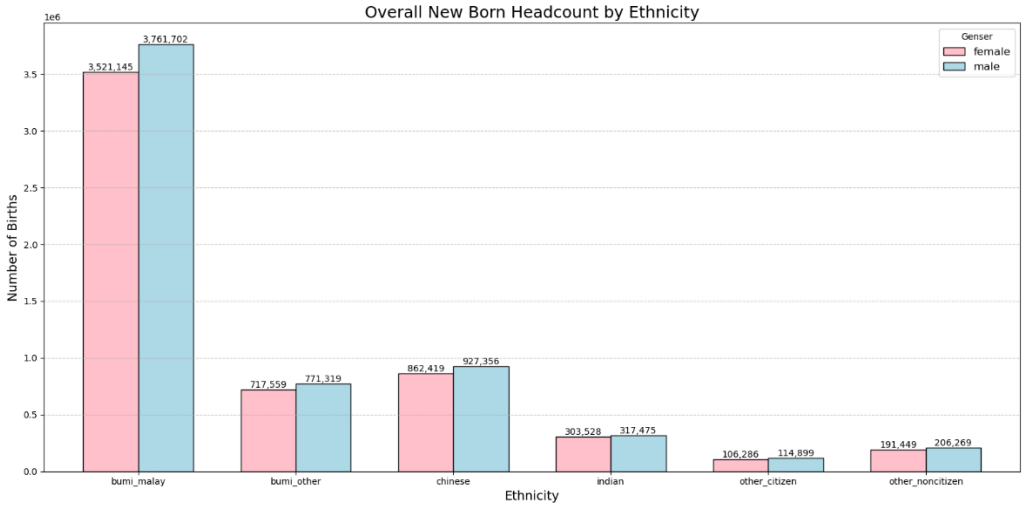


Figure 8: Overall New Born Headcount by Ethnicity

Breaking down the total births into ethnic group by summing over years and states for each ethnic category to shows the composition of births by gender. The largest share of births was by the bumi group (bumi_malay and bumi_other), followed by chinese and Indian. This reflects the ethnic composition of Malaysia's population Bumiputera being the majority. The other_citizen and other_noncitizen categories contribute a very small fraction of births.

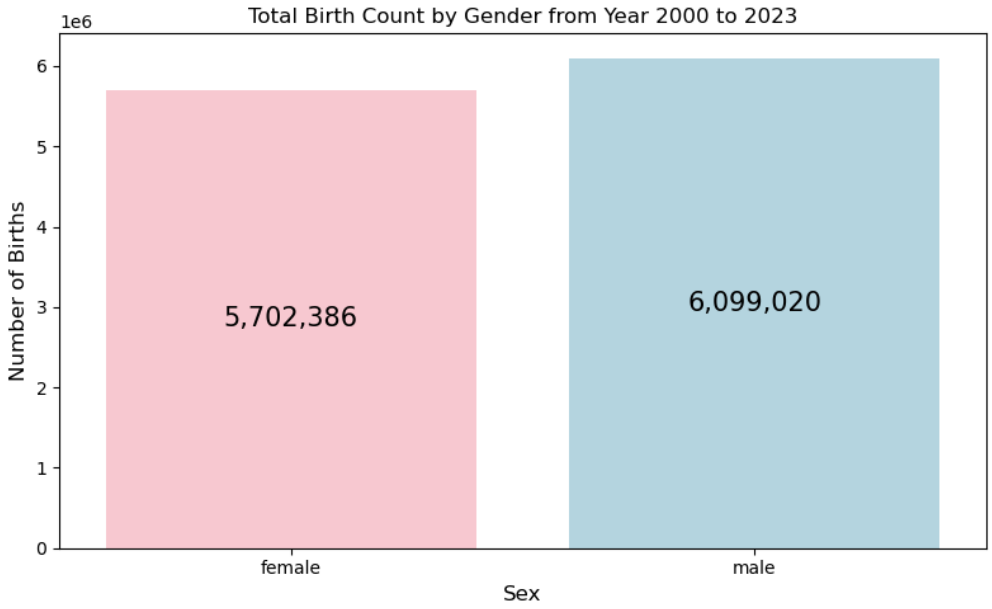


Figure 9: Total Birth Count by Gender from Year 2000 to 2023

Overall, there is a male majority in births. In the data, about 6.0 million (51.68%) of the births are male and 5.7 million (48.34%) were female, this is consistent with the biological norm of sex ratio at birth around 105 males per 100 females. Plotting total newborn headcount by sex showed males slightly above females.

iv) Distributions of the "abs" values which showed a right-skew typical of count data, hence justifying in log transform

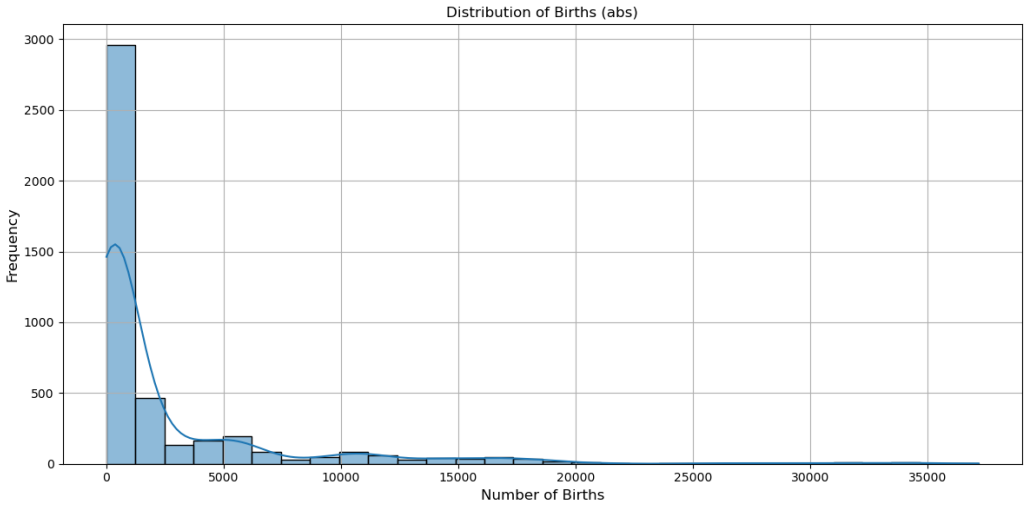


Figure 10: Distribution of Births (abs)

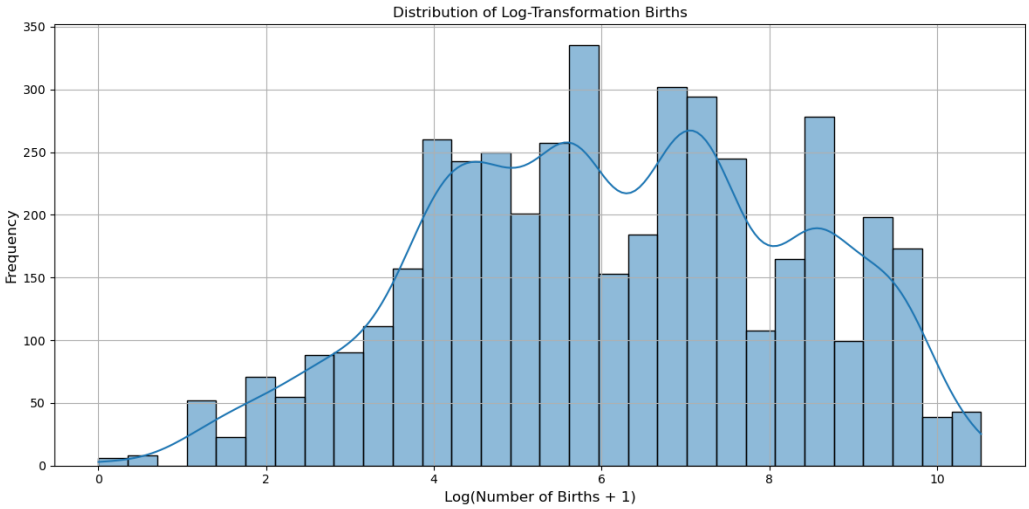


Figure 11: Distribution of Log-Transformation Births

Above figure shows plotted histogram of the abs values. The figure 10 shows distribution skewed to the right. Meaning most records like state, sex, ethnicity, and year have relatively moderate counts, but a fair number have very high counts and some have very low counts. This skew justified our usage of the log transform in modelling to not let those very large counts dominate to the error.

Analysis of Malaysia New Born Baby Dataset Report

Model Training

In this section will describe how each model was trained on the processed data. The modelling process were test with the original data models, evaluated results, then performed hyperparameter tuning especially for the more complex models and also training models on the log-transformed target (log_abs) to see is there any improved performance, since sometimes modelling outcome may in better fit.

Linear Regression

Performance on Original Data (abs):

The linear model explained only about 53% of the variance ($R^2 = 0.5319$) on the test data. This means it caught some patterns but still made many mistakes. The main weakness of this model is that it assumed the same rate of decline in births over time for all states, only changing the starting levels. It couldn't capture unique trends or more complicated patterns.

Performance on Log-Transformed Data (log_abs):

When predicting the log Transformed algorithm on births, the model performed better, explaining 73% of the variance ($R^2 = 0.\ 0.7257$). This improvement shows that birth data likely follows a percentage-based decline rather than a constant decrease. However, 73% is still far below compare with other model, indicating that more complex patterns like unique state trends or interactions between features might not capture by the simpler linear model.

In summary, linear regression are straightforward but limited. Using a log transformation improved performance by better matching exponential trends. However, this approach still missing crucial patterns that need more advance and non-linear modelling techniques.

No hyperparameter tuning needed for linear regression

Random Forest Regression

Performance on Original Data (abs):

The Random Forest model trained on the original data quickly achieve a very high accuracy. It showed a nearly perfect prediction with 99% ($R^2 = 0.9861$) on the test data. Such a high R^2 indicate the model learned the data exceptionally well, possibly because the prediction task was straightforward or due to hidden patterns or data leakage. Hyperparameter tuning was conducted by using grid search CV with 10-fold cross-validation, vary parameters like the number of trees (n_estimators: 50 or 100), maximum tree depth (max_depth: None or 10), and minimum samples per split (min_samples_split: 2 or 5). Surprisingly, the default parameters still performed best, maintaining the highest R^2 test at 0.9861.

Performance on Log-Transformed Data (log_abs):

On log-transformed data ("log_abs"), the Random Forest model also performed very well. After transform predictions back to the original scale, it achieved an R^2 of 0.9756 (same amount after hyperparameter tuning). While this is still excellent, it was slightly lower than the performance on the original scale.

Analysis of Malaysia New Born Baby Dataset Report

Support Vector Machine

Performance on Original Data (abs):

The Support Vector Regression (SVR) model performed poorly when predicting in the original birth numbers. With the default settings, it failed to learn meaningful patterns, achieve negative low R^2 (R^2 = -0.1770). The errors (RMSE) were 5338, the large error value showing the model was too simple for the data.

After tuning, SVR improved but still had limited success, it achieves R^2 in around 0.2594 and RMSE around 4234. This remained poor compared to other models. SVR struggled mainly because the dataset has many categorical variables, it created complex relationships were hard for learn, even with RBF kernel.

Performance on Log-Transformed Data (log_abs):

When predicting on log-transformed data the SVR model work much better. After tuning, R^2 has slightly dropped from 0.9750 to 0.9748 on the test set, it still nearly matching the Random Forest model. Using RBF kernel with higher flexibility allowed it to learn trends very accurately.

The significant improvement compared to the original data indicates that SVR is sensitive to large magnitude differences and variability, and transform the data into log-scale could simplified the relationships, making them easier to learn. However, it was relatively slow to train due to the large dataset, it requires more computation time compared to other models.

ARIMA Modelling

For time-series analysis, I took a different approach. Instead of using all available features, I focused only on total birth counts on each year for each state. I chose Perak (which is where I born) to demonstrate this method. The ARIMA model specifically for Perak to predict future birth counts and compare the forecast with results from our machine learning models.

Stationarity and Differencing:

```
In [122]: from statsmodels.tsa.stattools import adfuller
          adf_result = adfuller(Perak_df['abs'].dropna())


          # P-values lesser than 0.05, meaning Original is Stationary  , and I use Consist to use ARIMA because the data aren't given
          print("ADF Statistic:", adf_result[0])
          print("p-value:", adf_result[1])
          print("Number of Lags Used:", adf_result[2])
          print("Number of Observations:", adf_result[3])
          print("Critical Values:", adf_result[4])

          ADF Statistic: -3.64674607571649
          p-value: 0.004927722508232913
          Number of Lags Used: 0
          Number of Observations: 23
          Critical Values: {'1%': -3.7529275211638033, '5%': -2.998499866852963, '10%': -2.6389669754253307}
```

Figure 12: Coding for Augmented Dickey–Fuller Test and Result

Checked if this data was stationary by conducting an Augmented Dickey-Fuller (ADF) test. The test showed the series was stationary the p-value was 0.0049 lesser than 0.05, so reject the null hypothesis (H_0).

Thus, an ARIMA model with an integration component d=1 (first differencing) might appropriate. Besides, there is no obvious seasonality in yearly data aside from possibly a long-

term trend, so a simple ARIMA (p,d,q) should suffice because the dataset has no seasonal terms

Model Identification:

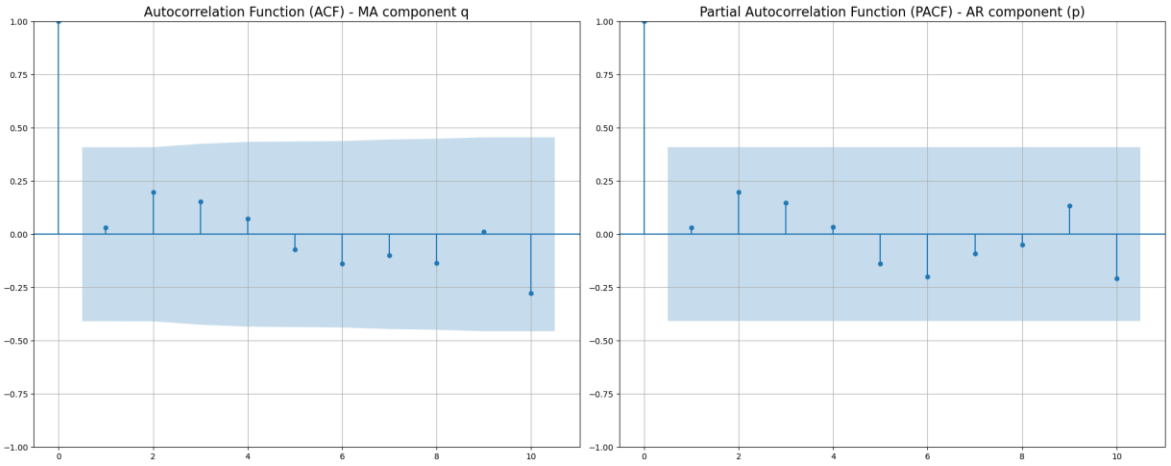


Figure 13: Visualization for Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

Autocorrelation and partial autocorrelation plots were examined to choose p and q. The autocorrelation of the differenced series showed a significant spike at lag 1, suggesting an AR(1) or MA(1) might be fitted. Hence, ARIMA (1,1,1) model was a reasonable choice.

SARIMAX Results

Dep. Variable:	abs	No. Observations:	24
Model:	ARIMA(1, 1, 1)	Log Likelihood	-193.703
Date:	Sun, 04 May 2025	AIC	393.407
Time:	10:11:14	BIC	396.813
Sample:	01-01-2000	HQIC	394.264
	- 01-01-2023		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.7844	0.045	17.330	0.000	0.696	0.873
ma.L1	-1.0000	0.282	-3.552	0.000	-1.552	-0.448
sigma2	1.287e+06	2.19e-07	5.89e+12	0.000	1.29e+06	1.29e+06

Ljung-Box (L1) (Q):	0.12	Jarque-Bera (JB):	0.94
Prob(Q):	0.73	Prob(JB):	0.62
Heteroskedasticity (H):	1.50	Skew:	-0.01
Prob(H) (two-sided):	0.58	Kurtosis:	2.01

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
[2] Covariance matrix is singular or near-singular, with condition number 1.17e+28. Standard errors may be unstable.

Figure 14: Summary Report for ARIMA (1,1,1) Model

After fitted an ARIMA (1,1,1) on the Perak total birth series using " statsmodels.tsa.arima". The model summary indicated coefficients and an AIC around 393.

Analysis of Malaysia New Born Baby Dataset Report

Forecasting with ARIMA:

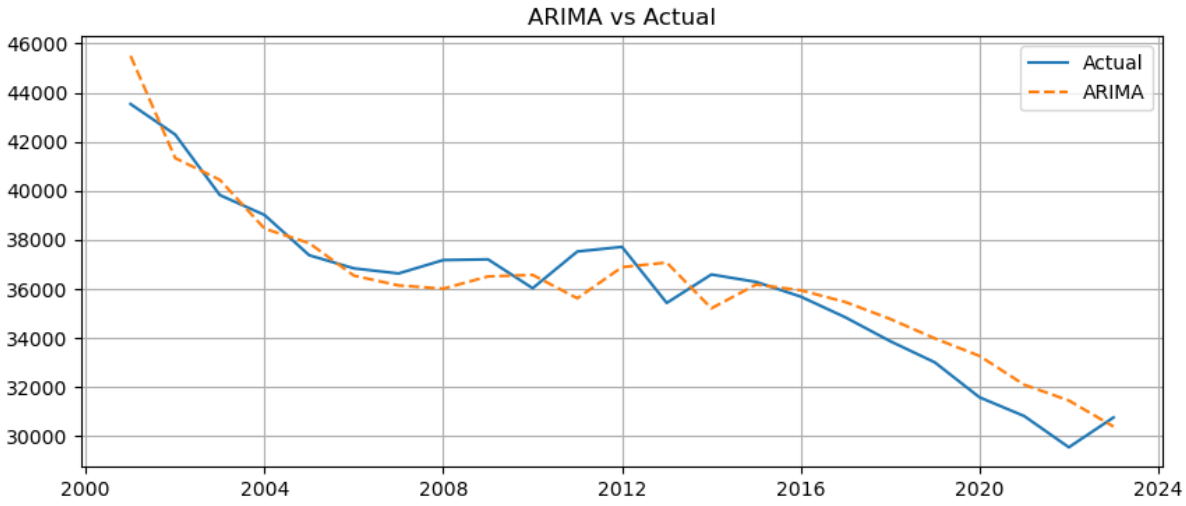
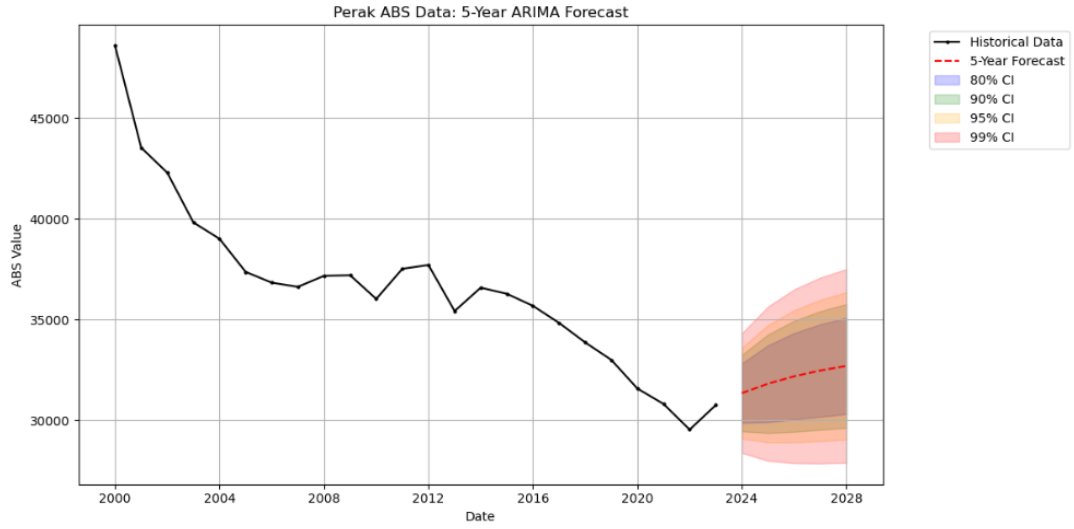


Figure 15: Visualization for Actual Value and Forecast Value Comparison

Before use the fitted ARIMA model to predict future values. Firstly, have to check the sample fit predictions align with actual values. Figure 15 plotted ARIMA vs actual for Perak birth count over the years, which showed the forecast line closely follow the actual with some lag. These predictions closely matched the actual birth counts, showing our model are worked well.



Out[128]:

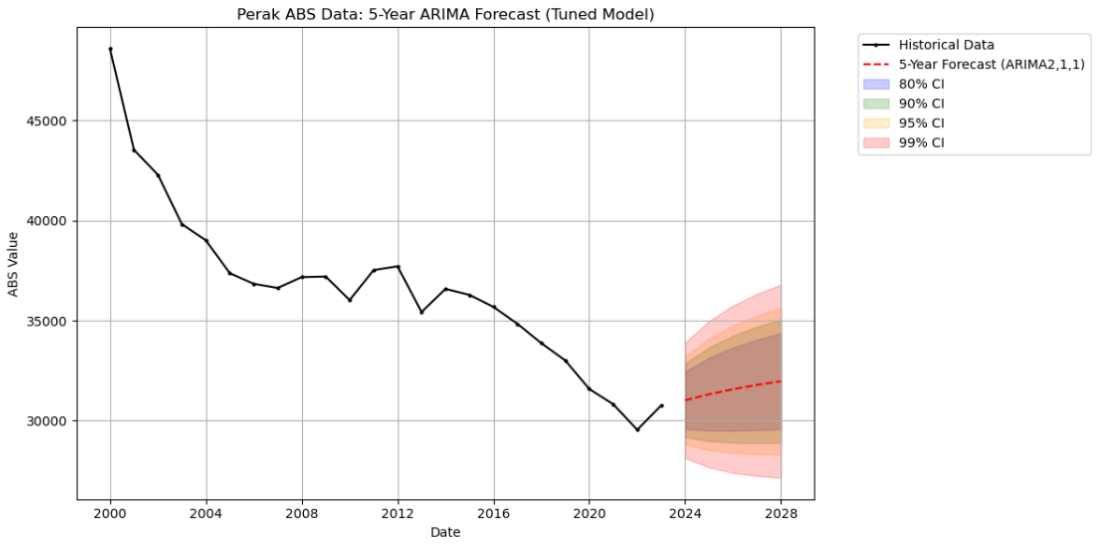
	Date	Forecasted ABS
0	2024-01-01	31354.527625
1	2025-01-01	31821.671989
2	2026-01-01	32188.109829
3	2027-01-01	32475.551359
4	2028-01-01	32701.026526

Figure 16: Visualization for Confidence Interval and Forecasted ABS Value

After tested the model with the existing data years (2000-2023). Then used the ARIMA model to forecast births for the next five years (2024–2028). The ARIMA forecast for Perak indicate a continuous slight increase in births, it projected the slowly upward trend. The visualized this forecast with confidence intervals as show in figure 16.

Analysis of Malaysia New Born Baby Dataset Report

Hyperparameter Tuning for ARIMA:



Out[142]:

	Date	Tuned Forecasted ABS
0	2024-01-01	31017.471880
1	2025-01-01	31319.892009
2	2026-01-01	31570.426144
3	2027-01-01	31785.810639
4	2028-01-01	31970.137845

Figure 17: Visualization for Tuned Confidence Interval and Forecasted ABS Value

Figure 17 shows tuned ARIMA Model. After tuning, The ARIMA model Identification switch to (2,1,1) model it gives a slightly different picture than the initial (1,1,1) fit:

1. Slight Rebound vs. Continued Upward
   Instead of projecting a curly upward, the tuned model has births are shows slightly flat. The predicted value from bottoming out around 31,000 in 2023 and then begin upward from 31,000 in 2024, 31,500 in 2026, and almost 32,000 by 2028.
2. Better In-Sample Fit
   Adding that extra AR(2) term then lower the AIC from 393 to 391, it smoothed out one-step-ahead residuals, so the model captures medium-term momentum slightly accurate.
3. Growing Uncertainty
   The confidence bands fan out rapidly by 2028 , the 95 % interval spans from roughly 28 000 to 36 000 births so while the point forecast nudges upward, the substantial are uncertain.

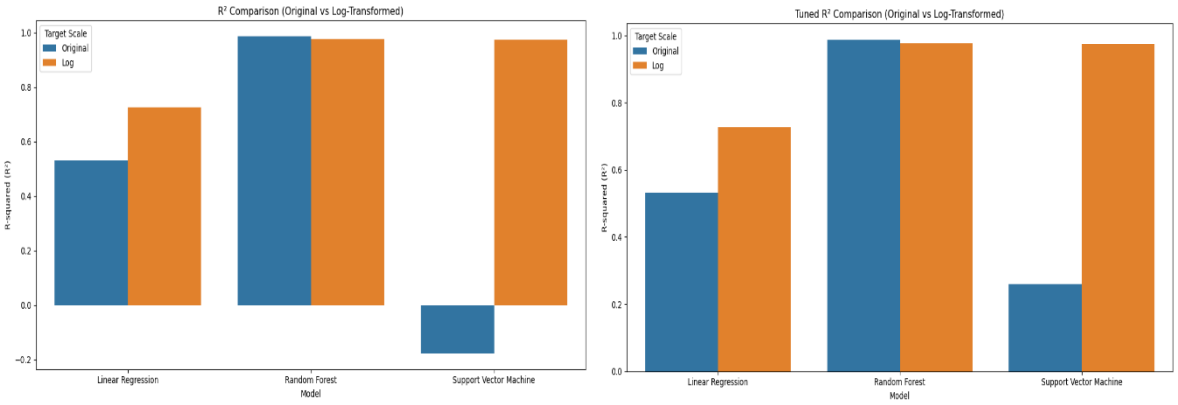
In Summary, to avoid a complex time series model that uses multiple variables (multivariate) because the machine learning models tested like Random Forest and Support Vector Machine already included features like year, state, sex, and ethnicity. These models effectively used these factors to predict births, making a separate multivariate time series approach is unnecessary.

ARIMA is a simpler time series model that focuses only on the time-based pattern. ARIMA are suitable because the data shows steady and consistent over time.

Analysis of Malaysia New Born Baby Dataset Report

Model Evaluation and Comparison

Regression Algorithm (Linear Regression, Random Forest and SVM)



Out[147]:

	Results					Tuned Results				
	Model	Scale	RMSE	MSE	R²	Model	Scale	RMSE	MSE	R²
0	Linear Regression	Original	3366.137775	1.133088e+07	0.531958	Linear Regression	Original	3366.137775	1.133088e+07	0.531958
1	Linear Regression	Log	1.137298	1.293446e+00	0.725790	Linear Regression	Log	1.137298	1.293446e+00	0.725790
2	Random Forest	Original	580.579076	3.370721e+05	0.986077	Random Forest	Original	580.579076	3.370721e+05	0.986077
3	Random Forest	Log	0.338873	1.148348e-01	0.975655	Random Forest	Log	0.338873	1.148348e-01	0.975655
4	Support Vector Machine	Original	5338.143646	2.849578e+07	-0.177067	Support Vector Machine	Original	4234.096656	1.792757e+07	0.259470
5	Support Vector Machine	Log	0.343392	1.179178e-01	0.975001	Support Vector Machine	Log	0.344276	1.185260e-01	0.974873

Figure 18: Visualization for Evaluation Metrics Result Comparison for Regression Algorithm

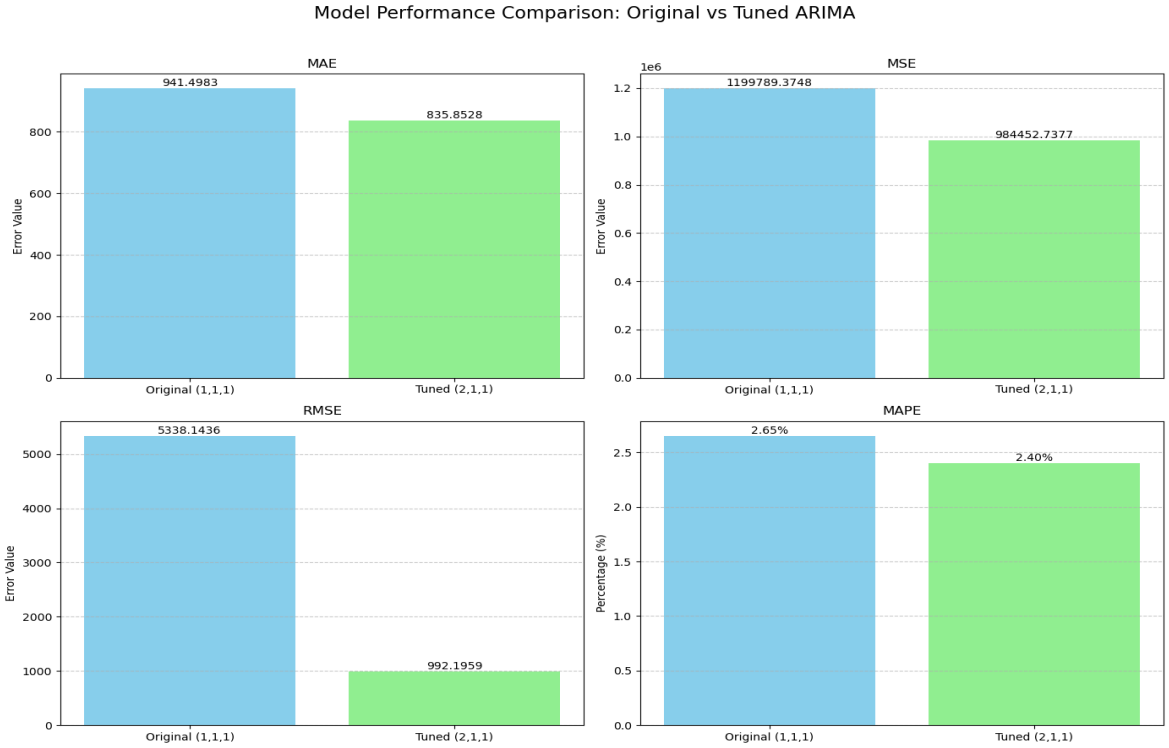
The figure 18 shows the compares performance of three models (Linear Regression, Random Forest, and Support Vector Machine) on predicting birth numbers using both original and log-transformed data. There's include:

1. Random Forest performed best in overall, achieve near-perfect accuracy ($R^2 = 0.9860$) on the original data, indicating it captured complex patterns that simpler models missed.
2. Linear Regression showed moderate performance ($R^2 = 0.5319$ on original data) but improved significantly when predict in log-transformed data ($R^2 = 0.7257$).
3. Support Vector Machine struggled with original data ($R^2 = 0.2594$ after tuning), log-transformed R^2 was dropped from 0.9750 to 0.9748 after tuning but nearly matched Random Forest's performance, highlighting its sensitivity to data scaling.
4. Log transformation consistently improved results for all models, reducing errors (RMSE) and aligning with exponential trends in the data.

In summary, Random Forest Regression is the best model for this task, while preprocess log transformation and parameter tuning are critical for optimizing performance. The results emphasize that model choice depends on data structure and need to balance accuracy with computational efficiency.

Analysis of Malaysia New Born Baby Dataset Report

Time Series Algorithm (ARIMA)



ARIMA Model Evaluation Metrics Comparison:

Out[148]:

	Metrics	Original Model	Tuned Model
0	MAE	941.4983	835.8528
1	MSE	1199789.3748	984452.7377
2	RMSE	5338.1436	992.1959
3	MAPE	2.65%	2.40%

Forecast abs Comparison:

Out[136]:

	Date	Forecasted ABS	Tuned Forecasted ABS
0	2024-01-01	31354.527625	31017.471880
1	2025-01-01	31821.671989	31319.892009
2	2026-01-01	32188.109829	31570.426144
3	2027-01-01	32475.551359	31785.810639
4	2028-01-01	32701.026526	31970.137845

Figure 19: Visualization for Evaluation Metrics Result Comparison for Time Series Algorithm

Figure19 shows the evaluation results compare forecasts and performance between the original and tuned models. Both models predict a gradual rise in births from 2024 to 2028, but the tuned model forecasts slightly lower values. Importantly, tuning improved accuracy across all metrics: the mean absolute error (MAE) dropped from 941 to 836 births, the root mean squared error (RMSE) fell sharply from 5,338 to 992, and the mean absolute percentage error (MAPE) improved from 2.65% to 2.40%. The significant reduction in MSE confirms that parameter adjustments effectively minimized large errors. While both models follow a similar upward trend, the tuned model's lower errors make its forecasts more reliable. This shows how refining the model parameters enhances precision without drastically altering trend predictions, balancing realistic forecasts with statistical accuracy.

In summary, the Random Forest model essentially performed as an automated nearest-neighbour predictor using the previous year data, while the ARIMA model capsulate the trend explicitly with a drift term. It is two different approaches: ARIMA is a parametric time series model assuming a linear trend plus noise, and Random Forest model is a non-parametric regression based on recent history.

Analysis of Malaysia New Born Baby Dataset Report

Deployment

In this academic project, the deployment phase is a theoretical exercise focused on finalizing the report rather than implementing a real-world solution. Since the project is for academic purpose, practical steps like deployment planning or ongoing monitoring and maintenance are not required.

Conclusion

This project aimed to predict newborn birth counts in Malaysia. By analyzing decades of data, it confirmed a steady annual decline in births. Following the CRISP-DM framework tested models like linear regression, Random Forest, Support Vector Regression, and ARIMA. Results showed Random Forest outperformed compare others, achieved near-perfect accuracy (R^2 = 0. 9860), while linear regression was less effective ($R^2 \approx$ 0. 5319). This proves advanced models better capture complex patterns. ARIMA provided reliable future forecasts, aligning with the overall slow increase.

The study tested Random Forest and Support Vector Regression made very small errors its only a few hundred births wrong, even when predicting tens of thousands of births. Linear regression had much larger errors. This difference shows that the data has complex patterns that simple models cannot capture, but advanced models like Random Forest can.

ARIMA are forecasts time series, it can't compare directly with other models into random test split. ARIMA (1,1,1) model predicted births from 2019 to 2023 with small errors, it forecast 32,093 births in 2021 when the actual number was 30,816 births, a difference of 1277 or approx. 4.1%. The RMSE was a few hundred births. If we used ARIMA for all states, larger states would show larger errors, but overall, it would still perform well. However, ARIMA treats each state separately and does not use information from other states. In contrast, the random forest model uses both year and state features, so it can learn each state's trend more effectively.

Analysis of Malaysia New Born Baby Dataset Report

Ethnic Consideration

1. **Privacy Protection**:
   The data is grouped and it does not apprears personal details like names or address. Because the resources are open for public, therefore privacy risks are low. However, it must still use the data responsibly to avoid misrepresenting it and credit the source.

2. **Avoiding Bias and Misinterpretation**:
   The data would show difference in birth numbers between ethnic groups and states. This is not because of any group inherent traits but due to factors like population size or access to healthcare. Analysts should focus on using the results to allocate resources fairly and not to blame or stereotype to the groups.

3. **Responsible of Policy Use**:
   If the governments use these predictions, they must ensure the outcome is to focus to support the community rather than punish the groups with lower birth rates. Policies should aim to help groups equally.

4. **Transparency and Communication**:
   The Random Forest model is accurate but complex. If the government use it for planning, the governments must clearly explain how it works and why decision are made. The explanations could help to build public trust.

Timeline

1. **January 2025 – February 2025: Data Understanding**
   Define research goal and study relevant research online and collect data from Malaysia's official public database (DOSM) and explored the data to identify patterns and plan for next steps.

2. **March 2025: Preparing the Data**
   Clean and preprocess the data for analysis and created charts and to deep understanding the relationship of each feature then summaries to guide model development.

3. **Late March 2025 – April 2025: Building and Testing Models**
   Developed prediction models and test different settings to improve accuracy as well record the progress and update the models step-by-step.

4. **Late April 2025: Evaluating Results**
   Test final models to check the performance, measure accuracy by using evaluation metrics and conduct extra checks to ensure reliability.

5. **Early of May 2025: Writing the Report and Poster**
   Start to drafting the thesis report, explain the methods and findings. Add charts, tables, and explanations to make the report and poster clear and organized sections to meet academic standards.

6. **Mid of May 2025: Final Submission**
   Complete the report, double check references and the format to ensure the report met college guideline for submission.

Reference

1. Bernama, Factors Driving Decline in Live Birth in Malaysia, 2024. Available at: https://www.bernama.com/en/news.php?id=2379875

2. Breiman, L., 2001. Random forests. Machine Learning, 45(1), pp.5-32.

3. Drucker, H., Burges, C.J., Kaufman, L., Smola, A.J. and Vapnik, V., 1997. Support vector regression machines. Advances in Neural Information Processing Systems, 9.

4. H2O.ai, Support Vector Machine (SVM). Available at: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/svm.html

5. IBM, ARIMA Model. Available at: https://www.ibm.com/think/topics/arima-model

6. Department of Statistics Malaysia (DOSM), 2023. Vital Statistics Malaysia 2023. Available at: https://www.dosm.gov.my/portal-main/release-content/vital-statistics-malaysia-2023

7. Al-Khassawneh, Y., Al-Rabadi, R. and Al-Zoubi, O., 2022. A hybrid machine learning approach for improving the accuracy of medical diagnostic systems. Journal of Statistics Applications & Probability, 11(2), pp. 345-356. Available at: https://www.ewadirect.com/proceedings/aemps/article/view/7293

8. Al-Hussein, M., Al-Kasasbeh, B. and Al-Sarayreh, K., 2021. Predictive analytics for healthcare using time-series forecasting models. Jordan Journal of Statistical Sciences, 14(1), pp. 23-45. Available at: https://digitalcommons.aaru.edu.jo/cgi/viewcontent.cgi?article=1621&context=jsap

9. Adeboye, N.B., Adedotun, A.F. and Ogunjobi, O., 2023. Modelling and forecasting urban population growth in Nigeria using autoregressive integrated moving average (ARIMA) models. African Journal of Applied Statistics, 10(1), pp. 78-95. Available at: https://www.researchgate.net/publication/384903487

10. Othman, N., Razali, F. and Ismail, M., 2023. Open data and machine learning for birth prediction and classification: A case study utilizing Malaysia's public sector open data portal. Journal of Data Science and Analytics, 15(3), pp. 210-225. Available at: https://www.researchgate.net/publication/388103270