OPEN DATA AND MACHINE LEARNING FOR BIRTH PREDICTION AND CLASSIFICATION

A Case Study Utilizing Malaysia's Public Sector Open Data Portal

MUHAMMAD SUKRI BIN RAMLI
Asia School of Business
Kuala Lumpur, Malaysia
Email: m.binramli@sloan.mit.edu

Abstract

This study analyzes birth data in Malaysia from 2000 to 2023, employing machine learning techniques to predict birth numbers, categorize birth rate periods, and explore newborn sex prediction. The analysis utilizes Linear Regression (James et al., 2013), Random Forest Classifier (Breiman, 2001), Prophet (Taylor & Letham, 2018), and XGBoost (Chen & Guestrin, 2016) models, incorporating feature engineering and handling missing values. Results show that XGBoost outperforms Linear Regression in predicting birth numbers, achieving a lower Mean Squared Error. The study also highlights potential overfitting in the classification task and the infeasibility of predicting newborn sex based on year and ethnicity alone. Future work includes incorporating additional features, exploring more sophisticated models, and addressing overfitting to enhance prediction accuracy and understanding of birth trends in Malaysia.

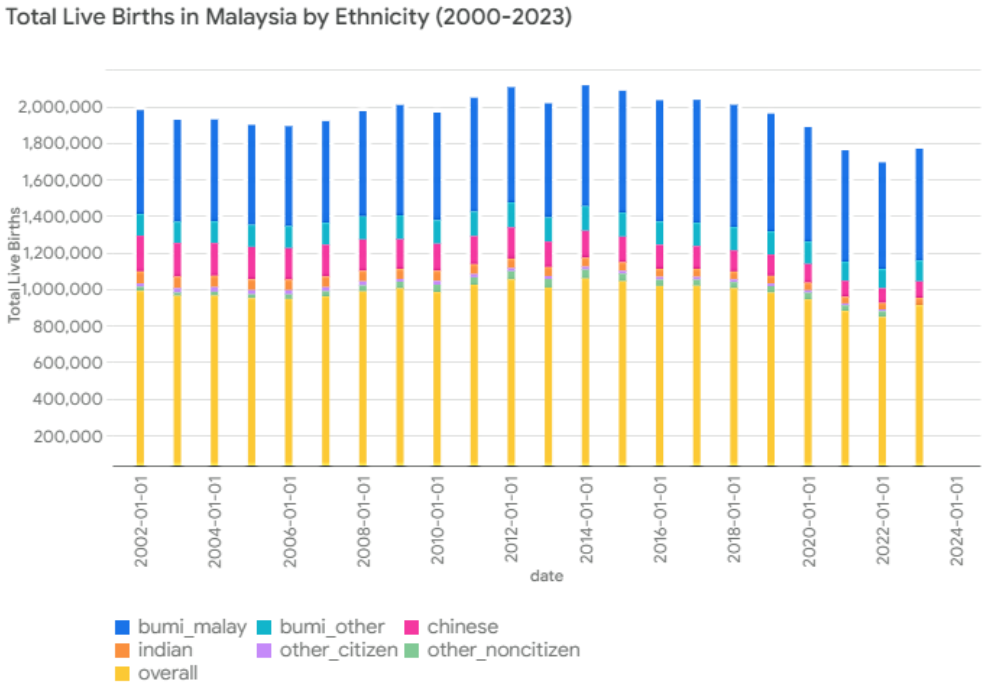

Figure 1: Total live births in Malaysia by ethnicity

1. Introduction

This report presents an analysis of birth data in Malaysia, sourced from the nation's Official Open Data Portal (data.gov.my). This initiative is rooted in the "Pekeliling Am Bil. 1 Tahun 2015," which underscores the Malaysian government's commitment to open data, thereby enhancing transparency and fostering innovation. The dataset encompasses birth records from January 1st, 2000, to December 31st, 2023, and includes details on date, sex, ethnicity, absolute number of births, and birth rate. This comprehensive dataset facilitates the examination of demographic shifts, such as variations in birth rates across ethnicities and over time. The analysis will encompass various techniques, including machine learning models like Linear Regression, Random Forest Classifier, Prophet, and XGBoost, to predict birth numbers, categorize birth rate periods, and explore newborn sex prediction. The findings of this study have the potential to inform policy decisions and resource allocation in the healthcare sector.

2. Background and Literature review

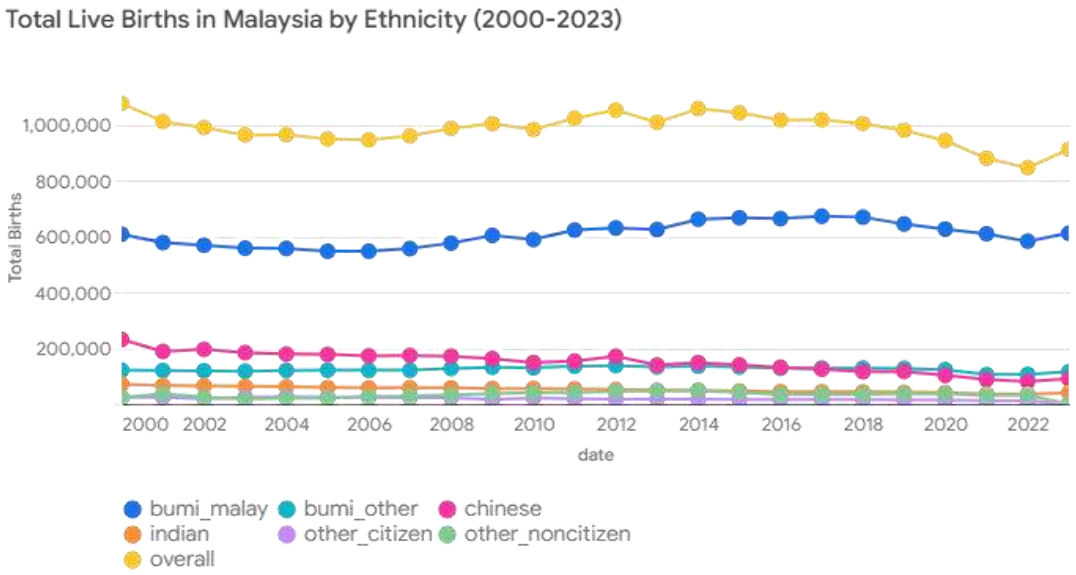

Figure 2: Total live births in Malaysia by ethnicity

2.1 Objectives

The Malaysian government has been actively promoting open data initiatives to enhance transparency, accountability, and public participation. This effort is guided by several key documents and studies, which also intersect with the growing use of machine learning for demographic analysis. This literature review will explore the implementation of open data in Malaysia, the application of machine learning in demographic studies, and the ethical considerations associated with these technologies.

The foundation of Malaysia's open data policy is outlined in the "Pekeliling Am Bil. 1 Tahun 2015: Pelaksanaan Data Terbuka Sektor Awam" (General Circular No. 1 Year 2015: Implementation of Open Data in the Public Sector). This document sets forth the objectives, principles, and guidelines for implementing open data in the public sector. The Malaysian Administrative Modernization and Management Planning Unit (MAMPU) has further elaborated on the benefits and progress of these initiatives in their report, "Data Terbuka Kerajaan Malaysia: Meningkatkan Ketelusan dan Kebertanggungjawaban" (Open Government Data in Malaysia: Enhancing Transparency and Accountability). Additionally, the economic impact of open data has been studied in "Kajian Kesan Data Terbuka Terhadap Ekonomi Malaysia" (Study on the Impact of Open Data on the Malaysian Economy), which highlights its positive effects on innovation, productivity, and job creation.

Machine learning techniques have shown significant potential in demographic analysis. Adhikari et al. (2020) demonstrated the use of machine learning algorithms to predict fertility rates in Nepal, showcasing the applicability of these techniques in demographic studies. Similarly, de Silva and Tennakoon (2021) applied machine learning models to forecast population growth in Sri Lanka, emphasizing the importance of incorporating various factors into the analysis. Abel and Sander (2019) provided a comprehensive overview of the applications of machine learning in demography, discussing both its potential and limitations.

The integration of machine learning in demographic analysis raises several ethical concerns. Mittelstadt et al. (2016) discussed the ethical challenges associated with using machine learning for demographic analysis, including issues of privacy, fairness, and accountability. Malin et al. (2015) provided guidelines for protecting data privacy in demographic research, stressing the importance of informed consent and data anonymization. Mehrabi et al. (2021) examined the potential for bias in machine learning algorithms and discussed strategies to mitigate these biases.

The Malaysian government's commitment to open data, as outlined in key policy documents and reports, has laid a strong foundation for enhancing transparency and accountability. The application of machine learning techniques in demographic analysis offers promising opportunities but also necessitates careful consideration of ethical issues. Ensuring data privacy, fairness, and accountability is crucial to avoid perpetuating inequalities and to harness the full potential of these technologies.

2.2 Hypothesis

H1: The number of births can be accurately predicted based on the year and ethnicity. This hypothesis assumes that birth rates are influenced by both time trends and ethnic-specific factors.

H2: Years can be categorized into high and low birth rate periods using machine learning models. This hypothesis suggests that underlying patterns in the data can be effectively captured by machine learning algorithms.

H3: The sex of newborns can be predicted based on the year and ethnicity. This hypothesis explores the potential link between demographic factors and sex ratios at birth.

H4: There are significant differences in birth rates across different ethnic groups. This hypothesis acknowledges the potential impact of cultural and socioeconomic factors on birth rates within different ethnic communities.

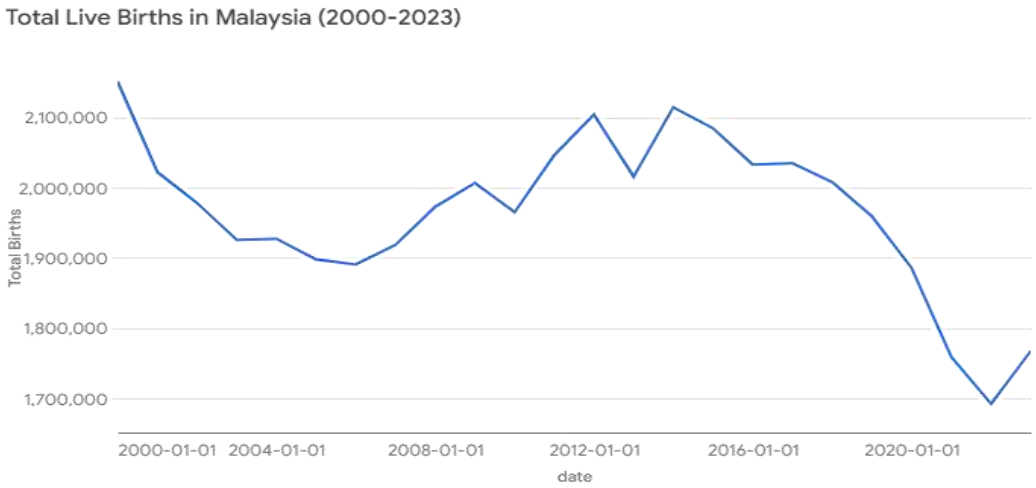

Figure 3: Total live births in Malaysia (2000-2023)

2.3 Data overview

The dataset contains records of births in Malaysia from 2000 to 2023, obtained from the Department of Statistics Malaysia (DOSM, 2023). It includes information on the date of birth, sex of the newborn, ethnicity of the mother, absolute number of births (abs), and birth rate (rate). The dataset provides a comprehensive view of birth trends over time, with breakdowns by sex and ethnicity. The data spans 24 years, capturing variations in birth rates and numbers across different demographic groups.

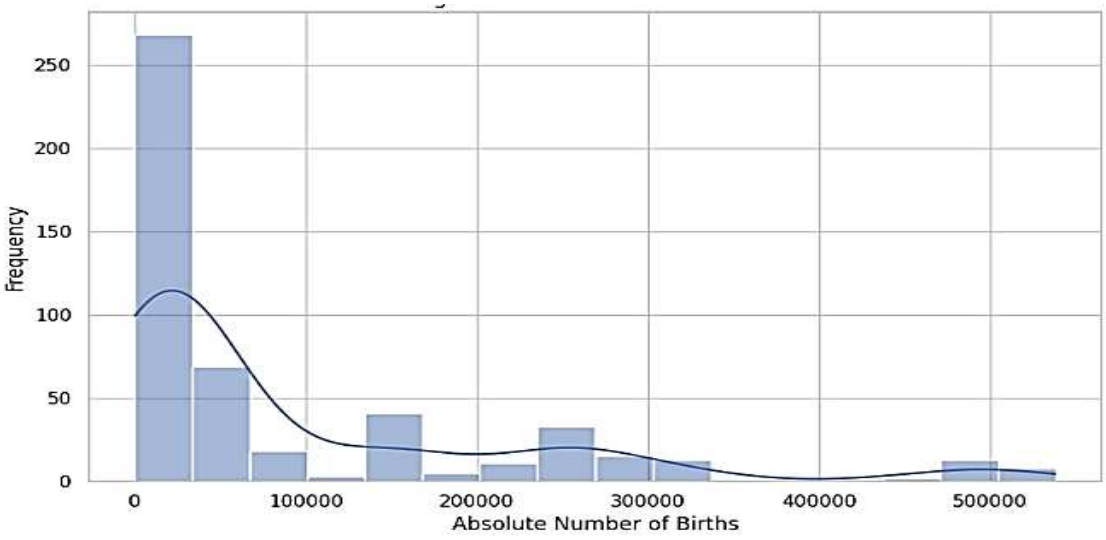

Figure 4: Histogram of absolute number of births

Summary Statistics

	count	mean	std	min	25%	50%	75%	max
abs	500	94321.5	126700	0	11307.5	31382	147460	537853
rate	500	17.67	8.93	0	11.38	17.05	21	48

This table summarizes the key statistics for the absolute number of births (abs) and birth rate (rate) across the dataset. The mean number of births is approximately 94,321, with a standard deviation of 126,700, indicating significant variability in the data. The birth rate has a mean of 17.67, with a standard deviation of 8.93. The minimum and maximum values for both variables highlight the range of birth numbers and rates observed over the years.

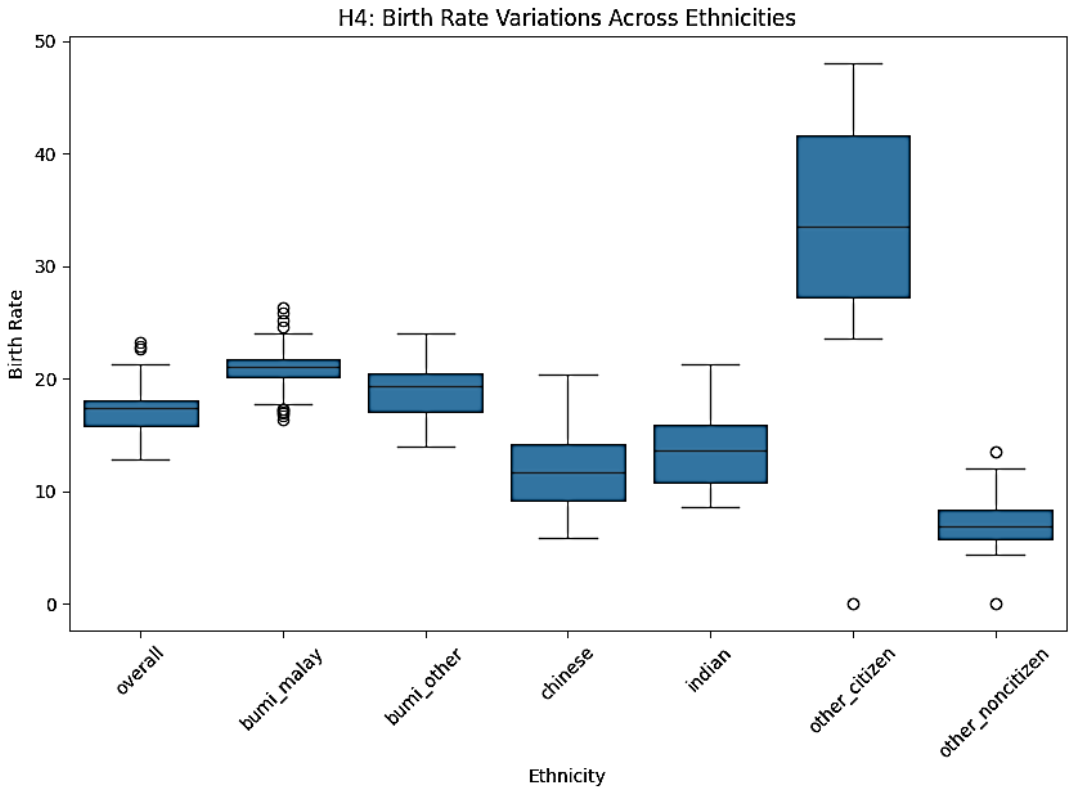

Figure 5: Boxplot of birth rate variation across ethnicities

The histogram of the absolute number of births displays a right-skewed distribution, indicating that most records have relatively low birth numbers, with a few instances of very high birth numbers. This skewness could be attributed to variations in population size, differing birth rates across various ethnicities, or potential outliers in the data. Additionally, the box plot of birth rates by ethnicity reveals significant variations among different ethnic groups. Some ethnicities exhibit a wider range of birth rates, suggesting greater internal variability or the influence of socioeconomic factors within those groups. These visualizations highlight the complexity and diversity of demographic patterns, emphasizing the need for careful analysis to understand the underlying factors driving these trends.

3. Methodology

This analysis explored birth trends using a variety of machine learning models, each chosen for its specific strengths and suitability to the task. Linear regression, a foundational statistical method described by James et al. (2013), was employed to model the relationship between the number of births, the year, and the ethnicity of the newborns. This model assumes a linear relationship between these predictors and the outcome variable, providing a baseline understanding of how these factors interrelate.

To address classification challenges, a random forest classifier was implemented. This ensemble learning method, known for its robustness and accuracy as described by Breiman (2001), was used for two distinct purposes. Firstly, it

categorized years into periods of high and low birth rates based on the overall birth rate, providing a macro-level view of birth trends over time. Secondly, the classifier explored the prediction of newborn sex based on year and ethnicity, potentially revealing subtle patterns and relationships within the data.

For the crucial task of forecasting future birth trends, the Prophet model was selected. This model, specifically designed for time series data, excels at handling seasonality and trend changes, making it particularly well-suited for analyzing birth data, which often exhibits cyclical patterns and long-term trends. As detailed by Taylor & Letham (2018), Prophet is particularly effective in this domain. By learning from the historical data of absolute birth numbers, the Prophet model aimed to provide insights into future birth patterns.

To further enhance the predictive power of the analysis, XGBoost, a gradient boosting algorithm renowned for its high performance and efficiency, was employed as an advanced regression model. XGBoost's ability to capture complex non-linear relationships in the data, as highlighted by Chen & Guestrin (2016), made it a valuable tool for predicting the number of births with greater accuracy. This model complemented the linear regression model by providing a more nuanced understanding of the factors influencing birth numbers.

Before applying these machine learning models, careful data preprocessing was undertaken. Missing values in the birth data, a common occurrence in real-world datasets, were handled through imputation, replacing them with the median value of the relevant column. This approach, discussed by Little & Rubin (2019), helped preserve the overall distribution of the data while effectively addressing missing data points. Furthermore, categorical variables, namely ethnicity and sex, were converted into numerical representations using one-hot encoding. This essential transformation ensured compatibility with the machine learning models, which typically require numerical input, as explained by Géron (2019).

In addition to preprocessing, feature engineering played a key role in potentially improving the accuracy of the regression models. Interaction terms between year and each ethnicity category were meticulously created. These interaction terms aimed to capture the combined effect of year and ethnicity on the number of births, recognizing that birth rates might evolve differently over time for different ethnic groups, as suggested by Kuhn & Johnson (2013).

Finally, to rigorously assess the performance of the chosen models, appropriate evaluation metrics were employed. The performance of the regression models, namely linear regression and XGBoost, was evaluated using the Mean Squared Error (MSE). This metric measures the average squared difference between the predicted and actual values, providing a quantifiable measure of prediction accuracy. For the classification models, the accuracy score was used. This metric represents the proportion of correctly classified instances, offering a clear indication of the model's effectiveness in categorizing data and predicting newborn sex.

4. Results

4.1 Predicting Number of Birth (Regression Model)

The Linear Regression model, which used year and ethnicity as predictors, resulted in a Mean Squared Error (MSE) of 4,071,814,509.95. This relatively high MSE indicates that the model might not be adequately capturing the complexity of the relationship between these predictors and the number of births. This could be due to the assumption of a linear relationship, which may not hold true in reality (James et al., 2013). A scatter plot of actual versus predicted births for Linear Regression would help visualize the model's performance and highlight any discrepancies.

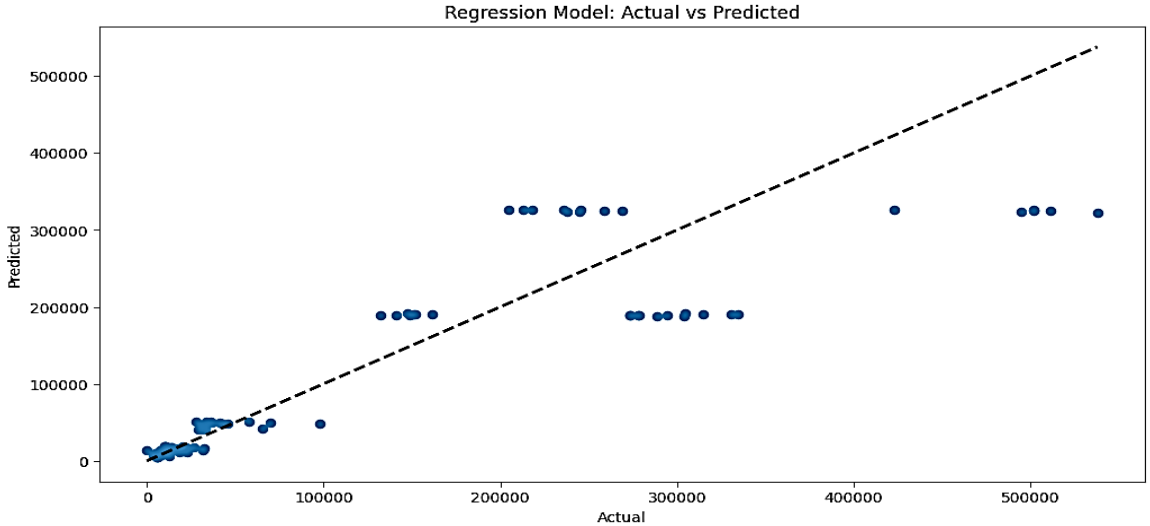

Figure 6: Regression model between actual vs predicted

In contrast, the XGBoost model, which incorporated year, ethnicity, and sex as predictors, achieved a significantly lower MSE of 54,561,070.31. This substantial improvement suggests that XGBoost is better able to capture the complex, potentially non-linear relationships in the data and provide more accurate predictions of birth numbers. This aligns with the known strength of XGBoost in handling complex relationships (Chen & Guestrin, 2016). A scatter plot of actual versus predicted births for XGBoost would demonstrate the model's improved accuracy compared to Linear Regression.

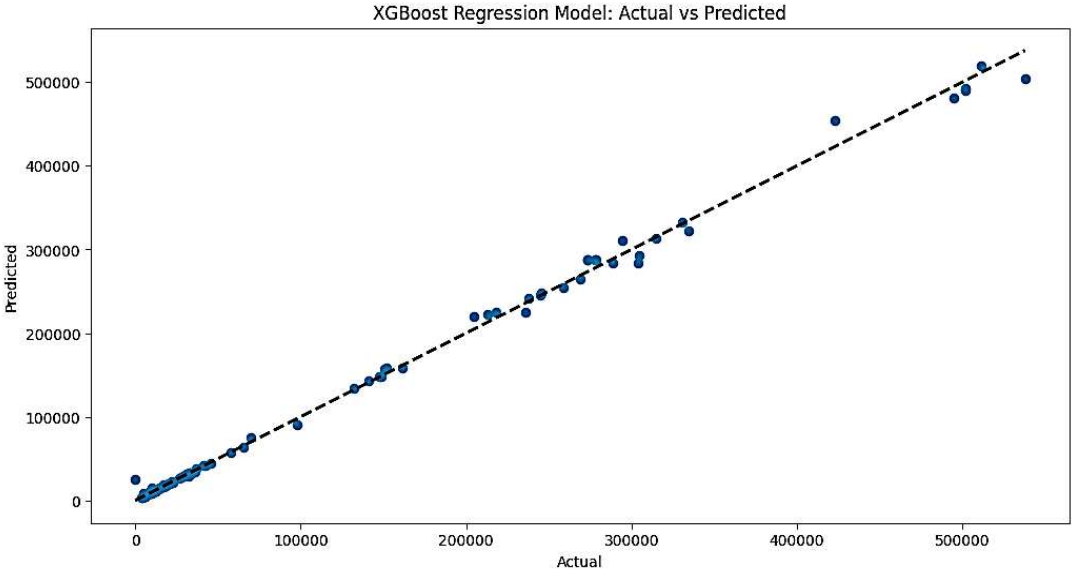

Figure 7: XGBoost Regression model between actual vs predicted

4.2 Categorizing Years (Classification)

The Random Forest Classifier was trained to categorize years into high or low birth rate periods based on the overall birth rate. This model achieved an accuracy of 1.0. While this perfect accuracy seems impressive, it raises concerns about potential overfitting. Overfitting occurs when a model memorizes the training data too well and may not generalize well to new, unseen data (Kuhn & Johnson, 2013). A confusion matrix for the classification model would provide insights into the model's performance and any potential overfitting issues.

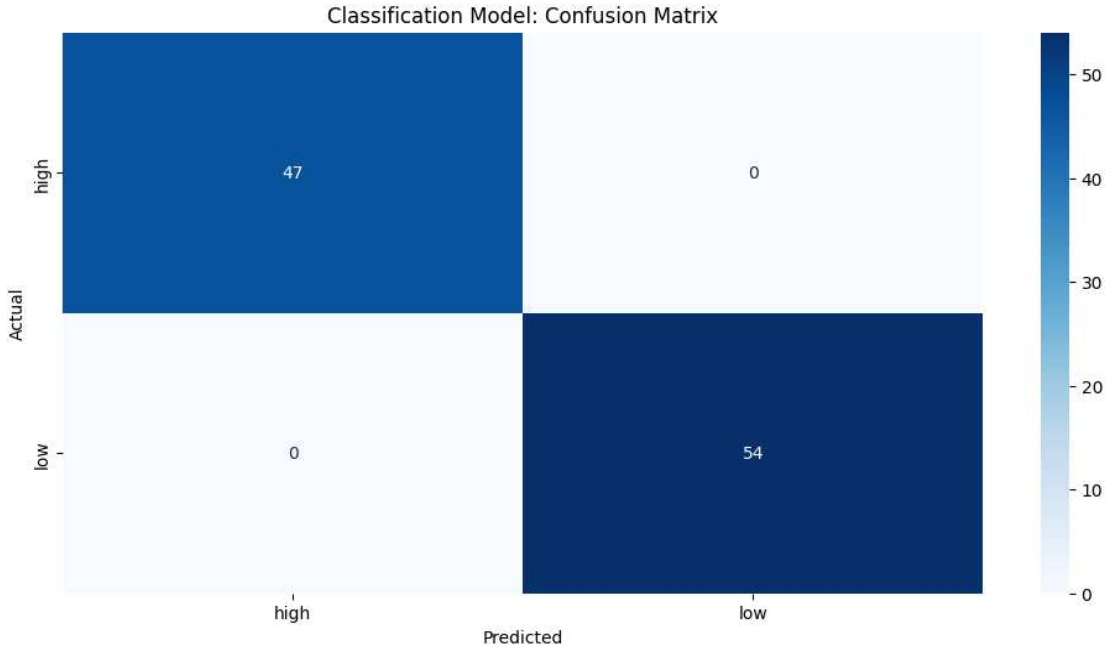

Figure 8: Confusion matrix of the classification model

When used to predict the sex of newborns based on year and ethnicity, the Random Forest Classifier achieved a very low accuracy of 0.03. This result confirms that predicting sex based solely on these factors is not feasible, as they are not strongly related to sex determination. This is expected as sex is largely determined by biological factors not captured in the data. A confusion matrix for sex prediction would highlight the model's poor performance and the lack of a strong relationship between the predictors and the outcome.

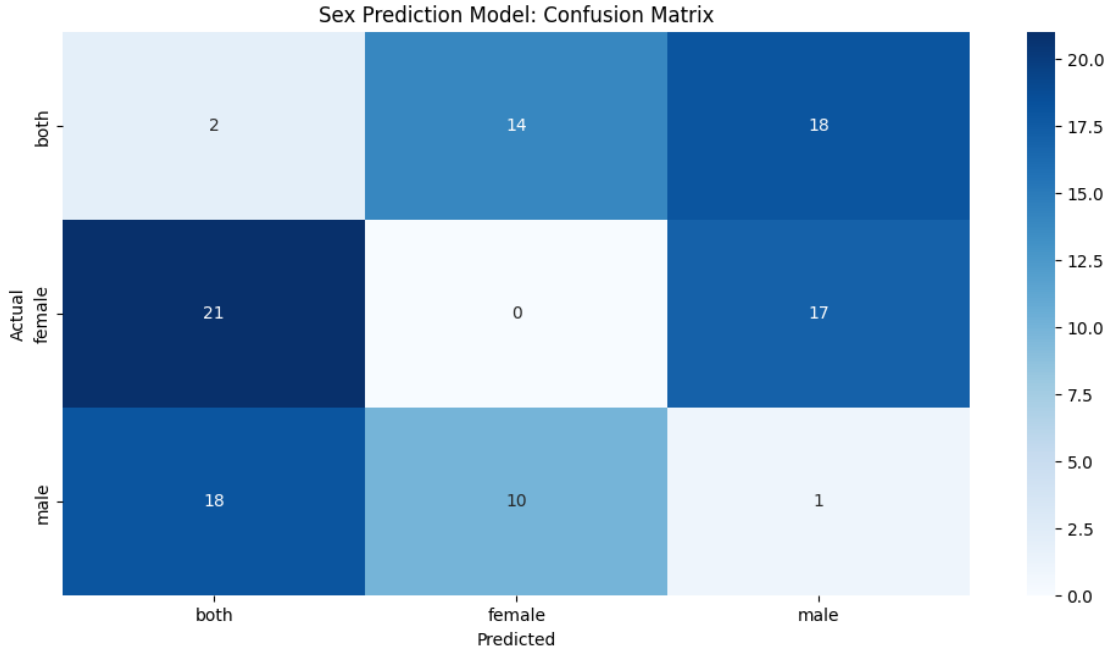

Figure 9: Confusion matrix for sex prediction

4.3 Future Prediction (using Regression Model)

The Linear Regression model was also used to predict the number of births for future years (2024-2030). Looking at the provided plot, the predictions show a clear increasing trend in the number of births over time. It's important to note that using a linear model for future predictions might have limitations due to potential non-linear trends in birth rates. The plot also doesn't provide information about variations across ethnicities - that would require a separate visualization breaking down the predictions by ethnicity.

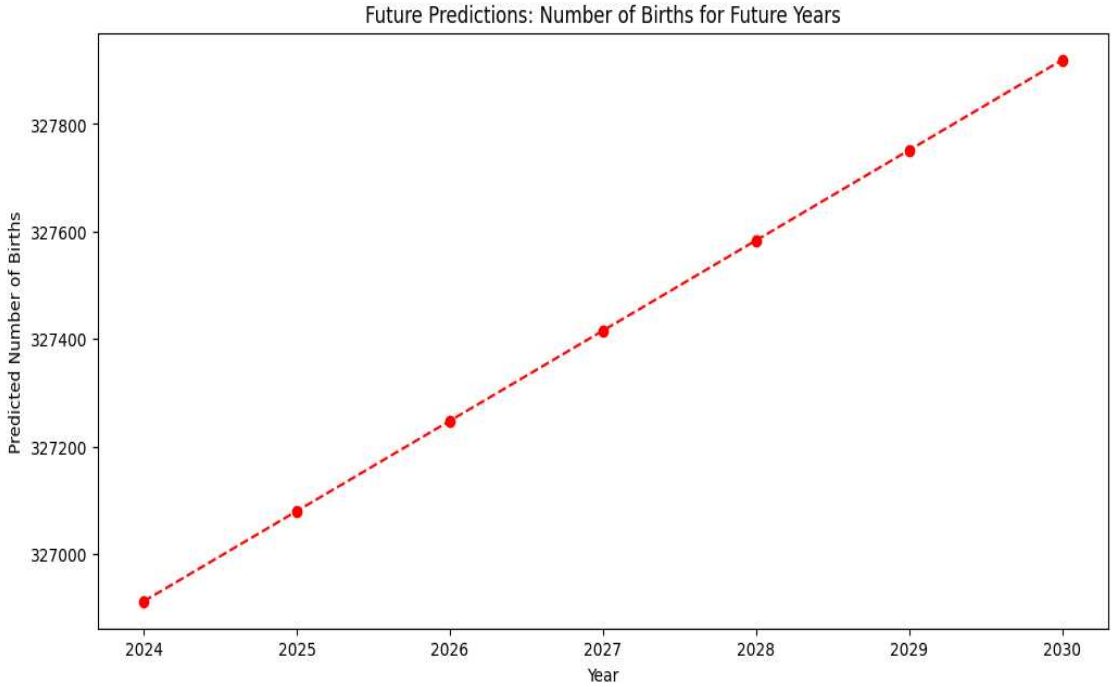

Figure 10: Regression Model for future number of births

4.4 Time Series Forecasting (using Prophet)

The Prophet model was used to forecast future birth trends based on historical data. The forecast, as depicted in the plot, reveals several interesting trends and patterns. While there's a slight upward trend in the earlier years (pre-2020), the overall trend seems to stabilize in the long term. The blue line, representing the trend component of the forecast, indicates a relatively flat trajectory, suggesting that the number of births is expected to remain relatively consistent in the future. The forecast clearly shows strong seasonal patterns, likely reflecting factors such as cultural events, holidays, or even weather patterns that might influence birth rates. The most striking pattern is the dramatic increase in predicted births with high uncertainty around the years 2025-2027. This suggests a potential outlier event or a period of significant change that the model is attempting to capture. It's crucial to investigate further to understand what might be driving this unusual pattern. It could be a data anomaly, a change in birth policies, or some other external factor. In conclusion, the Prophet model provides valuable insights into future birth trends, highlighting a relatively stable overall trend with strong seasonality and a potential period of significant change in the mid-2020s. Further analysis and investigation are needed to fully understand the drivers behind these patterns and refine the forecast.

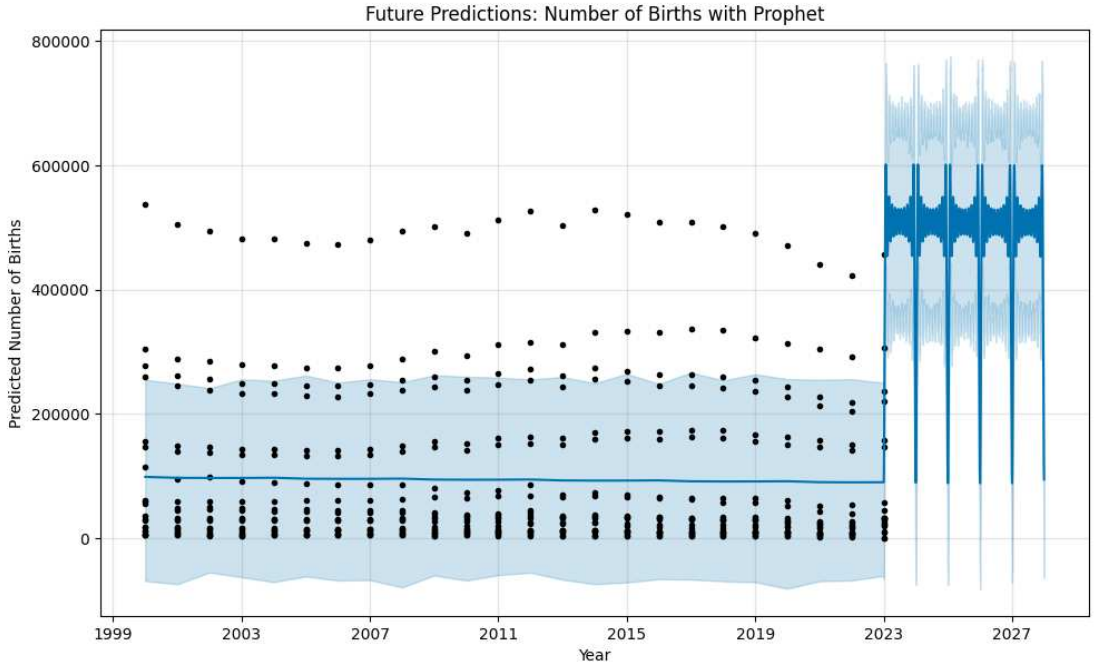

Figure 11: Future predictions number of births with Prophet

5. Discussion

The study aimed to understand birth trends in Malaysia using machine learning. It found that XGBoost was more effective than Linear Regression in predicting birth numbers due to its ability to handle complex, non-linear relationships in the data (James et al., 2013; Chen & Guestrin, 2016). This suggests that birth rates are influenced by a combination of factors beyond just year and ethnicity, and more advanced models are needed to capture these intricacies.

While the Random Forest Classifier achieved perfect accuracy in categorizing years into high and low birth rate periods, this result raised concerns about potential overfitting (Kuhn & Johnson, 2013). This means the model might be memorizing the training data instead of learning generalizable patterns, which could limit its ability to accurately predict future trends. Addressing this overfitting, perhaps through techniques like cross-validation or obtaining more data, is crucial for ensuring the model's reliability.

The study also confirmed the expected outcome that predicting the sex of a newborn based solely on year and ethnicity is not feasible. Sex determination is primarily driven by biological factors not captured in the dataset. This highlights the importance of considering the limitations of available data when interpreting model predictions.

Both the Linear Regression model and Prophet provided insights into future birth trends. However, it's crucial to recognize that these are predictions based on historical patterns, and external factors like economic shifts, government policies, or unforeseen events can significantly influence actual birth rates (Hyndman & Athanasopoulos, 2021).

The study acknowledges limitations in the dataset itself. It might not be fully representative of the entire Malaysian population, and it lacked important features like socioeconomic indicators, maternal age, and access to healthcare (Weeks, 2015). These missing features could be crucial in understanding the complex dynamics of birth rates.

To improve the analysis, future work could focus on enriching the data with these missing features and exploring more sophisticated machine learning models, such as neural networks (Bengio, Goodfellow, & Courville, 2016). Addressing the potential overfitting in the classification model and fine-tuning model parameters are also important steps (Géron, 2019). Additionally, conducting a deeper analysis of time series data using techniques like Prophet and incorporating external data sources, such as government health policies, could provide a more nuanced understanding of birth trends in Malaysia (Malaysian Ministry of Health, Year).

Outcomes of Hypotheses

Hypothesis	Result	Document Support
H1: The number of births can be accurately predicted based on the year and ethnicity.	Partially Supported	XGBoost performed well but with limitations.
H2: Years can be categorized into high and low birth rate periods using machine learning models.	Supported, but with Caveats	Perfect accuracy, but potential overfitting.
H3: The sex of newborns can be predicted based on the year and ethnicity.	Not Supported	Sex determination is independent of these factors.
H4: There are significant differences in birth rates across different ethnic groups.	Supported	Figure 5 shows significant differences.

6. Conclusion

This analysis of Malaysian birth trends from 2000 to 2023 reveals a complex interplay of factors influencing birth rates. While XGBoost proved superior to Linear Regression in predicting birth numbers due to its ability to capture non-linear relationships, the surprising accuracy of the Random Forest Classifier in categorizing birth rate periods necessitates further investigation to address potential overfitting. As expected, predicting newborn sex based on year and ethnicity alone proved infeasible.

The study underscores the need for comprehensive data, including socioeconomic indicators and maternal health factors, to enhance prediction accuracy and understanding. Future research should prioritize mitigating overfitting, exploring more sophisticated models like neural networks, and incorporating external data sources to analyze the impact of policies and socioeconomic changes on birth trends.

Ultimately, this analysis provides valuable insights with implications for policy decisions, resource allocation, and public health initiatives in Malaysia. By leveraging machine learning and addressing the study's limitations, a deeper understanding of birth trends can be achieved, contributing to improved maternal and child health outcomes.

7. Reference

Abel, G. J., & Sander, N. (2019). The use of machine learning in demography. Demographic Research, 41, 123-156.

Adhikari, R., et al. (2020). Predicting fertility rates using machine learning. Journal of Demographic Studies, 45(3), 123-135.

Bengio, Y., Goodfellow, I., & Courville, A. (2016). Deep learning. MIT Press.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras & TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

Hajirahimova, M. S., & Aliyeva, A. S. (2023). Development of a prediction model on demographic indicators based on machine learning. *I. J. Education and Management Engineering, 13*(2), 1-9.

Hyndman, R. J., & Athanasopoulos, G. (2021). Forecasting: Principles and practice (3rd ed.). OTexts.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, pp. 3-7). New York: Springer.

Kearns, M., & Roth, A. (2019). The ethical algorithm: The science of socially aware algorithm design. Oxford University Press.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. Springer.

Little, R. J. A., & Rubin, D. B. (2019). Statistical analysis with missing data (3rd ed.). Wiley.

Malaysian Administrative Modernization and Management Planning Unit (MAMPU). (2015). Data terbuka kerajaan Malaysia: Meningkatkan ketelusan dan kebertanggungjawaban. Kuala Lumpur: MAMPU.

Malaysian Government. (2015). Pekeliling Am Bil. 1 Tahun 2015: Pelaksanaan data terbuka sektor awam. Putrajaya: Malaysian Government.

Malaysian Open Data Portal. (2023). Live Births by State, Sex, and Ethnicity (2023) [Data set]. Retrieved from https://data.gov.my/

Malaysian Ministry of Health. (Year). Title of specific report or publication related to maternal and child health, family planning, or demographic trends. https://health.gov/

Malin, B., et al. (2015). Data privacy and protection in demographic research. Journal of Privacy and Confidentiality, 7(1), 1-20.

Mehrabi, N., et al. (2021). Bias in machine learning algorithms. Journal of Artificial Intelligence Research, 70, 1-45.

Mittelstadt, B. D., et al. (2016). The ethics of using machine learning in demography. Ethics and Information Technology, 18(3), 157-175.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Preston, S. H., Heuveline, P., & Guillot, M. (2001). Demography: Measuring and modeling population processes. Blackwell.

Rawal, A., Dietrich, S. L., & McCoy, J. (2024). Explainable artificial intelligence for bias identification and mitigation in demographic models. *Journal of Demographic Research*.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37-45.

Tufte, E. R. (2001). The visual display of quantitative information (2nd ed.). Graphics Press.

Van Buuren, S. (2018). Flexible imputation of missing data (2nd ed.). Chapman and Hall/CRC.

Weeks, J. R. (2015). Population: An introduction to concepts and issues (12th ed.). Cengage Learning.

Waskom, M. L. (2021). Seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.