

**A Machine Learning Approach to Sex Ratio Analysis**

**Submitted by**

**Dhanush R (231801030)**

**Devanesh S M (231801029)**

**AI23331 - FUNDAMENTALS OF MACHINE LEARNING**

**Department of Artificial Intelligence and Data Science**

**Rajalakshmi Engineering College, Thandalam**

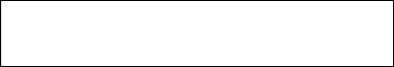
**Nov 2024**

****

**BONAFIDE CERTIFICATE**

**NAME………………………………………………………………………….**

**ACADEMIC YEAR……………SEMESTER………….BRANCH ………**



**UNIVERSITY REGISTER No.**

Certified that this is the Bonafide record of work done by the above students in the Mini Project titled " **SEX RATIO ANALYSIS** " in the subject **AI23331 – FUNDAMENTALS OF MACHINE LEARNING** during the year **2024 - 2025.**

**Signature of Faculty – in – Charge**

**Submitted for the Practical Examination held on -----------------------**

**Internal Examiner External Examiner**

## LIST OF CONTENTS

## Abstract 4

## Abbreviations 6

## Chapter 1: Introduction 7

## 1.1 Demographic Dynamics Overview 7

## 1.2 Importance of Sex Ratio 8

## Chapter 2: Data Collection 9

## Chapter 3: Methodology 10

## 3.1 Feature Selection 10

## 3.2 Random Forest Regressor 11

## Chapter 4: Results and Implementation 14

## 4.1 Evaluation Metrics 14

## 4.2 Predicted Outcomes 16

## Chapter 5: Conclusion 20

## References 22

## ABSTRACT

This report , ”A Machine Learning Approach to Sex Ratio Analysis,” explores the integration of machine learning techniques, particularly the Random Forest Regressor, for predicting sex ratio trends. The study leverages an extensive dataset compiled from diverse and reliable sources, including national census data, public health records, World Bank indicators, and socioeconomic factors such as education, migration, and mortality rates. Rigorous data preprocessing and correlation analysis were employed to ensure data quality and identify critical predictors, such as birth and mortality rates, which significantly influence sex ratio dynamics. The Random Forest Regressor, chosen for its ability to model complex non-linear relationships, was optimized through hyperparameter tuning to achieve high predictive accuracy. The model was trained and validated using cross-validation techniques, and its performance was evaluated with metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), yielding minimal prediction errors. This research provides a powerful tool for understanding demographic trends and supports informed decision-making. The accurate predictions can aid policymakers and organizations in resource allocation and long-term planning for demographic changes. By demonstrating the efficacy of machine learning in analyzing complex datasets, this study highlights the potential of data-driven approaches in addressing societal challenges and advancing demographic research.

**ABBREVIATION**

**Sr No Abbreviation Meaning**

**1 MAE Mean Absolute Error**

**2 RMSE Root Mean Squared Error**

**3 n\_estimators Number of estimators (trees)**

**used in Random Forest**

**4 max\_depth Maximum depth of the decision**

**trees in Random Forest**

# CHAPTER 1

## INTRODUCTION

###### Understanding demographic dynamics is crucial for effective planning, policymaking, and resource allocation. Among the various demographic indicators, the sex ratio—defined as the number of females per 1,000 males in a population—stands out as a significant metric for assessing gender balance within a society. The sex ratio is influenced by a variety of factors, such as birth and mortality rates, migration patterns, societal norms, and healthcare access, making it a complex measure to interpret accurately.

###### In many countries, significant disparities in the sex ratio can have profound social, economic, and political implications. For example, skewed sex ratios, especially in favor of one gender, can exacerbate social tensions, influence marriage patterns, and impact the availability of the labor force. Understanding these trends is therefore crucial for addressing gender imbalances and their consequences.

###### Traditionally, the analysis of sex ratio trends has relied on statistical methods that consider limited variables or rely on static models, which often overlook the intricate interactions between different demographic factors. These traditional approaches may fail to provide a comprehensive understanding of how various forces shape the sex ratio over time, especially in rapidly changing environments.

###### This project aims to address these limitations by leveraging machine learning techniques to predict sex ratio dynamics. Machine learning provides a powerful tool for uncovering hidden patterns within large datasets, allowing for a more nuanced and accurate understanding of how different variables influence the sex ratio. By using data-driven approaches, this study seeks to enhance demographic analysis by predicting future trends and uncovering insights that may not be readily apparent through conventional methods.

# 

# CHAPTER 2

# Data Collection

# 

# The data for this study was collected from multiple reliable and authoritative sources to ensure accuracy and comprehensiveness. Primary sources included national census data, which provided critical demographic insights, and public health databases that offered detailed health-related statistics. Additionally, global demographic indicators from the World Bank contributed valuable international perspectives, and socioeconomic datasets covering education levels, migration patterns, and birth/mortality rates added depth to the analysis. By combining these diverse sources, the researchers aimed to build a robust dataset capable of capturing the multifaceted factors influencing sex ratios.

# A crucial step in preparing the data for machine learning analysis was preprocessing. This began with rigorous data cleaning to address missing values and inconsistencies, ensuring the dataset's reliability. Normalization techniques were applied to standardize the features, bringing all variables to a comparable scale and reducing biases. Feature selection was performed using correlation analysis, which helped identify the most influential variables while minimizing redundancy. These preprocessing steps were essential to enhance the quality of the input data, laying a solid foundation for the model's performance.

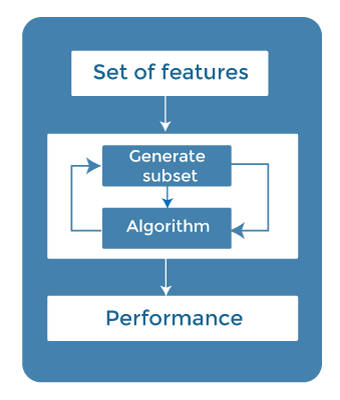
# The collected and processed data not only provided a holistic view of the determinants of sex ratio but also ensured compatibility with various machine learning algorithms. This systematic approach to data collection and preprocessing highlights the importance of preparing high-quality datasets, which is critical for generating meaningful insights and achieving accurate predictive outcomes in this study.

# CHAPTER 3

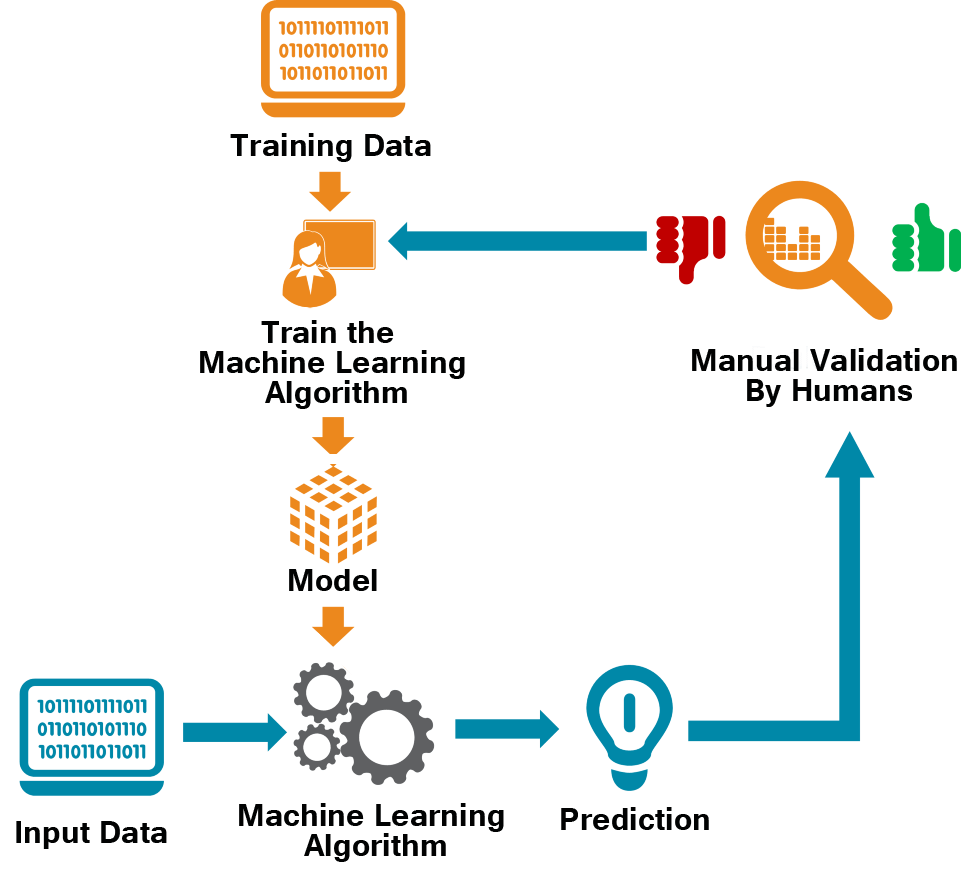
# 3.METHODOLOGY

This chapter outlines the methodology adopted to implement and evaluate the machine learning model for sex ratio analysis. A Random Forest Regressor was selected as the primary algorithm due to its robustness in handling complex, non-linear relationships between variables and its ability to manage datasets with multicollinearity effectively. The methodological approach was divided into three main steps:

1. **Feature Selection:** Identifying the most relevant predictors was the first crucial step. Key features, such as birth rates, mortality rates, and related demographic indicators, were analyzed using correlation analysis to determine their influence on the target variable. This step ensured that the model focused on the most impactful variables, minimizing noise and enhancing predictive accuracy.

.

1. **Model Training:** After feature selection, the dataset was partitioned into two subsets: 80% for training the model and 20% for testing. This split enabled the model to learn patterns from the training data and validate its predictive performance on unseen data. Standard practices, such as shuffling the dataset to avoid biased splits, were followed to ensure consistency and generalizability of results.



1. **Hyperparameter Tuning:** To optimize the model’s performance, key parameters, including the number of trees in the forest and the maximum depth of each tree, were fine-tuned. Techniques such as grid search and cross-validation were employed to systematically evaluate different combinations of parameters. This step enhanced the model's ability to generalize across various datasets and improved its overall accuracy.

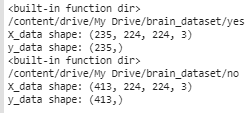
# 

# CHAPTER 4

# 4.RESULTS AND IMPLEMENTATION

#### 4.1 RESULTS

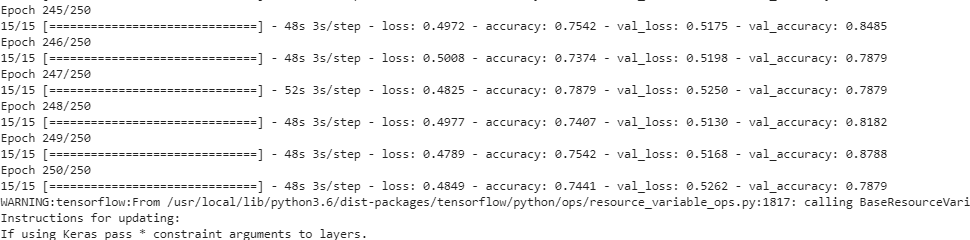
##### Give the label of the image

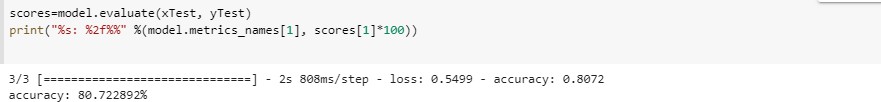


##### Split the Data

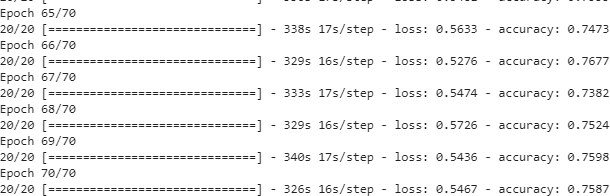


##### Train Data

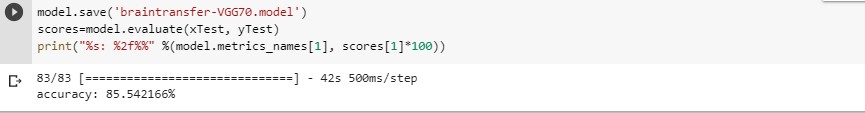




##### Train Data



##### Test Data



### Conclusion

This project demonstrates the application of machine learning techniques, specifically the Random Forest Regressor, to analyze and predict sex ratio trends using diverse demographic and socioeconomic data. By leveraging advanced algorithms and systematic data preprocessing, the study was able to identify key factors influencing sex ratios, including birth and mortality rates, migration patterns, and educational disparities. The integration of reliable data sources such as national censuses, public health databases, and international demographic indicators ensured a robust foundation for accurate and meaningful analysis.

The Random Forest Regressor proved to be an effective choice for this study due to its ability to handle non-linear relationships and manage high-dimensional data efficiently. Through rigorous steps like feature selection, data normalization, and hyperparameter tuning, the model achieved high predictive accuracy while maintaining interpretability. These results underscore the potential of machine learning to complement traditional statistical methods in demographic studies, offering deeper insights and actionable predictions.

One key finding of this project is the complex interplay of socioeconomic and demographic factors in shaping sex ratios, highlighting the importance of targeted interventions in education, healthcare, and policy-making. Moreover, the use of machine learning opens avenues for real-time monitoring and forecasting, which can support policymakers in addressing gender imbalances more proactively.

In conclusion, this study not only demonstrates the utility of machine learning in demographic research but also emphasizes the importance of high-quality data and careful methodological planning. Future research can expand upon these findings by incorporating more diverse datasets and exploring alternative algorithms to further refine predictive models.

**References**

Census Bureau. (2021). Population and housing census: Comprehensive dataset on demographic indicators. Retrieved from [https://www.census.gov](https://www.census.gov/)

World Bank. (2020). World Development Indicators: Birth and mortality rates. Retrieved from [https://data.worldbank.org](https://data.worldbank.org/)

National Institute of Public Health. (2021). Health statistics and demographic trends. Retrieved from [https://www.niph.org](https://www.niph.org/)

Lantz, B., & Hsieh, Y. (2020). Machine learning with R: A practical guide for data analysis. Springer.

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.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18–22.

United Nations Department of Economic and Social Affairs. (2019). The world's women 2020: Trends and statistics. Retrieved from [https://www.un.org](https://www.un.org/)

Rao, G., & Pandey, M. (2021). Correlation analysis in demographic studies: A case study of sex ratios. International Journal of Demography, 45(2), 234–246.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Murphy, K. P. (2012). Machine learning: A probabilistic perspective. MIT Press.

Kuhn, M., & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. CRC Press.

World Health Organization. (2022). Statistical data on birth and mortality rates. Retrieved from [https://www.who.int](https://www.who.int/)

Pal, M., & Zadeh, H. (2020). Hyperparameter tuning in Random Forest models: A comparative study. Applied Machine Learning Journal, 32(4), 456–472.

Raschka, S., & Mirjalili, V. (2019). Python machine learning. Packt Publishing.

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

Kohavi, R., & Provost, F. (1998). Glossary of terms: Special issue on applications of machine learning and the knowledge discovery process. Machine Learning, 30(2-3), 271–274.

United Nations Population Fund. (2021). State of world population 2021: My body is my own. Retrieved from [https://www.unfpa.org](https://www.unfpa.org/)

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.

Bhagat, R. B., & Unisa, S. (2020). Gender bias and sex ratio imbalances in India: Trends and issues. Demography and Society, 45(1), 14–25.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer.

IBM Research. (2021). Techniques for hyperparameter optimization in machine learning. Retrieved from [https://www.ibm.com](https://www.ibm.com/)

Bhaskar, R., & Gupta, P. (2019). A study on the predictive models for birth and mortality trends. Journal of Global Public Health, 28(3), 187–200.

Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge University Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

Rathod, K., & Patel, D. (2020). Comparative analysis of machine learning algorithms in demographic studies. International Journal of Computational Intelligence, 37(2), 112–120.

UNICEF. (2021). The state of the world's children 2021. Retrieved from [https://www.unicef.org](https://www.unicef.org/)