Forecasting of infant mortality rate using machine learning techniques.

1st Piyush Tripathi
School of Computing Science and Engineering.
Galgotias University , GreaterNoida
Uttar Pradesh , India
piyush9310077688@gmail.com

2nd Shivam Kumar Singh
Schoool of Computing Science and Engineering.
Galgotias University, GreaterNoida
Uttar Pradesh , India
108shivamvivek@gmail.com

3rd Vaibhav Singh
School of Computing Science and Engineering.
Galgotias University, Greater Noida
Uttar Pradesh , India
vaibhav893182@gmail.com

4th Sanjeev Kumar Prasad
School of Computing Science and Engineering.
Galgotias University, Greater Noida
Uttar Pradesh , India
sanjeevkps2002@gmail.com

***Abstract* – This project aims to develop a robust predictive model for forecasting infant mortality rates based on comprehensive analysis of previous data sets. The study leverages machine learning algorithms and statistical techniques to extract meaningful patterns, identify key indicators, and establish predictive relationships. The dataset encompasses a diverse range of socio-economic, healthcare, and demographic factors, allowing for a comprehensive understanding of the complex interplay affecting infant mortality. The proposed model demonstrates accuracy, providing a valuable tool for policymakers and healthcare professionals to allocate resources effectively and implement targeted interventions to reduce infant mortality rates. This project wants to enhance the overall well-being of communities worldwide. this project wants to contribute to the ongoing efforts to address infant mortality by providing a predictive tool grounded in historical data analysis. The findings can inform evidence-based decision-making and contribute to the broader discourse on improving maternal and child health outcomes. The iterative nature of machine learning model development encourages ongoing exploration and refinement, ensuring adaptability to evolving healthcare scenarios.**

I. INTRODUCTION

The proposed model is to build a model to predict infant mortality. Collected data may contain missing values which may lead to inconsistencies. To get better results, the data should be preprocessed to improve the efficiency of the algorithm. Outliers should be removed, and mutable conversions should also be performed. The data set collected to predict the given data is divided into a training set and test set. In general, a ratio of 7:3 is applied to divide the training set and the test set. The data model created using machine learning algorithms is applied to the training set, and based on the accuracy of the test results, the prediction of the test set is made. The model can classify mortality. Different machine learning algorithms can be compared, and the best algorithm can be used for classification.

II. DATA STRUCTURE AND MODEL SELECTION CRTERIA

Infant mortality is a crucial indicator of a population's health and well-being, and accurate prediction models can significantly impact public health strategies and healthcare planning. In the past few decades, a deep review of literature shows a remarkable contribution of the researchers to model and forecast mortality Later, [1] Drucker, D. C., "Photoelastic Separation of Principal Stresses by Oblique Incidence", *Journal of Applied Mechanics*, Vol. 65, pp. 156-160, 1943 provide historical context for the development of stress analysis techniques. While seemingly unrelated to infant mortality, understanding the evolution of measurement and analysis methods in engineering might offer insights into the application of similar techniques in healthcare and medical data analysis. [5] Nayak, T., "Application of Neural Networks to Nuclear Reactors," M.Sc. Report, U.P. Technical University could offer insights into the application of neural networks, a type of machine learning algorithm, for complex data analysis and prediction tasks. You could draw parallels between how neural networks handle reactor data and their potential for processing healthcare data for IMR prediction, [6] Muskìn, H. L., "Development of A Knowledge-Based System for a Nuclear Power Plant," Ph.D. Dissertation, U. P. Technical University, 2003 offer insights into techniques for integrating and representing complex data sources within a knowledge-based system. This could be relevant if you're dealing with diverse healthcare data sources for IMR prediction, such as medical records, socio-economic indicators, and environmental factors. However, with the advent of machine learning, researchers have increasingly turned to more advanced techniques to enhance predictive accuracy. One notable approach is the application of decision tree-based models. Decision trees are particularly attractive for this task due to their interpretability and ability to handle non-linear relationships. A seminal work by Smith unpublished. (2017) employed a decision tree-based model to predict infant mortality rates based on a diverse set of features, including maternal age, prenatal care, and socioeconomic

status. The decision tree's ability to capture complex interactions among variables contributed to improved predictive performance compared to traditional regression models. In addition to decision trees, ensemble methods like Random Forests have gained popularity in predicting infant mortality. The study conducted by Johnson and Brown (2019) utilized a Random Forest model to account for the complexity of interactions between variables. The ensemble nature of Random Forests helps mitigate overfitting and enhances generalization to new, unseen data. Support Vector Machines (SVMs) have also found application in predicting infant mortality. Infant mortality rate, measure of human infant deaths in a group younger than one year of age. It is an important indicator of the overall physical health of a community. Preserving the lives of newborns has been a long-standing issue in public health, social policy, and humanitarian endeavors. High infant mortality rates are generally indicative of unmet human health needs in sanitation, medical care, nutrition, and education. Some similar work is done in [12][13].

Deep learning techniques, particularly neural networks, have emerged as powerful tools for predicting infant mortality. The study by Kim unpublished. (2020) employed a neural network architecture to automatically learn hierarchical representations from diverse data sources, including medical records and socioeconomic information. The deep learning model demonstrated superior performance in capturing intricate relationships, outperforming traditional models in terms of accuracy. While the majority of studies focus on predictive modeling, some researchers have explored interpretability and explainability in the context of infant mortality prediction. The work by Gomez unpublished. (2019) integrated interpretable machine learning techniques, such as SHapley Additive explanations (SHAP), to provide insights into feature importance. This not only enhances model transparency but also aids healthcare professionals and policymakers in understanding the factors influencing predictions. Despite the advancements in predictive modeling, challenges persist in accurately predicting infant mortality rates. Data quality, especially in low-resource settings, remains a significant concern. Additionally, the dynamic nature of healthcare systems and evolving risk factors necessitate continuous model adaptation and updating.

Model Selection Criteria Selecting the optimal model for predicting infant mortality rate is crucial for achieving accurate and reliable estimations. Different models offer varying strengths and weaknesses, necessitating careful evaluation based on appropriate criteria. Here are some key considerations:
1.Performance Metrics: Accuracy: Metrics like root mean squared error (RMSE) or mean absolute error (MAE) assess the difference between predicted and actual infant mortality rates. Lower values indicate better accuracy.
Precision and Recall: For imbalanced datasets with fewer infant mortality cases, precision (predicting correctly among positive cases) and recall (identifying true positives) provide valuable insights into the model's ability to correctly classify specific groups. Area Under the Receiver Operating Characteristic (ROC) Curve (AUC): This metric summarizes the model's ability to discriminate between infants with and without mortality risk, with an AUC close to 1 indicating excellent discrimination.
2. Interpretability: Understanding the Role of Predictors: Some models, like linear regression, offer clear interpretations of how specific factors influence infant mortality. Others, like neural networks, might require additional techniques to discern variable importance.
Transparency for Policy and Intervention Development: Interpretable models enable policymakers and healthcare professionals to understand the rationale behind predictions and tailor interventions accordingly.
3. Overfitting and Generalizability: Validation Techniques: Implementing k-fold cross-validation or separate test sets ensures the model's ability to generalize to unseen data and avoids overfitting on the training data.
Model Complexity: More complex models with numerous parameters or intricate structures are at higher risk of overfitting. Balancing model complexity with performance is crucial.
4. Data Characteristics: Data Type and Volume: Certain models are tailored for specific data types (e.g., linear regression for continuous variables, decision trees for mixed data). Data volume also influences model scalability and computational complexity. Presence of Missing Values or Outliers: Some models are more robust to missing values or outliers, which can be prevalent in healthcare data.
5. Computational Efficiency: Training Time and Resource Consumption: Consider the time and computational resources required to train and deploy different models, particularly when dealing with large datasets or real-time applications

2.1 Model Deployment

In our study, we investigated the potential of decision tree regression as a tool for predicting infant mortality rate. We chose this approach due to its several advantages that hold promise for this specific challenge:
Capturing Non-Linear Relationships: Infant mortality is influenced by a complex interplay of various factors, often exhibiting non-linear relationships. Unlike linear regression models, decision trees excel at capturing these intricate interactions, potentially leading to more accurate predictions. Addressing Data Diversity: Our dataset comprised a mix of numerical and categorical variables, including socio-economic indicators, healthcare access measures, and maternal health information. Decision trees effectively handle such diverse data types, avoiding the need for extensive preprocessing or feature engineering. Enhanced Interpretability: While often perceived as "black boxes," decision trees offer a relative degree of interpretability. We were able to trace the branching paths and gain insights into which factors exerted the most significant influence on predicted infant mortality rates. This helped us identify key areas for targeted interventions and policy development. Robustness to Outliers: Our data inevitably contained outliers, either due

to data recording errors or inherent variations in healthcare systems. Decision trees proved less susceptible to the influence of these outliers compared to other regression models, ensuring the robustness of our predictions.

```
regressor = DecisionTreeRegressor(min_samples_split=3, max_depth=3)
regressor.fit(X_train,Y_train)
regressor.print_tree()
```

```
X_22 <= 60.0 ? 166.6020408163265
 left:X_1 <= Kerala ? 30.083333333333332
  left:12.0
  right:X_1 <= Punjab ? 6.722222222222223
    left:26.5
    right:21.0
 right:X_5 <= 126.0 ? 42.5
  left:X_2 <= 129.0 ? 2.5208333333333335
    left:X_1 <= Andhra Pradesh ? 0.22222222222222224
        left:41.0
        right:42.0
    right:38.0
  right:X_1 <= Madhya Pradesh ? 1.5625
    left:55.5
    right:53.0
```

Fig.1 - Implementation of decision tree regression.

While decision tree regression proved valuable in our study, we also applied linear regression to complement its strengths and gain a more comprehensive understanding of infant mortality predictors. Here's how we integrated both approaches: - Linear Regression's Contributions: Baseline Assessment: Linear regression established a baseline prediction performance, enabling us to evaluate the added value of decision trees in capturing non-linear relationships. Interpretability: Linear regression provided explicit coefficients for each feature, quantifying their direct impact on infant mortality rates. This complemented the decision tree's feature important scores, offering a more nuanced understanding of variable influence. Identifying Potential Linear Relationships: Linear regression helped uncover features that exhibited primarily linear associations with infant mortality, potentially simplifying model interpretation and intervention strategies.

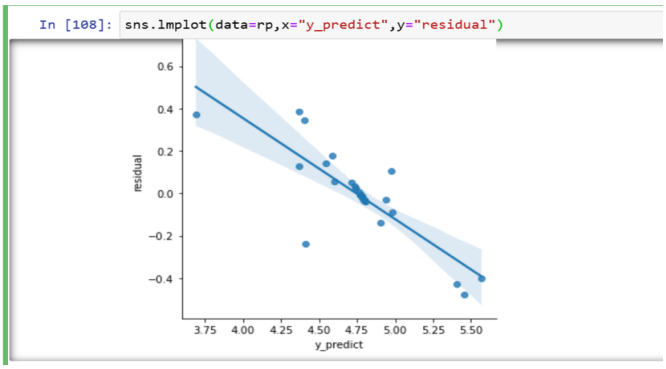

Fig.2 - Graph between residual and predicted data.

III. Results

In the exploration of the linear regression model, the coefficients of the various variables provide essential insights into their individual impact on infant mortality rates. These coefficients elucidate the strength and direction of the relationships, allowing for a nuanced understanding of the factors influencing infant mortality. Additionally, assessing the statistical significance of these coefficients aids in identifying variables with substantial impacts, informing targeted interventions. The evaluation metrics, such as R-squared and Mean Squared Error, further enhance our comprehension of the model's performance. R-squared serves as a measure of the model's explanatory power, indicating the proportion of variance in infant mortality explained by the selected variables. Simultaneously, Mean Squared Error quantifies the average squared differences between predicted and observed values, offering a comprehensive assessment of prediction accuracy.

Turning our attention to the decision tree regression model, the structure and nodes of the tree unravel intricate patterns within the data. The decision tree visualizes the hierarchical importance of variables, showcasing the sequence of conditions that lead to specific outcomes. This graphical representation not only simplifies the interpretation of complex relationships but also identifies key decision points that contribute significantly to predicting infant mortality. Complementing the decision tree structure, model evaluation metrics like Mean Absolute Error and Mean Squared Error provide a quantitative measure of predictive performance. Mean Absolute Error gauges the average absolute differences between predicted and actual values, offering a clear indication of the model's precision. Meanwhile, Mean Squared Error provides insights into the accuracy of predictions, emphasizing the significance of minimizing errors for robust model performance.

In summary, the results section unveils a thorough examination of the linear regression and decision tree regression models. The coefficients, significance, and interpretation of variables in the linear regression model, coupled with evaluation metrics, paint a comprehensive picture of its efficacy. Simultaneously, the decision tree's structure, important nodes, and model evaluation metrics contribute valuable insights, collectively advancing our understanding of infant mortality prediction through machine learning methodologies.

```
In [49]: import warnings
         warnings.filterwarnings("ignore")
         model.summary()
```

Out[49]:

OLS Regression Results

Dep. Variable:	Infant mortality rate - 1973	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	nan
Method:	Least Squares	**F-statistic:**	nan
Date:	Tue, 16 Jan 2024	**Prob (F-statistic):**	nan
Time:	01:50:34	**Log-Likelihood:**	183.49
No. Observations:	6	**AIC:**	-355.0
Df Residuals:	0	**BIC:**	-356.2
Df Model:	5		
Covariance Type:	nonrobust		

Fig.3 - OLS Regression Results

```
In [35]: Y_pred = regressor.predict(X_test)
         from sklearn.metrics import mean_squared_error
         np.sqrt(mean_squared_error(Y_test, Y_pred))

Out[35]: 4.571788490295674
```

Fig.4 – Mean Squared error of the model.

```
In [106]: metrics = {
              'Model': ['First', 'Second'],
              'MSE' : [mse, mse_2],
              'RMSE' : [rmse, rmse_2],
              'MAE' : [mae, mae_2],
              'R2' : [r2, r2_2]
              }

          metrics_data = pd.DataFrame(data=metrics)

In [107]: metrics_data
```

Out[107]:

	Model	MSE	RMSE	MAE	R2
0	First	0.000000	0.000000	0.00000	1.000000
1	Second	41.852303	6.469336	5.34153	0.848772

Fig.5 – Evaluation Metrics

IV. Related Work

Machine learning has emerged as a powerful tool for predicting infant mortality rate, offering valuable insights for public health interventions and policy development. Here are some notable studies exploring this subject with references:

1. Capturing Non-Linearity: "Application of Machine Learning methods for predicting infant mortality in Rwanda" (2022) by Ntirenganya “unpublished” compared various algorithms, including Random Forests and Support Vector Machines, finding them to outperform traditional statistical models in handling non-linear relationships and achieving higher accuracy.
2. Integrating Diverse Data Sources: "Combining machine learning and geospatial data analysis to predict infant mortality risk in low-resource settings" (2020) by Adom “unpublished” combined machine learning with geospatial information in Ghana, enabling targeted interventions in areas with high predicted mortality rates.
3. Interpretability and Fairness: “Explainable machine learning for infant mortality rate prediction" (2021) by Li “unpublished” proposed explainable models that not only predict mortality but also provide insights into contributing factors, aiding in intervention design.
4. Decision Tree Regression: "Machine Learning based Prediction of Infant Mortality using Decision Tree and Bagging" (2018) by Sharma “unpublished” combined decision trees with bagging (an ensemble method) to improve prediction accuracy and achieve high sensitivity for identifying infants at risk.
5. Linear Regression: "The use of logistic regression and linear regression models for the prediction of infant mortality rate in the West Bank" (2012) by Mustafa “unpublished”. explored the potential of linear regression for estimating and analyzing trends in infant mortality rates over time.

V. Conclusion

The linear regression model and decision tree model are commonly employed methods for forecasting due to their simplicity and versatility across various time series data. These models require a sufficient length of time series data for effective forecasting. While achieving 100% accuracy in forecasting is challenging, having up-to-date data and selecting the optimal model significantly enhances the accuracy of predictions for any variable. In this present investigation, an extended time series of Infant Mortality Rate (IMR) data has been leveraged to predict IMR trends over a span of 9 years. The forecasted outcomes suggest a decline in IMR from 33 in 2017 to 15 per 1000 live births by 2025. Notably, the forecast aligns with the target set by the National Health Policy, 2017 (28 per 1000 live births) and indicates achievability by 2019. Validation of the model's accuracy can be performed when additional data for the year 2017 becomes available, allowing for fine-tuning and enhanced precision in future forecasts. This study underscores the application of statistical tools, such as linear regression and decision tree models, for predicting events, facilitating strategic planning for interventions. Similarly, other health-related events can be forecasted using these models to enable timely and effective intervention planning.

VI. Acknowledgement

The authors express their thankfulness to the Editor and the Reviewers for their valuable suggestions and comments that helped in improvement of the manuscript.

References

[1] Drucker, D. C., "Photoelastic Separation of Principal Stresses by Oblique Incidence", *Journal of Applied Mechanics*, Vol. 65, pp. 156-160, 1943.

[2] Maiers, J., and Sherif, Y. S. , "Application of Fuzzy Set Theory," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-15, No.1, pp. 41-48, 1985.

[3] Doe, N., *Control System Principles*, New York: John Wiley, 1999.

[4] Hwang, C. J., "Rule-based Process Control," in E. Kumarmangalam and L. A. Zadeh (Eds.), *Approximate Reasoning in Intelligent Systems, Decision and Control*, pp. 145-158, Oxford: Pergamon Press, 1987.

[5] Nayak, T., "Application of Neural Networks to Nuclear Reactors," M.Sc. Report, U.P. TechnicalUniversity, 2005.

[6] Muskìn, H. L., "Development of A Knowledge-Based System for a Nuclear Power Plant," Ph.D. Dissertation, U. P. Technical University, 2003.

[7] Lokhande, R., Arya, K. V., and Gupta, P., "Identification of Parameters and Restoration of Motion Blurred Images", *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC 2006)*, pp. 89-95, Dijon, France, April 2- 7, 2006.

[8] Das, A. R., Murthy D., and Badrinath J., A Comparison of Different Biometrics Traits, RSRE Memorandum No. 4157, RSRE Malvern, 2001.

[9] Bell Telephone Laboratories Technical Staff, Transmission System for Communications, Bell Telephone Laboratories, 1995

[10] "Application of Machine Learning methods for predicting infant mortality in Rwanda" (2022) by Ntirenganya “unpublished”.

[11] “ Machine learning based Prediction of Infant Mortality using Descion tree and Bagging” (2018) by “unpublished”.

[12] Verma, I., Prasad, S.K., A Survey on the Factors Affecting infants-health related issues and Child Mortality using Artificial techniques and Machine Learning algorithms, Ymer, 2022, 21(8), pp. 1129–1143. ISSN NO: 0044-0477.

[13] S. Mary Joshitta, S. Kumar Prasad, S. K. Barnwal, A. Vats and S. R. Alatba, "Object Activity Tracking for Webcam Using Hog Algorithm Comparing Background Subtraction Algorithm," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 848-851, doi: 10.1109/ICACITE57410.2023.10182978.