Predicting Childhood Routine Immunization Status in Ethiopia Using Ensemble Machine Learning Algorithms

Alexander Takele Mengesha[e,1], Muhammed Alkader[f,1], Assefa Chekole Addis[f]

[a]*Department of Information Science, University of Gondar, 196, Ethiopia*
[b]*Department of Information Science, University of Gondar, 196, Ethiopia*
[c]*Department of Information Science, University of Gondar, 196, Ethiopia*

Abstract

The study aims to develop a predictive model for assessing childhood immunization status in Ethiopia using ensemble machine learning techniques.The research follows an experimental approach. Data from the Ethiopian Demographic and Health Survey (EDHS), collected at five-year intervals, was preprocessed for quality assurance. The study employed several ensemble machine learning algorithms, including Extreme Gradient Boosting (XGBoost), CatBoost, Random Forest (RF), and Gradient Boosting, with a One-Versus-Rest class decomposition method. A total of 35,512 instances with 18 features were used, with an 80/20 training/testing dataset split. The models were evaluated based on accuracy, with XGBoost achieving the highest performance at 88.30%, followed by RF (87.17%), CatBoost (86.92%), and Gradient Boosting (84.16%). SHAP (Shapley Additive Explanations) values were used to identify the most significant factors influencing immunization status, including child's age, region, mother's occupation, current parity, and mother's age.Using XGBoost and SHAP values extracted decision rules based on feature importance. These rules reveal specific patterns related to immunization status, providing evidence-based insights for policymakers.XGBoost was selected as the best predictive model for childhood immunization status in Ethiopia. The study highlights key factors for targeted vaccination programs and scheduling, emphasizing the importance of early immunization and maternal characteristics in improving child vaccination coverage.

Keywords: Childhood Immunization, Ensemble Machine Learning, Vaccination, Predictive Model, Explainable AI

Highlights

Predicting Childhood Routine Immunization Status in Ethiopia Using Ensemble Machine Learning Algorithms

Alexander Takele Mengesha, Muhammed Alkader, Assefa Chekole Addis

- Identification of significant factors influencing childhood routine immunization status in Ethiopia.
- Development of a predictive model using ensemble machine learning algorithms for immunization status prediction.
- Implications for improving vaccination programs and formulating data-driven policies.

Email addresses: alexander.takele@uog.edu.et (Alexander Takele Mengesha), muhammedalkader@gmail.com (Muhammed Alkader), assefa.chekole@uog.edu.et (Assefa Chekole Addis)

Predicting Childhood Routine Immunization Status in Ethiopia Using Ensemble Machine Learning Algorithms

Alexander Takele Mengesha[e], Muhammed Alkader[f,1,], Assefa Chekole Addis[f]

[d]*Department of Information Science, University of Gondar, 196, Ethiopia*
[e]*Department of Information Science, University of Gondar, 196, Ethiopia*
[f]*Department of Information Science, University of Gondar, 196, Ethiopia*

Keywords: Intimate Partner Violence, Machine Learning, Data Analytics, SHAP Explainable AI, Data-driven Policy Formulation

1. Introduction

Childhood immunization is one of the most cost-effective public health interventions. Basic immunizations are estimated to prevent more than 2.5 million annual child deaths globally, primarily due LinkedIn, Medium, Portfolio 6th Deep Learning Indaba Conference (DLI 2024).to the prevention of measles, pertussis, and tetanus. Vaccination of children also is expected to avert adult deaths by preventing hepatitis B virus (HBV)-related cirrhosis and liver cancer and human papillomavirus (HPV)-related cervical cancer [1]. According to a World Health Organization (WHO) study in 2022, 14.3 million infants missed their first DTP vaccine dose, with an additional 6.2 million partially vaccinated. Almost 60 of these children reside in Angola, Brazil, Ethiopia, India, Indonesia, Mozambique, Nigeria, Pakistan, and the Philippines, highlighting pressing issues in healthcare accessibility [2]. Even while free routine vaccinations are available in low and middle-income countries (LMICs), many children miss receiving all recommended vaccinations, receive them too late for their age, or drop receiving them altogether[3]. To achieve the Millennium Development Goal Four (MDG4) of reducing children's deaths by two-thirds in 2015, Ethiopia has adopted strategies such as sustainable outreach service and reaching every district that focuses on identifying bottlenecks and developing community ownership of the services to improve routine immunization services and increasing coverage[1]. The complete vaccination coverage varies across administrative regions ranging from 21% in the Afar regional state to 89% in the Amhara regional state. Despite promising improvements in child vaccination coverage in Ethiopia since 2011, due to its large size population, the country still has many unvaccinated children and there are huge variations in immunization coverage across regions [4]. Recently, machine learning has been used to predict healthcare outcomes including cost, utilization, and quality [10] [11]. It has also been used to predict which patients are most likely to experience hospital re-admission for congestive heart failure and related conditions [12]. To solve this problem, significant numbers of research has been performed. Some of them were Fareeha Sameen et al [], Hiwot Abebe [19], Abadi Girmay [8], and Tenaw Gualu [7]. Most of the statistical researches conducted on child immunization were considered to identify the determinant factors and most of the previous research conducted using data mining considered the vaccination types of BCG, DPT-HepB-Hib, polio, and Measles, however, the government of Ethiopia introduced the pneumococcal conjugate vaccine (PCV) and monovalent human rotavirus vaccine (RV) into the national infant immunization program in November 2011 and October 2012, respectively. Most of the study relies on data obtained from a single source, they did not include some of the

Email addresses: alexander.takele@uog.edu.et (Alexander Takele Mengesha), muhammedalkader@gmail.com (Muhammed Alkader), assefa.chekole@uog.edu.et (Assefa Chekole Addis)

potential factors such as the influence of cultural beliefs and practices, access to mass media, or healthcare providers' attitudes on a child's immunization status, used classic machine learning algorithms, and the performance of the developed predictive model was not high, which may not be reliable enough for making critical decisions and they didn't consider the explainability of the child immunization prediction model. Besides, the lack of previous research that used different ensemble machine learning algorithms to develop a child Motivated by these gaps, our study seeks to address the underrepresentation of certain vaccine types (PCV and RV) in the Ethiopian national immunization program and incorporate cultural beliefs, and mass media access. We aim to develop an ensemble machine learning predictive model tuned for optimal performance, focusing on explainability to inform targeted strategies and policies for enhancing child immunization in Ethiopia. To this end, this study aims to investigate the following research questions: Which Ensemble machine learning algorithm is best for predicting a child's routine immunization status? What are the determinant features that contribute to the lowering of child immunization in each region of Ethiopia? What are the underlying patterns and decision rules learned by the ensemble learning model for child routine immunization status? The rest of this document is organized as follows: Section II related works, Section III materials and methods used, Section IV experimental setup and result discussion, and Section V presents the conclusion.

1.1. Related work

Research on predicting childhood immunization status has utilized machine learning and statistical tools to enhance routine immunization determinants, coverage, and vaccination rates [11] [12]. Such as Fareeha Sameen applied supervised ML algorithms to forecast immunization adherence and analyze contributing factors, achieving notable accuracy with the random forest model. However, the study's reliance on a single data source limits its representativeness, suggesting the exploration of additional ML techniques for improved performance and external validation of the model. Hiwot Abebe employed data mining techniques to assess infant immunization in Ethiopia, employing four classification algorithms. Despite identifying key factors, the study's accuracy and consideration of additional factors like cultural beliefs and provider attitudes suggest potential improvements. Abadi Girmay examined immunization coverage in the Sekota Zuria district, finding positive correlations with antenatal care, maternal education, and proximity to health facilities. However, the study's 2limitations include potential recall bias and the inability to establish causal relationships, indicating a need for more comprehensive approaches. Tenaw Gualu and Abdinasir Abdullahi Jama explored vaccination coverage in Debre Markos and Somalia, respectively, highlighting factors like child sex, maternal care, and distance to health facilities. Neeta Singh and Merga Dheresa further investigated immunization coverage about socio-demographic variables, revealing significant associations with maternal education, birth order, and place of delivery. These previous statistical studies [89] [7] [90] [91] [92] used local clinical data, and covered limited geographical areas. So, we are motivated to fill these gaps, by developing a predictive model using homogeneous ensemble machine learning algorithms to develop a predictive model of a child's immunization status in Ethiopia with more data that was previously undone.

2. Methodology

2.1. Data Collection and Preprocessing

In this study, an experimental research design approach was implemented based on the flow chart presented in Fig. 1. The data was obtained from the Ethiopian Demographic Health Survey (EDHS) which was collected in 2016 and 2019 by the Ethiopian Central Statistical Agency (ECSA). Data preprocessing tasks such as data cleaning, data transformation, feature encoding, class balancing, and feature engineering were employed to get the best results. Missing

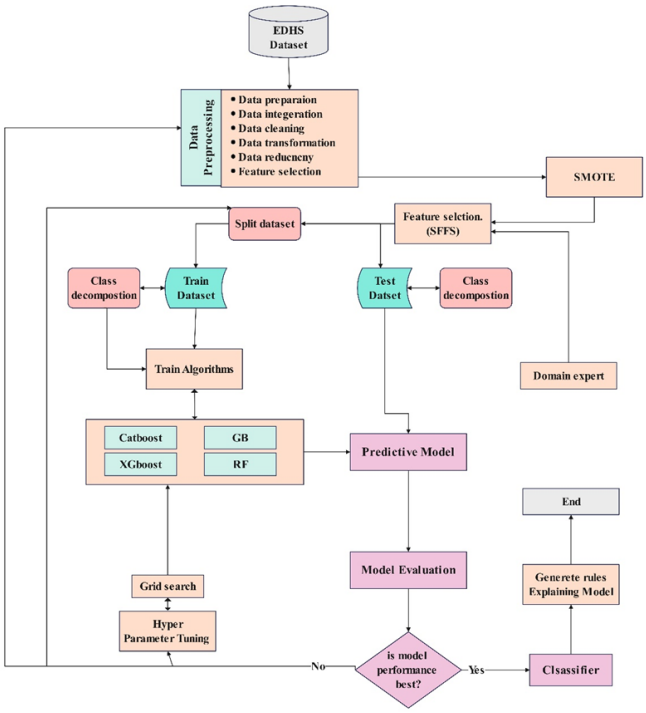

Figure 1: Proposed Model Architecture

values were handled using mode imputation and mean imputation techniques for categorical data and numerical data, respectively. The ordinal encoding techniques were used to encode the categorical features. Data discretization and data normalization techniques were applied to attributes with continuous values to minimize distinct values. Feature engineering methods namely forward and backward feature selection methods were selected for the experiment. From these experiments, sequential forward selection achieves the highest performance with 81.55% accuracy using a random forest classifier, and the 18 best features were selected from 46 features (see Table 1). The training dataset's class level was imbalanced, which needed to be treated using class-balancing techniques to avoid biased learning. The researcher used the Synthetic Minority Oversampling Technique (SMOTE) and the dataset increased from 16394 instances to 35511 records without any duplicate instances of each class. Finally, a total of 35511 instances with 18 attributes were selected for further analysis and prediction model development, and then the dataset was split into training and testing datasets based on the 80/20% ratio.

2.2. Data Collection and Preprocessing

In this study, an experimental research design approach was implemented based on the flow chart presented in Fig. 1. The data was obtained from the Ethiopian Demographic Health Survey (EDHS) collected in 2016 and 2019 by the Ethiopian Central Statistical Agency (ECSA). Data preprocessing tasks such as data cleaning, data transformation, feature encoding, class balancing, and feature engineering were employed to achieve the best results. Missing values were handled using mode imputation and mean imputation techniques for categorical and numerical data, respectively. The ordinal encoding technique was used to encode categorical features. Data discretization and normalization techniques were applied to attributes with continuous values to minimize distinct values.

Feature engineering methods, namely forward and backward feature selection methods, were used for the experiment. Sequential forward selection achieved the highest performance with 81.55% accuracy using a random forest classifier, and 18 best features were selected from 46 features. These selected features are presented in Table 1.

The training dataset's class level was imbalanced, requiring class-balancing techniques to

avoid biased learning. The Synthetic Minority Oversampling Technique (SMOTE) was employed, increasing the dataset from 16,394 instances to 35,511 records without any duplicate instances of each class. Finally, a total of 35,511 instances with 18 attributes were selected for further analysis and prediction model development, and the dataset was split into training and testing datasets based on an 80/20% ratio.

Table 1: Selected Features

No	Forward Selection	Backward Selection
0	Women's age in groups	Women's age in groups
1	Region	Region
2	Highest educational level	Highest educational level
3	Religion	Religion
4	Ethnicity	Ethnicity
5	Age of household head	Age of household head
6	Wealth index	Literacy
7	Total children	Wealth index
8	Currently breastfeeding	Number of living children
9	Distance to a health facility	Total children
10	Husband's education level	Currently breastfeeding
11	Husband's occupation	Distance to a health facility
12	Women's occupation	Husband's occupation
13	Sex of child	Women's occupation
14	Current child of age	Husband's age
15	Place of delivery	Current child of age
16	The child's age in months	Place of delivery
17	Media	The child's age in months

2.3. Predictive Model Development

To construct a predictive model, we employed Ensemble Machine Learning Algorithms such as Extreme Gradient Boosting (XGBoost), CatBoost, Random Forest (RF), and Gradient Boosting Machine (GBM). The experiment utilized a one-vs-the-rest class decomposition approach. These algorithms were chosen for their ability to achieve optimal performance by mitigating bias, variance, and overfitting issues. XGBoost employs regularization techniques and early stopping to minimize overfitting and enhance prediction accuracy. CatBoost excels in handling data with categorical features, improving model performance while reducing overfitting and minimizing parameter tuning time with fast predictions. RF is effective with both categorical and continuous values, reducing overfitting and sensitivity to outliers, thereby improving prediction accuracy. The GBM algorithm is an efficient learner, known for its quick training and high performance.

2.4. Predictive Model Development

To construct a predictive model, we employed Ensemble Machine Learning Algorithms such as Extreme Gradient Boosting (XGBoost), CatBoost, Random Forest (RF), and Gradient Boosting Machine (GBM). The experiment utilized a one-vs-the-rest class decomposition approach. These algorithms were chosen for their ability to achieve optimal performance by mitigating bias, variance, and overfitting issues. XGBoost employs regularization techniques and early stopping to minimize overfitting and enhance prediction accuracy. CatBoost excels in handling data with categorical features, improving model performance while reducing overfitting, and minimizing parameter tuning time with fast predictions. RF is effective with both categorical and continuous values, reducing overfitting and sensitivity to outliers, thereby improving prediction accuracy. The GBM algorithm is an efficient learner, known for its quick training and high performance.

3. Experimental Results and Discussion

3.1. Which Ensemble machine learning algorithm is best for predicting a child's routine immunization status

To identify the best-performing model, a set of ensemble machine learning algorithms such as XGBoost, CatBoost, GB, and RF was used with grid search with a 5-fold cross-validation to optimize the model's performance with their best hyperparameters. The performance of the model was evaluated using evaluation metrics accuracy, ROC, F1 Score, Precision, and Recall. Based on the experiment, the XGBoost algorithm outperformed the others across all considered metrics. So, XGBoost was identified as the most suitable ensemble machine learning algorithm for developing a predictive model RIS in this study. The following figure 2 defines the performance of each model.

Table 1: Comparison of Ensemble Machine Learning Models

Evaluation Metrics	**CatBoost (%)**	**RF (%)**	**XGBoost (%)**	**Gradient Boost (%)**
Accuracy	86.92	87.17	88.30	84.16
Precision	87.35	87.66	88.42	84.82
Recall	86.97	87.21	88.33	84.19
F1-score	86.83	87.05	88.22	84.12
ROC	96.19	96.02	96.64	95.02

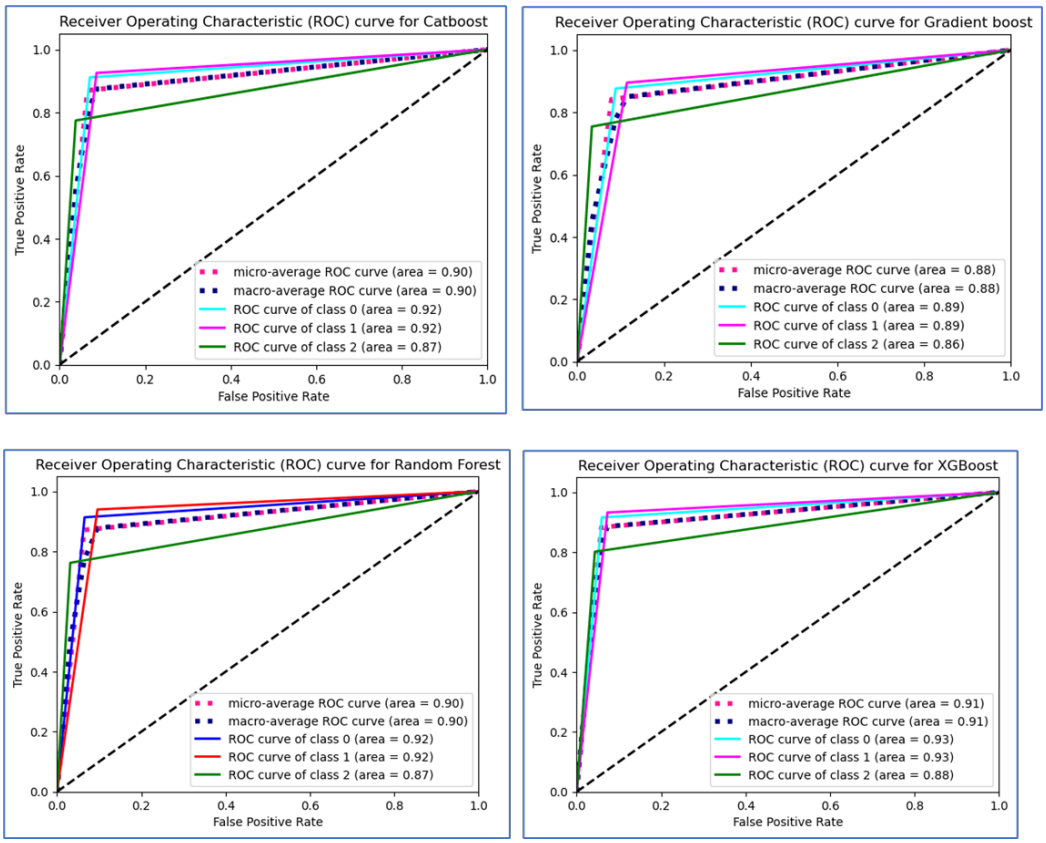

Figure 2: Performance comparison chart

3.2. What are the determinant features that contribute to the lowering of child immunization status in Ethiopia?

This experiment aims to identify the influential factors that impact predictive outcomes. To get the most significant factors, we conducted an experiment measuring the importance of all

features in the dataset through a feature importance analysis using SHapely. Hence XGBoost is the most effective predictive model for determining RI. To further explain the XGBoost model, we employed SHapley Additive exPlanations (SHAP) explainable AI methods. The following figure defines the significant factors of each variable in the model.

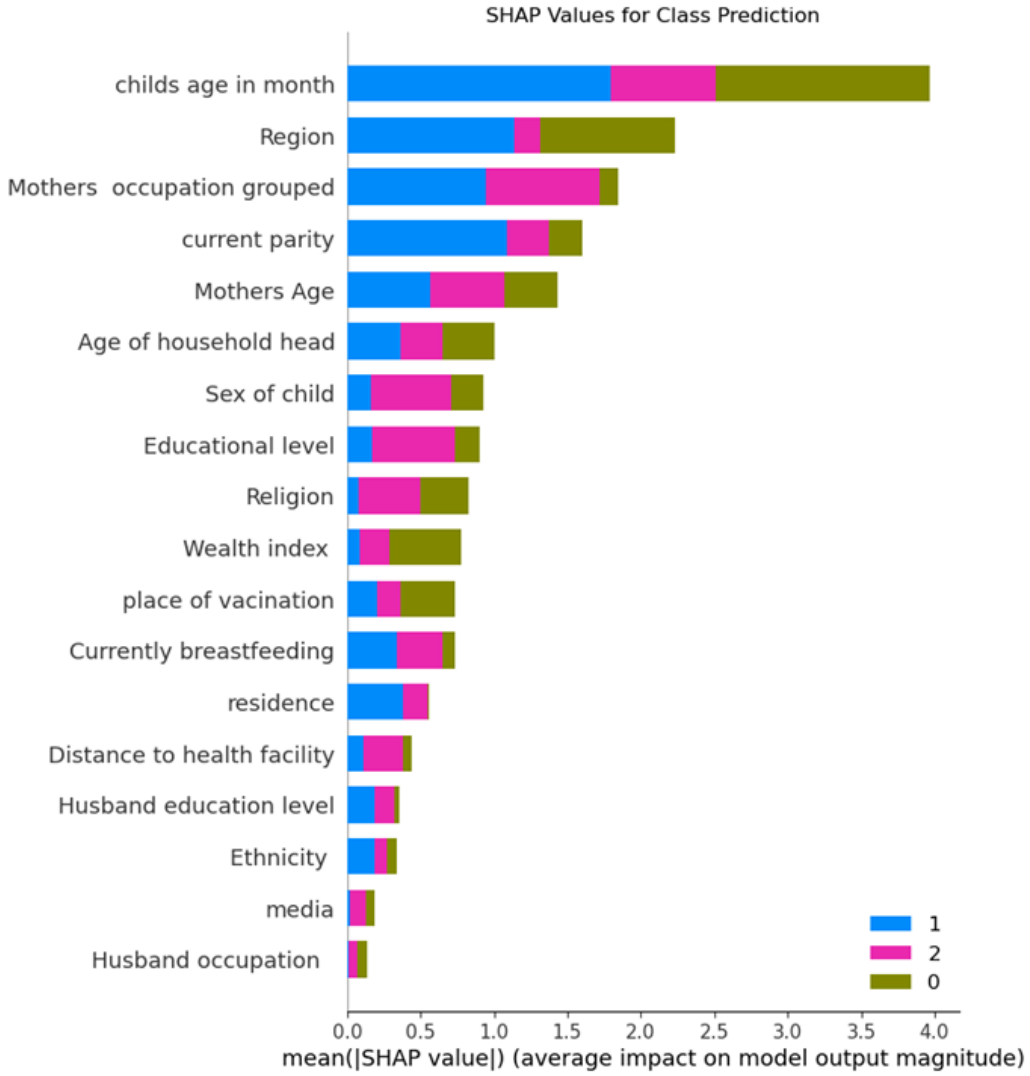

Figure 3: Performance comparison chart

3.3. How does the ensemble model (XGBoost) identify key factors influencing childhood immunization status?

Using XGBoost and SHAP values, we extracted decision rules based on feature importance. These rules reveal specific patterns related to immunization status, providing evidence-based insights for policymakers. This transparent approach enhances understanding and informs decision-making. The following are some of the most important rules/patterns extracted from the XGBoost predictive model.

As shown in Figure 4, the following rule predicts the class **Partially Immunized**:

IF Region = 1 & Age of household head = 4 & Child's age in months = 3 & Educational level = 3 & Place of vaccination = 2 & Religion = 0 & Wealth index = 1 & Ethnicity = 3 & Mother's occupation = 0 & Mother's age = 4 & Distance to health facility = 1 & Sex of child = 2 & Husband's education level = 0 & Residence = 2 & Current parity = 2 & Current breastfeeding = 0 & Media = 0 & Husband's occupation = 1 **THEN** Partially Immunized.

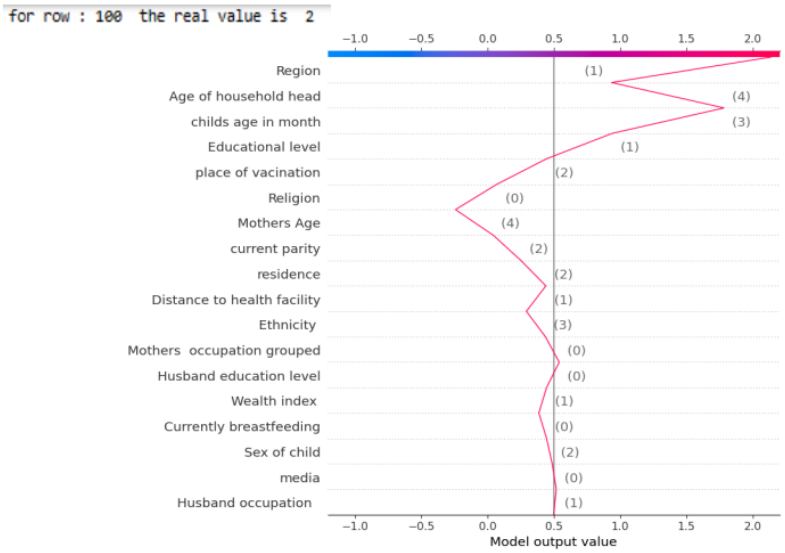

Figure 4: Sample rule predicting **Partially Immunized** (Class 2).

IF Region = 5 & Religion = 1 & Place of vaccination = 1 & Wealth index = 1 & Ethnicity = 2 & Mother's occupation = 0 & Mother's age = 2 & Child's age in months = 3 & Educational level = 1 & Distance to health facility = 1 & Age of household head = 1 & Sex of child = 1 & Husband's education level = 0 & Residence = 2 & Current parity = 2 & Current breastfeeding = 0 & Husband's occupation = 1 & Media = 0 **THEN** Not Immunized.

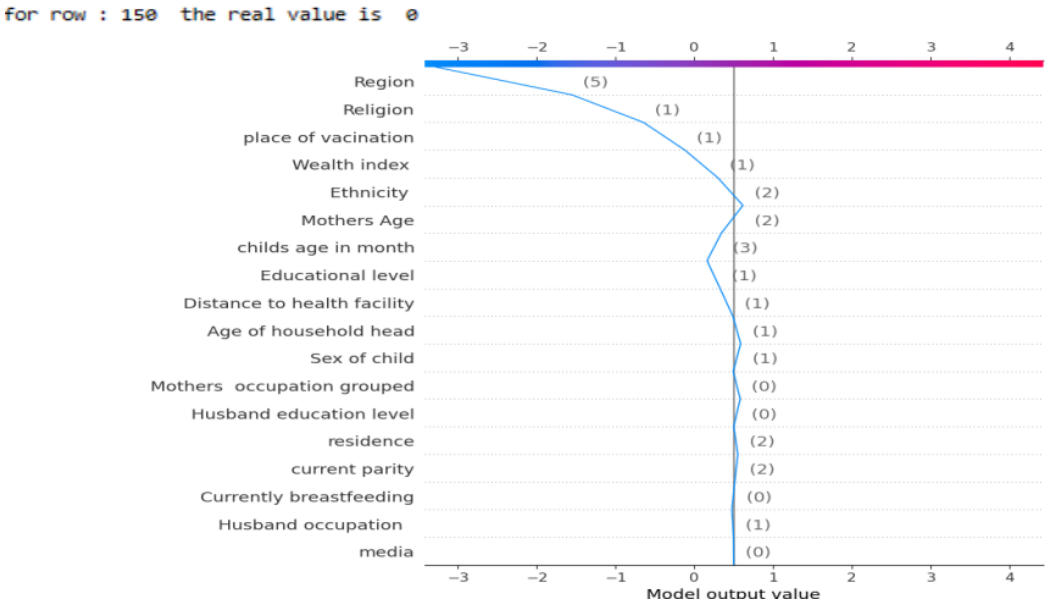

Figure 5: Sample rule predicting **Not Immunized** (Class 0).

Finally, Figure 6 presents a rule predicting the class **Fully Immunized**:

IF Child's age in months = 2 & Residence = 1 & Place of vaccination = 2 & Mother's age = 2 & Media = 1 & Religion = 0 & Educational level = 1 & Wealth index = 3 & Ethnicity = 6 & Mother's occupation = 0 & Distance to health facility = 1 & Age of household head = 3 & Sex of child = 1 & Husband's education level = 2 & Current parity = 1 & Current breastfeeding = 1 & Region = 10 & Husband's occupation = 1 **THEN** Fully Immunized.

3.4. Model Explainability

To enhance the explainability of the predictive model, we used LIME and SHAP frameworks to interpret how the model makes predictions. The XGBoost model, which performed the best,

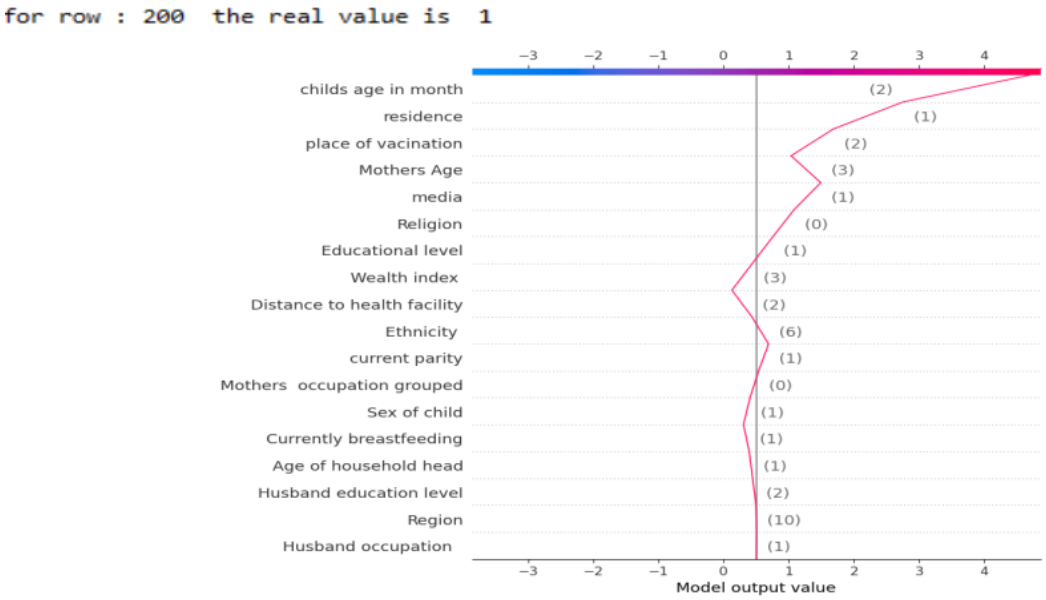

Figure 6: Sample rule predicting **Fully Immunized** (Class 1).

was analyzed using SHAP values to identify key factors influencing immunization status. Figure 7 illustrates the model's prediction of "Fully immunized" (class 1) for row 1750 with a 99% probability, based on features such as child age, vaccination place, mother's educational level, and child's sex. Figure 5 shows the model predicting "Not immunized" (class 0) for row 1200 with a 98% probability, influenced by religion, sex, and mother's education. Figure 6 presents a 71% probability for "Partially immunized" (class 2) for row 1100, based on similar features. These figures demonstrate the key factors such as age, region, and maternal characteristics that drive the model's predictions. This transparency enhances the understanding of the model's decision-making process and provides valuable insights for improving child immunization programs in Ethiopia.

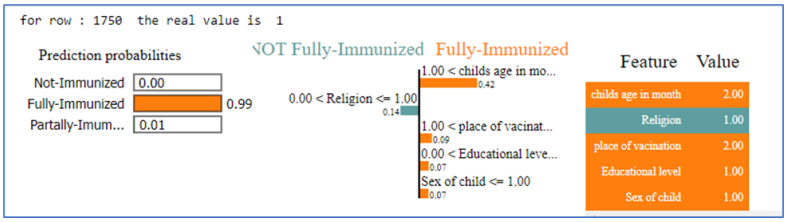

Figure 7: Explainability of the XGBoost model for row 1750

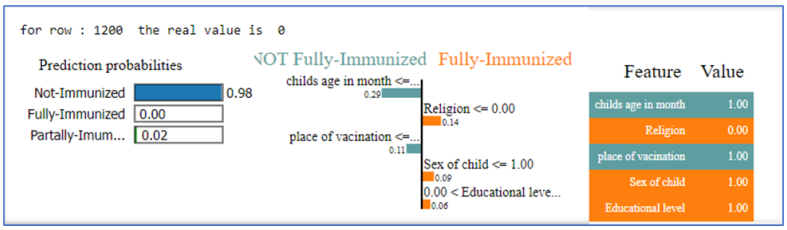

Figure 8: Explainability of the XGBoost model for row 1200

4. Conclusion

Despite efforts to ensure all children receive necessary vaccinations, there remains a significant gap in immunization coverage globally and within Ethiopia specifically. Many children,

particularly in developing regions, miss out on critical vaccinations, leading to preventable disease outbreaks and fatalities. In 2021 alone, millions of children worldwide did not complete the recommended DTP CV series, indicating a pressing need for improved access to and completion of vaccination schedules.

A recent study aimed to understand the factors influencing childhood routine immunization status in Ethiopia. By analyzing data from the Ethiopian Demographic and Health Survey (EDHS) and employing advanced machine learning techniques, researchers identified key predictors of immunization coverage. These included demographic factors such as the child's age and region, as well as socioeconomic indicators like the mother's occupation and education level.

The findings underscore the complexity of ensuring high immunization rates, highlighting the interplay between individual, community, and systemic factors. Policymakers and healthcare providers can leverage these insights to develop targeted interventions aimed at improving immunization coverage and ultimately reducing the burden of preventable diseases among children. The study's conclusions offer valuable guidance for stakeholders involved in child health initiatives.

Recognizing the multifaceted nature of immunization challenges, it becomes clear that comprehensive strategies are needed. These should encompass not just direct efforts to increase vaccine uptake but also broader social and economic interventions to support families and communities. By doing so, we can move closer to achieving universal immunization coverage and safeguarding the health and well-being of future generations.

5. Acknowledgment

We would like to acknowledge the Ethiopian Central Statistics for providing us with the data and a dataset description.

6. Funding

The research was supported by the University of Gondar Research and Community Service Vice President's Office.

7. Ethics Approval and Consent to Participate

All methods used in this study followed guidelines and regulations. The Central Gondar Zone City Administration Women, Children, and Youth Affairs Office approved this study.

8. Authors' Contributions

Muhammed conceived and designed the study, participated in data analysis, wrote the report, finished the model requirements, Alexander carried out a deep analysis of the experiment results, drafted and revised the initial manuscript, and revised the manuscript. Assefa managed the quality and progress of the whole study.

Appendix A. Appendix: Data and Methods

This section provides additional details on the data used in the study, including the source of the EDHS data, preprocessing steps, and the specific parameters used for model training. Declaration of generative AI and AI-assisted technologies in the writing process

Appendix A.1. Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used GPT 3.5 + in order to rewrite the sentence and gammer only. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1] Centers for Disease Control and Prevention (CDC), "Global immunization strategies framework 2011–2015," 2015.

[2] World Health Organization (WHO), "Immunization coverage," 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/immunization-coverage. [Accessed: 2 Dec. 2022].

[3] World Health Organization (WHO), "State of the world's vaccines and immunization," 3rd ed., Geneva, 2009.

[4] T. Gualu and A. Dilie, "Vaccination coverage and associated factors in children aged 12–23 months in Debre Markos Town, Amhara Regional State, Ethiopia," *Hindawi Advances in Public Health*, 2017.

[5] A. Girmay and A. F. Dadi, "Full immunization coverage and associated factors among children aged 12-23 months in hard-to-reach areas of Ethiopia," *Hindawi International Journal of Pediatrics*, 2019.

[6] S. Qaz and M. Usman, "Smart healthcare using data-driven prediction of immunization defaulters in expanded program on immunization (EPI)," *Computers, Materials Continua*, 2020.

[7] D. S. Weng, "Introducing machine learning for healthcare research," 2021.

[8] A. Panesar, *Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes*, Coventry, UK, 2019.

[9] A. N. Chard and M. Gacic-Dobo, "Routine vaccination coverage," US Department of Health and Human Services/Centers for Disease Control and Prevention, 2020.

[10] T. Y. Nour and A. M. Farah, "Immunization coverage in Ethiopia among 12–23 month old children: systematic review and meta-analysis," *BMC Public Health*, 2020.

[11] K. E. Al, "Full vaccination coverage among children aged 12–23 months in Ethiopia: a systematic review and meta-analysis," *BMC Public Health*, 2020.

[12] World Health Organization (WHO). (2022). *Immunization Coverage and Vaccination Trends*. Retrieved from https://www.who.int

[13] Müller, P., & Højsgaard, S. (2020). Applying Machine Learning to Predict Immunization Status: A Review of Recent Trends and Methods. *International Journal of Public Health, 65*(4), 319-328. https://doi.org/10.1007/s00038-020-01447-6

[14] Duan, Y., Zhao, J., & Yang, Z. (2021). A Comparison of Machine Learning Models in Predicting Immunization Coverage in Low-Income Countries. *Journal of Global Health, 11*(3), 1-8. https://doi.org/10.7189/jogh.11.03001

[15] Sameen, F., & Zainab, M. (2018). Predicting Immunization Adherence using Random Forest and Decision Trees. *Journal of Biomedical Informatics, 78*, 39-47. https://doi.org/10.1016/j.jbi.2018.01.003

[16] Girmay, A., & Tesfaye, S. (2019). Factors Influencing Child Immunization in Ethiopia: A Socio-Demographic and Economic Analysis. *Ethiopian Journal of Health Sciences, 29*(5), 613-623. https://doi.org/10.4314/ejhs.v29i5.4

[17] Rathore, M., & Sharma, D. (2020). Feature Selection Techniques for Predictive Modeling of Immunization Status in Healthcare Data. *Journal of Healthcare Engineering, 2020*, 1-12. https://doi.org/10.1155/2020/7159061

[18] Kibret, M., & Woldie, M. (2019). Impact of Socioeconomic and Health Service Factors on Immunization Coverage in Ethiopia: A Data Mining Approach. *PLOS ONE, 14*(11), e0225104. https://doi.org/10.1371/journal.pone.0225104

[19] Zhang, L., & Liu, Y. (2022). A Review on Machine Learning Methods for Healthcare Predictive Models: Applications in Immunization Status Prediction. *Healthcare Informatics Research, 28*(4), 218-226. https://doi.org/10.4258/hir.2022.28.4.218