PREDICTING MALAYSIA'S NEWBORN BIRTH TRENDS BY USING MACHINE LEARNING

01 INTRODUCTION

This study aims to predict birth numbers in Malaysia using different machine learning techniques. By analyzing historical data from 2000 to 2023 and explain how different models can help understand population changes and predict future birth count.

02 OBJECTIVE

1. Illustrate Exploratory Data Analysis (EDA) to identify trends.
2. Develop and compare multiple predictive models.
3. Evaluate model performance using standard metrics.
4. Predict the future birth number

03 METHODOLOGY

Details the overall CRISP-DM process and maps the project activities to its phases:

1. Data Understanding
2. Data Pre-processing and Analysis
3. Model Training
4. Model Evaluation and Comparison
5. Deployment

04 DATA UNDERSTANDING

1. Total 5 features with 4 features is **object** and 1 feature is **float**
2. Total 7854 entries, abs features has 64 **missing value**, total file size is **306.9kb**.
3. Contain overlap data in sex and ethnicity feature

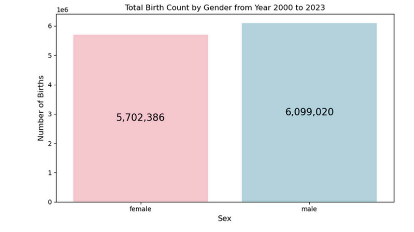
05 DATA PREPROCESSING AND ANALYSIS

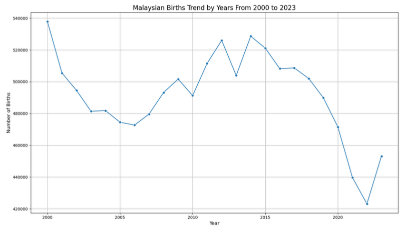
1. **Drop overlap data** sex = both and ethnicity = overall
2. Convert “date” to **datetime format**
3. Filling missing value by **grouped median** of state, sex, and ethnicity
4. Perform **pipeline** column transfer to run **imputer** and **one hot encoder** for categorical data
5. Perform **log-transformation** (skew right in distribution)

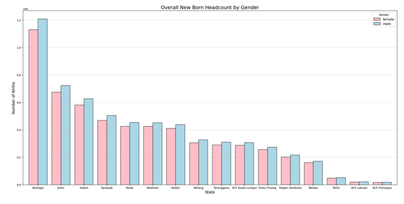
06 MODEL TRAINING

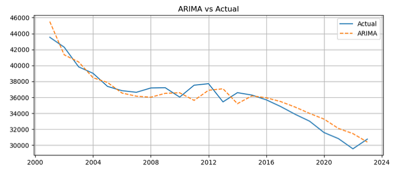
Different machine learning models were trained to predict birth accuracy and number (abs) from the data. The approaches included:

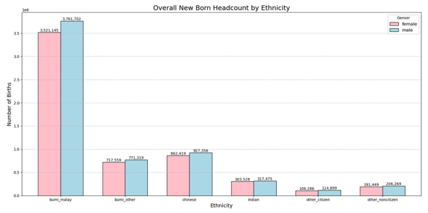
- **Linear Regression** - Basic Regression Model
- **Random Forest Regression** - The ensemble tree-based model
- **Support Vector Machine (SVM)**
- **ARIMA** Model - Time series prediction

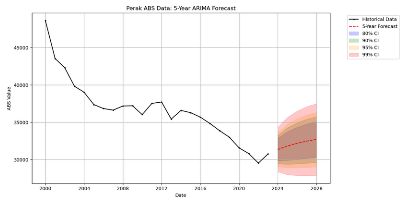




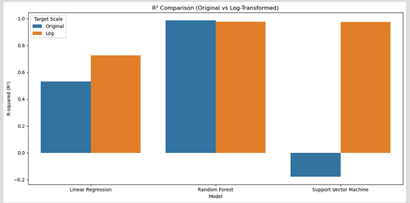


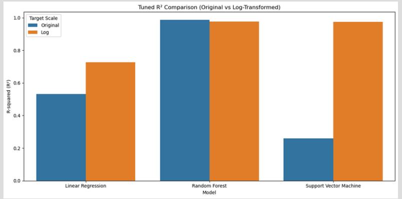


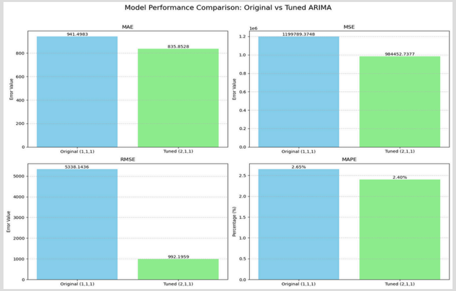




07 MODEL ANALYSIS AND EVALUATION COMPARISON







1. **Random Forest Regression**: Accuracy with R² = 0.9860 on original scale R² = 0.9756 on log-transformed scale, remain the same after hyperparameter tuning
2. **Linear Regression**: Original scale R² = 0.5319 but with log transformation slightly well perform (R² = 0.7257).
3. **Support Vector Machine** : Struggled with original data but performed well after hyperparameter tuning (R² = -0.1770 to 0.2594), log transformation perform moderate but dropped after tuning (R² = 0.9750 to 0.9748)
4. **ARIMA Model**: Provided reliable future forecasts, predicting a slight increase in births from 2024 to 2028 with lower MAPE error from 2.65% to 2.40% after tuning.

08 RESULTS/ FINDINGS

1. Overall Malaysia experienced a **steady annual decrease** in new-born births over decades.

2. **Model Performance**:

- **Random Forest** excelled (R² ≈ 0.99), making errors of only a few hundred births.
- **Linear Regression** struggled (R² ≈ 0.53), showing larger inaccuracies.
- ARIMA provided forecasts (≈ 2.4% error) and predicted 5 years birth number for Perak (**2024≈31017**, **2025≈31319**, **2026≈31570**, **2027≈31785**, and **2028≈31970**)

3. ARIMA are performing **short-term forecasts** cannot leverage cross-state data like Random Forest

4. **Regional Insights**: Larger states may see higher prediction errors, but trends are still reliably captured.

09 CONCLUSION

The Random Forest model had 99% accuracy (R²), which was much higher than linear regression (53%). This is also understood the detailed patterns in different states. The ARIMA model are nearly match the observed decrease in births trend over the year.

Advanced methods like Random Forest are better perform from the past data, while ARIMA helps predict future trends, proving both model are give most accurate results for Malaysia’s birth data.