Multi-Factor Health Risk Assessment for Newborns and Adolescents

Vishvapavani Bhupathi
MS in Data Science
University of Colorado
Boulder, CO
+1 720-350-1796
vibh3203@colorado.edu

Sai Aravind Nelluri
MS in Data Science
University of Colorado
Boulder, CO
+1 303-931-3515
sane9125@colorado.edu

Sai Annessha Veluri
MS in Data Science
University of Colorado
Boulder, CO
+1 512-412-9196
save5451@colorado.edu

ABSTRACT

Child or newborn health outcomes serve as a critical measure of the overall health and equity of a community. Despite progress in healthcare accessibility and awareness, disparities in child health remain a pressing concern, particularly across different social, economic and geographic groups. These challenges are not limited to individual health issues but reflect larger inequalities that affect access to care, quality of services, and community health infrastructure. Understanding the factors that contribute to poor child health outcomes is therefore essential for guiding preventive strategies and improving health standards. The core objective of this project is to support the early identification of risk-prone groups and promote targeted interventions that can reduce disparities in the health of children. By analyzing relationships among variables such as race, neighborhood and insurance coverage, the study highlights how interconnected factors collectively affect the well-being of the younger generation. The study ultimately demonstrates how data science can be utilized not only to interpret health records but also to guide real-world decisions that enhance care quality and outcomes for the most vulnerable members of society.

1. DATA ACQUISTION

The dataset we have collected is called Maternal, Child, and Adolescent Health Needs Assessment (2023-2024), which can be found on the Data.gov open data portal of the topic health (https://catalog.data.gov/dataset/maternal-child-and-adolescent-health-needs-assessment-2023-2024). This open-access data set includes aggregated health indicators of maternal, infant, and adolescent population in San Francisco and was published by the San Francisco Department of Public Health (SFDPH). It has a direct relation to the aim of our project to create data-based health risk assessment framework that covers newborns and adolescents in San Francisco because it incorporates the most important demographic, social, and clinical variables assisting in revealing the discrepancies in the health outcomes in the early years of life. Our research objectives are supported by the dataset as it will allow analyzing the relationships between demographic characteristics (race, insurance type, and geographic region) and infant health risks, as well as identify high-risk groups and determine specific public health interventions.

The dataset contains 62,714 records and 28 features, encompassing demographic, clinical, and socio-economic variables. It offers the details of health outcomes of newborns, mothers and adolescents in San Francisco, which is obtained through various public health data systems.

2. DATA PREPROCESSING

The data processing in this project progresses through a series of steps to clean, prepare, and standardize the data for analysis. First, the dataset entitled Maternal, Child, and Adolescent Health Needs Assessment (2023-2024) was initially assessed. The shape of the entire dataset is (62714, 28). Missing values in each feature were identified, and appropriate imputation methods were applied — numerical features were filled using their median values, while categorical features were filled using their mode values. As a result, the dataset is free of missing values, and ready for further processing. Table 1 shows the details of missing values before and after removal.

Table 1. Number of Missing Values Before and After Cleaning

Feature	Number of Missing Values Before Cleaning	Number of Missing Values After Cleaning
Age_group	1	0
Sex	0	0

Topic	1	0
Data_source	0	0
Health_condition-Data_source	1	0
Period	1	0
Year	30635	0
Denominator	188	0
Denominator_is	120	0
Number_with_outcome	119	0
Numerator_is	15	0
Rate_95CI	148	0
Rate_SF_pop	148	0
CI_low	209	0
CI_high	209	0
Rate_is	124	0
Insurance	1	0
Zip_code	1	0
Race_ethnicity	1	0
Rate_title	1	0
Trend_title	24	0
Filter_1000_or_more	0	0
Sort_1000_or_more	0	0
Primary_Neighborhood	10	0
Cause of death rank	62676	0
Death_tooltip	62670	0
Latest_data	62680	0

The dataset was analyzed to separate features into categorical and numerical columns based on their data types. Columns containing object data types were classified as categorical, while those with integer or float data types were classified as numerical. Since the 'Year' column represents a categorical time-based feature rather than a continuous numeric value, it was moved from the numerical to the categorical list.

Categorical features: Age_group, Sex, Topic, Data_source, Health_condition-Data_source, Period, Denominator_is, Numerator_is, Rate_95CI, Rate_is, Insurance, Zip_code, Race_ethnicity, Rate_title, Trend_title, Filter_1000_or_more, Primary_Neighborhood, Death_tooltip, Year

Numerical features: Denominator, Number_with_outcome, Rate_SF_pop, CI_low, CI_high, Sort_1000_or_more, Cause of death rank, Latest_data, Unique_row_id

For the outlier analysis, the Interquartile Range (IQR) method was selected to identify and remove extreme values from the numerical columns. Outliers are assessed using the IQR technique where data records that fall beyond 1.5 times the IQR from the quartiles, are classified as outliers. Outliers were evident in multiple variables, with the highest counts found in Number_with_outcome (8,475), Denominator (8,392), and Rate_SF_pop (6,530). Columns including Latest_data (N/A) and Unique_row_id (N/A) had no outliers. After removing extreme values, the dataset size decreased from (62,714, 28) to (43,742, 28), removing about 30.25% of the rows in the dataset.

Categorical variables were encoded using the LabelEncoder from the scikit-learn library to convert text-based categories into numeric values suitable for model training. Each categorical column was transformed individually to maintain consistent encoding. The numerical features were then standardized using the StandardScaler, ensuring that all numeric values were on a similar scale with a mean of zero and a standard deviation of one. This preprocessing step helped improve model performance and training efficiency. Sample data after encoding is shown in Figure 1.

	Age_group	Sex	Topic	Data_source	Health_condition-Data_source	Period	Year	Denominator	De
0	0	0	5	0	32	43	22	-0.198251	
1	0	0	5	0	32	47	23	-0.227846	
2	0	0	5	0	32	50	24	-0.261700	
3	0	0	5	0	32	52	25	-0.287534	
4	0	0	5	0	32	27	22	-0.314219	

Figure 1. Sample data after encoding

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while preserving most of its variance. The process began by standardizing all numerical features to ensure equal contribution from each variable. PCA was then performed, and the explained variance plot indicated that the first two principal components captured more than 95% of the total variance, suggesting that most of the dataset's information could be represented in just two dimensions. Although PCA was applied to reduce dimensionality, the reduced features were not used alone because they replaced all the original

numerical variables, which contained important information. To preserve these essential features, the first two principal components (PC1 and PC2) were added to the dataset instead of replacing it. This approach maintained the interpretability of the original variables while enriching the dataset with new features capturing underlying variance patterns. Combining both sets of features ensured that no valuable information was lost and improved the overall data representation for further analysis.

The summary statistics that followed data cleaning and standardization process indicate that all numerical variables were scaled appropriately with all features showing a mean close to 0, and standard deviation of perhaps 1, indicating proper standardization successfully completed. Beyond this, most features appeared to have mild positive skewness and kurtosis levels, such as within Denominator, Number_with_outcome, Rate_SF_pop, CI_low, CI_high (denominator is sort of extra, so decide whether you remove it or not). These statistics reflect distributions that might be slightly right-skewed with some peakedness. External to this, a few columns such as Sort_1000_or_more had higher variability representing wider ranges of values, suggesting they have distinct values existing even within the previous transformation. External to this, there are variables such as Cause of death rank and Latest_data which are perfectly constant distributions (zero variance) in any meaning and convey no extra information beyond relatively the values of the counts. In conclusion, the pre-processed data set was properly standardized where most features were normally distributed and can be used in subsequent dimensionality reduction or modeling.

As part of the data assessment process, we evaluated the dataset for missing values to identify potential data quality issues. The results showed that approximately 25.46% of the data was missing across all features, indicating a consistent pattern of missingness rather than isolated cases. Visualizations using the Missing Data Overview and Missing Data Matrix clearly depicted the extent and uniform distribution of missing entries across columns. These visual analyses helped confirm that missing data affected nearly every feature at a similar rate. Identifying this pattern was an important step before imputation, ensuring that appropriate methods (median for numerical and mode for categorical data) could be applied consistently during preprocessing.

The dataset was checked again for duplicate rows, inconsistent categorical labels, and numerical inconsistencies. The results showed that there were **no duplicate records**, confirming the dataset's integrity. Each categorical column was inspected for inconsistencies, such as variations in capitalization or spacing, and no major irregularities were observed. For numerical columns, the minimum and maximum values were analyzed to ensure that they fell within valid and expected ranges. Overall, this assessment confirmed that the dataset was clean, consistent, and ready for further modeling and analysis.

To further evaluate the dataset, a data type verification and distribution analysis were performed. The datatype check confirmed that all variables were stored as float64, ensuring consistency across the dataset. Next, the distribution of categorical variables such as Race_ethnicity, Insurance, Sex, and others was examined to identify any class imbalance. The results revealed notable disparities in class proportions—for instance, some categories like Insurance (0.0) and Death_tooltip (1.0) were highly dominant, indicating potential imbalance issues. Numerical features were summarized using descriptive statistics, which confirmed that the data was standardized with a mean near zero and a standard deviation close to one. This comprehensive assessment helped confirm that the dataset maintained structural consistency while highlighting areas where category imbalance could affect model training.

The dataset aligns closely with our problem statement on multi-factor health risk assessment for newborns and adolescents as it offers both numerical and categorical features that enable diverse analytical approaches. Numerical indicators such as Rate_SF_pop, CI_low, CI_high, and Number_with_outcome are valuable for identifying clusters of communities or demographic groups with similar health outcomes. Categorical features like Race_ethnicity, Insurance, Age_group, Sex, and Health_condition support classification models to detect patterns and disparities in newborn health risks. Additionally, outcome-based variables allow regression analysis to predict continuous health risk rates, while the inclusion of the Year feature facilitates temporal trend analysis over time. Moreover, location-based attributes such as Zip_code and Primary_Neighborhood add a layer of granularity, enabling localized insights into health disparities and community-level risk variations. However, potential

biases may exist due to uneven data collection and reporting, particularly in marginalized or underrepresented communities, where health conditions may be underreported. Regional disparities in data completeness, such as fewer records for certain zip codes or demographic groups, could reflect differences in healthcare access rather than actual health outcomes. To address these concerns, imputation was applied carefully and only where appropriate, avoiding artificial generation of demographic data. Despite these precautions, results derived from this dataset should be interpreted as indicative rather than conclusive, acknowledging the influence of social, economic, and systemic factors beyond the dataset's scope. This awareness ensures responsible use and ethical interpretation of the data in assessing newborn health risks. The sample data before and after cleaning is shown in Figure 2 and Figure 3 respectively.

Age_group	Sex	Topic	Data_source	Health_conditior	Period	Year	Denominator
0to12months	ALL	HEALTH CONDI	CDPH BIRTH RI	ASSISTED VEN	2019	2019	8307
0to12months	ALL	HEALTH CONDI	CDPH BIRTH RI	ASSISTED VEN	2020	2020	7890
0to12months	ALL	HEALTH CONDI	CDPH BIRTH RI	ASSISTED VEN	2021	2021	7413
0to12months	ALL	HEALTH CONDI	CDPH BIRTH RI	ASSISTED VEN	2022	2022	7049
0to12months	ALL	HEALTH CONDI	CDPH BIRTH RI	ASSISTED VEN	2014-2016		6673

Figure 2. Data before cleaning

Age_group	Sex	Topic	Data_source	Health_conditior	Period	Year	Denominator
0	0	5	0	32	43	22	-0.1982505417
0	0	5	0	32	47	23	-0.2278460182
0	0	5	0	32	50	24	-0.2616998367
0	0	5	0	32	52	25	-0.2875337778
0	0	5	0	32	27	22	-0.3142193874

Figure 3. Data after cleaning

3. DATA VISUALIZATION

The correlation heatmap in Figure 4 shows the relationships among numerical variables that affect newborn health outcomes. A strong positive correlation exists between Rate_SF_pop, CI_low, and CI_high. This suggests that changes in newborn health rates are consistently reflected in their confidence intervals, which helps ensure data reliability. Moderate correlations between Number_with_outcome and Sort_1000_or_more indicate that larger population groups are linked to higher recorded outcomes. On the other hand, the negative correlation between Denominator and rate-related variables means that larger population bases relate to lower standardized rates. Overall, this analysis points out important connections between variables. It helps identify redundant features and guides the choice of the most influential factors for assessing multi-factor health risks among newborns.

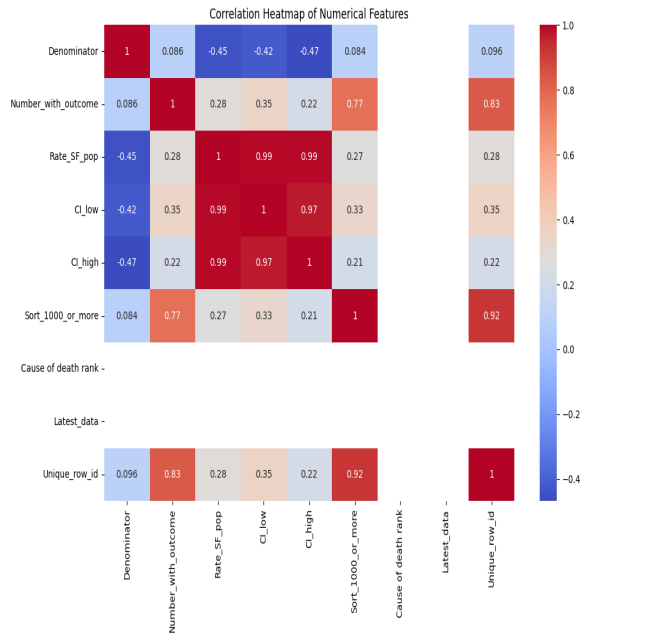

Figure 4. Correlation heatmap of numerical variables

The PCA visualizations shown in Fig. 5 and Fig. 6 illustrate how the dataset's variation is captured by two principal components. As seen in Fig. 5 the explained variance curve shows that the first two components account for nearly 100% of the total variance, indicating that most information in the original dataset is retained even after dimensionality reduction. This demonstrates that the dataset has strong internal relationships among features, allowing for effective compression without significant information loss. Fig. 5 provides a 2D projection of the data along these two principal components, where the spread of points reflects variations in health-related indicators across different records. Together, these visualizations confirm that the data structure can be effectively represented in two dimensions, simplifying analysis while maintaining the dataset's core patterns and variability. Although the reduced features were not utilized, as most numerical variables were excluded leaving primarily categorical features, the principal components PC1 and PC2 were concatenated with the dataset to enhance model learning and capture underlying data variability.

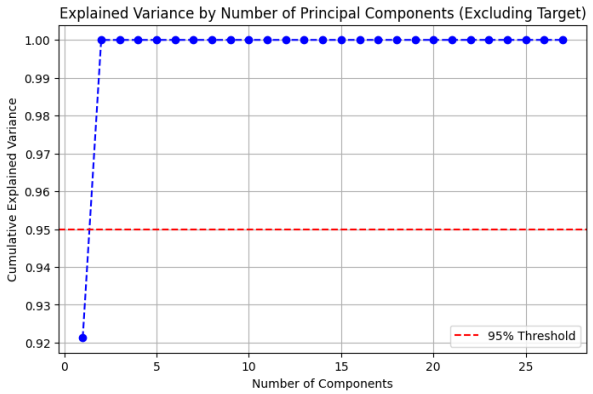

Figure 5. Explained variance of principal components

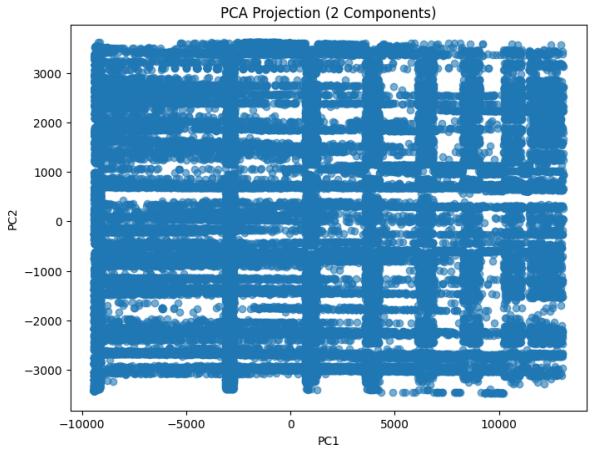

Figure 6. 2D PCA projection of dataset

The line graph in Figure 7 shows the trend of the average Rate_SF_pop over the years. It illustrates how newborn health risk rates have changed in San Francisco. The visualization indicates a fairly stable trend from 1997 to about 2020, with some moderate fluctuations. This suggests that risk levels stayed relatively consistent over time. However, after 2020, there is a sharp increase, peaking around 2024 to 2025. This points to a significant rise in newborn health risks in recent years. This sudden rise may be due to new socio-economic challenges, gaps in healthcare, or effects from the pandemic that affected the quality of care for mothers and newborns. Overall, the graph shows an important change in recent years that needs further investigation and targeted health efforts to tackle the growing risks to newborn health outcomes.

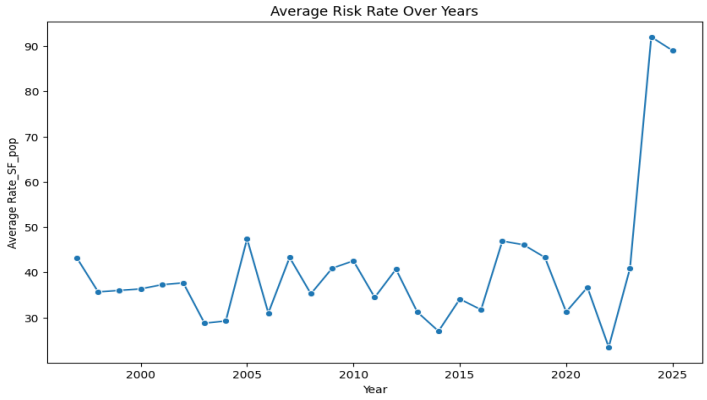

Figure 7. Trend of average newborn risk rate over the years.

The QQ plot for Rate_SF_pop in Figure 8 reveals that the data distribution deviates significantly from normality. The points curve sharply upward at the right tail, indicating strong right-skewness. This suggests that most newborn health rates in the dataset are concentrated at lower values, while a few areas or groups exhibit exceptionally high rates. Such skewness highlights disparities in health outcomes among different populations or regions, implying that a small subset of newborns may be at considerably higher risk compared to the overall population.

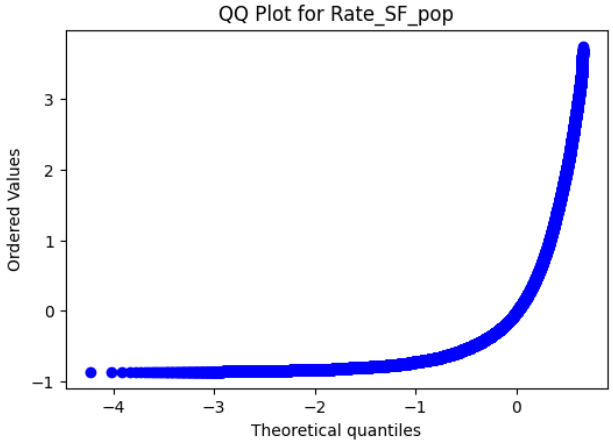

Figure 8. QQ Plot for Rate_SF_pop

The box plot in Figure 9 illustrates the distribution of the Rate_SF_pop variable across various age groups, showing substantial variation in risk levels among different populations. Newborns (0–12 months) and pregnant individuals exhibit notably higher median rates and a wider range of values, suggesting greater vulnerability in these groups. In contrast, adolescent and adult groups display more stable and lower

median rates, indicating comparatively lower associated risks. The presence of numerous outliers, especially in early life stages, highlights variability and potential data heterogeneity across demographic segments.

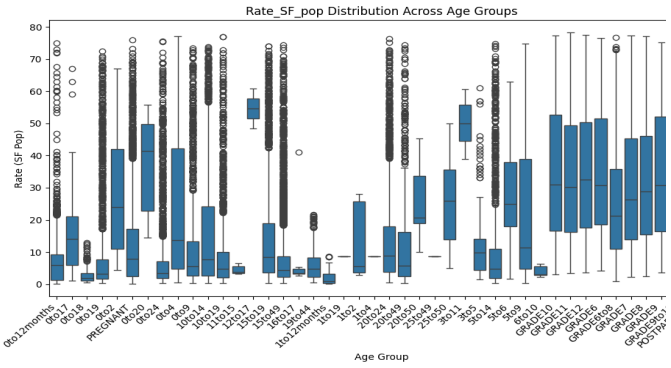

Figure 9. Distribution of Rate_SF_pop across different age groups

The violin plot in Figure 10 presents the distribution of Rate_SF_pop across different racial and ethnic groups, showing noticeable variability in risk rates. Groups such as Pacific Islander and Unknown exhibit higher median values and broader distributions, suggesting greater fluctuation or potential disparities in outcomes. In contrast, Asian, White, and Hispanic or Latino/a groups display lower and more concentrated distributions, indicating relatively consistent and lower risk levels. The presence of wider spread in certain categories highlights demographic differences that could be influenced by underlying social or healthcare factors.

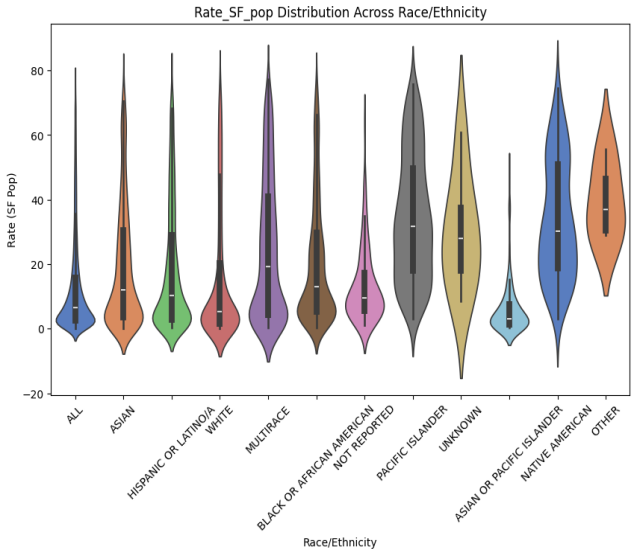

Figure 10. Distribution of Rate_SF_pop across racial and ethnic groups

This line graph in Figure 11 illustrates the yearly trend of the Rate_SF_pop across different sex categories: Male, Female, and All over time. The overall pattern shows moderate fluctuations across the years, with all three groups following a generally similar trend. Between 2000 and 2010, the rates remained relatively stable with minor variations, followed by a slight decline around 2015. After this period, the rates began to increase again, showing a noticeable spike in recent years, particularly for the "All" category. The overlapping intervals suggest that while differences exist between males and females, the overall risk trends are relatively consistent across sexes.

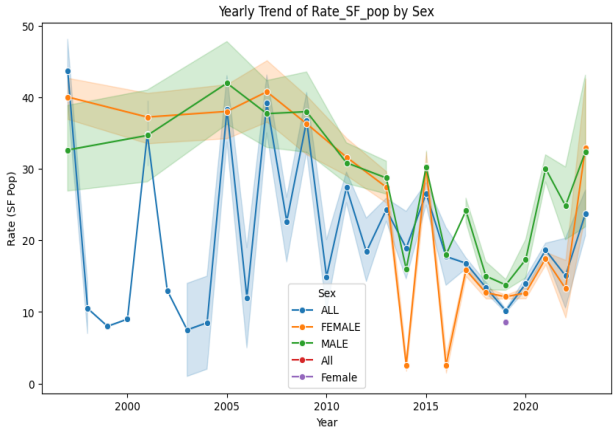

Figure 11. Distribution of Rate_SF_pop across racial and ethnic groups

This visualization in Figure 12. presents a comparative bubble plot showing the distribution of Denominator, Number with Outcome, and Rate_SF_pop across various categories. Each circle represents grouped data, with the bubble size reflecting the magnitude of values within each metric. The plot indicates significant variation among these three measures, with the Denominator showing the largest scale differences, likely representing population size or sample base. The Number with Outcome values remain moderate, while Rate_SF_pop displays lower relative values but consistent clustering patterns. This visualization highlights the proportional relationships among population size, recorded outcomes, and associated rates, providing a clear overview of data magnitude and balance across metrics.

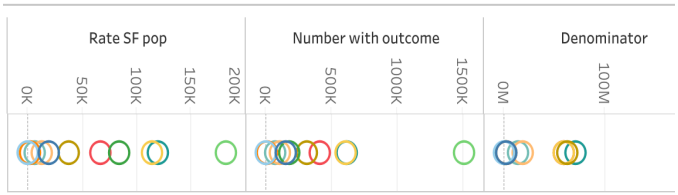

Figure 12. Bubble plot illustrating the comparative distribution of Denominator, Number with Outcome, and Rate_SF_pop

This bar chart in Figure 13 llustrates the average Rate_SF_pop across different insurance types. It shows that individuals covered under SFHP Head Start or Early Head Start programs have the highest average risk rate, followed by ALL and Public insurance categories. In contrast, Private and Not Public insurance types exhibit significantly lower rates. This suggests that newborns associated with public or community-based insurance programs may experience higher health-related risks compared to those under private coverage.

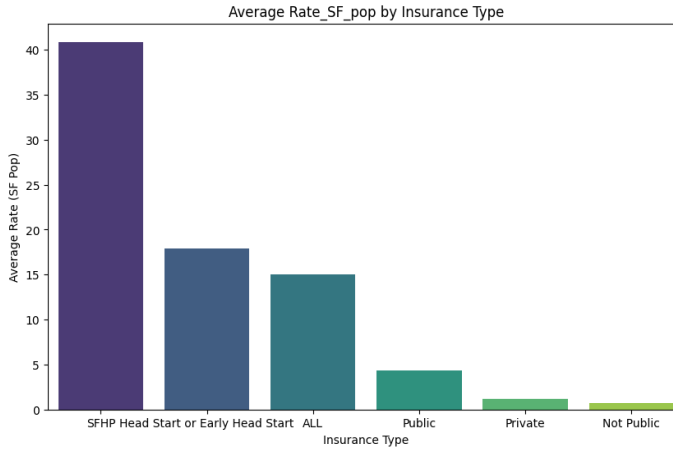

Figure 13. Average Rate_SF_pop across different insurance type

4. MODEL TRAINING

The modeling framework for this study involved both supervised and unsupervised approaches to assess the health risks among newborns and adolescents. Initially, unsupervised clustering algorithms were employed to derive risk categories from unlabeled data for subsequent classification tasks. Following this, supervised classification models were trained to predict risk levels using the generated labels, while regression models were developed to estimate continuous health outcomes such as Rate_SF_pop. Model selection, preprocessing strategies, and hyperparameter optimization were carefully aligned with algorithm-specific requirements to ensure performance across all predictive tasks.

4.1 Clustering for Risk Label Generation

Unlabeled health data for newborns and adolescents was initially processed using unsupervised clustering algorithms to define risk categories. Two clustering approaches were applied: K-Means and Gaussian Mixture Models (GMM). K-Means clustering was configured with three clusters (n_clusters=3) and a fixed random seed to ensure reproducibility. Cluster centroids were analyzed for the feature Rate_SF_pop and assigned as Low Risk (0), Medium Risk (1), and High Risk (2) based on ascending order of mean values. K-Means was selected due to its efficiency in clustering numerical features and its suitability for large datasets with relatively well-separated groups.

Gaussian Mixture Models were also applied with three components (n_components=3) and a full covariance structure. Cluster labels were subsequently mapped to risk categories using the mean of Rate_SF_pop for each component. Evaluation of clustering results using Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score suggested that K-Means produced slightly better cluster separation, and its labels were adopted for subsequent supervised learning. GMM was chosen for their flexibility in capturing clusters with different shapes and variances.

4.2 Supervised Classification Models

Classification models were trained to predict risk categories derived from clustering. The dataset was split into training and testing sets with an 80:20 ratio, maintaining stratification to preserve class distribution. Classification models did not need to go through any data transformation as the processed data was compatible with all the models. Hyperparameter optimization was performed using GridSearchCV with three-fold cross-validation. Key hyperparameters explored are included in Table 2.

Table 2. Hyperparameters for Classification

Model	Hyperparameters
Logistic	C: [0.01, 0.1, 1, 10], penalty: ['l2'],

Regression	solver: ['lbfgs', 'liblinear']
Decision Tree	max_depth: [3,5,10,20,None], min_samples_split: [2,5,10], min_samples_leaf: [1,2,4], max_features: ['sqrt','log2',None], max_leaf_nodes: [None,10,20,50]
Random Forest	n_estimators: [100,200,300], max_depth: [5,10,20,None], min_samples_leaf: [1,2,4]
Support Vector Machine	C: [0.1,1,10]
K-Nearest Neighbors	n_neighbors: [3,5,7,11], weights: ['uniform','distance']
XGBoost Classifier	n_estimators: [200,300,500], max_depth: [3,5,7]
MLP Neural Network	hidden_layer_sizes: [(50,),(100,),(100,50)], activation: ['relu','tanh']

Logistic Regression was chosen due to its efficiency in modeling linear relationships between features and the target, providing a strong baseline with interpretable coefficients for risk prediction. Decision Tree was included for its ability to capture nonlinear patterns and interactions between features, enabling intuitive visualization of decision paths for different risk categories. Random Forest was selected to improve stability and generalization by aggregating multiple decision trees, reducing overfitting while maintaining high predictive accuracy. Support Vector Machine was employed to maximize the margin between classes, which is useful for distinguishing subtle differences among risk levels in the dataset. K-Nearest Neighbors was applied to leverage instance-based learning, allowing classification based on similarity to known data points, useful for capturing local patterns in the feature space. XGBoost was incorporated for its gradient boosting capability, combining weak learners to produce a strong model, particularly effective in handling complex feature interactions with high predictive performance. MLP Neural Network was included to model complex nonlinear relationships through multiple layers and neurons, capturing intricate patterns that may exist in the health risk data and providing robust predictions across classes.

Each model was trained using a pipeline combining preprocessing and classifier steps. Model performance was evaluated using accuracy, precision, recall, and F1-score on both training and test sets. Confusion matrices were visualized for all models to assess class-wise prediction performance.

4.3 Regression Models for Numeric Health Outcome Prediction

Regression models were trained to predict the continuous feature Rate_SF_pop. The dataset was divided into training and test sets with an 80:20 split. Numerical features were standardized, and categorical features were one-hot encoded. A ColumnTransformer was used to apply these transformations consistently across all models. The data before and after transformation is shown in Figure 14 and 15.

	Age_group	Sex	Topic	Data_source	Health_condition-Data_source	Period	Year	Denominator
0	0	0	5	0	32	43	22	-0.198251
1	0	0	5	0	32	47	23	-0.227846
2	0	0	5	0	32	50	24	-0.261700
3	0	0	5	0	32	52	25	-0.287534
4	0	0	5	0	32	27	22	-0.314219

Fig 14. Data before transformation

	0	1	2	3	4	5	6	7
0	-1.495695	-0.944862	-0.612787	-2.617098	-1.110875	0.381436	0.255485	-0.201022
1	-1.495695	-0.944862	-0.612787	-2.617098	-1.110875	0.841662	0.618848	-0.230424
2	-1.495695	-0.944862	-0.612787	-2.617098	-1.110875	1.186832	0.982211	-0.264057
3	-1.495695	-0.944862	-0.612787	-2.617098	-1.110875	1.416945	1.345575	-0.289722
4	-1.495695	-0.944862	-0.612787	-2.617098	-1.110875	-1.459468	0.255485	-0.316233

Fig 15. Data after transformation

The regression algorithms included: Linear Regression, Ridge Regression, Lasso Regression, Decision Tree Regressor, Random Forest Regressor, Support Vector Regressor, K-Nearest Neighbors Regressor, XGBoost Regressor and MLP Neural Network Regressor.

Hyperparameter tuning was performed using GridSearchCV, optimizing model-specific parameters. Key hyperparameters included are in Table 3:

Table 3. Hyperparameters for Regression

Model	Hyperparameters
Linear Regression	None
Ridge Regression	alpha: [0.1, 1, 10]
Lasso	alpha: [0.0001, 0.001, 0.01]

Regression	
Decision Tree Regressor	max_depth: [3,5,10,None], min_samples_split: [2,5,10], min_samples_leaf: [1,2,4], max_leaf_nodes: [None,10,20,50]
Random Forest Regressor	n_estimators: [100,200,300], max_depth: [5,10,20,None]
Support Vector Regressor	C: [0.1,1,10]
K-Nearest Neighbors Regressor	n_neighbors: [3,5,7,9], weights: ['uniform','distance']
XGBoost Regressor	n_estimators: [200,300], max_depth: [3,5,7], learning_rate: [0.01,0.05,0.1]
MLP Neural Network Regressor	hidden_layer_sizes: [(100,),(100,50)], activation: ['relu','tanh']

Linear Regression was used as a basic model to understand how features relate to the target and to provide an easy-to-interpret baseline. Ridge Regression was applied to reduce overfitting when features are closely related, helping the model stay stable. Lasso Regression helped identify the most important features by reducing the impact of less relevant ones. Decision Tree Regressor captured complex patterns and interactions between features, showing clear decision rules. Random Forest Regressor combined many trees to improve accuracy and handle noisy data. Support Vector Regressor focused on separating data points effectively in high-dimensional space. K-Nearest Neighbors Regressor predicted values based on similarity to nearby points, capturing local patterns. XGBoost Regressor built strong predictions by combining many weak models, handling complex relationships efficiently. MLP Neural Network Regressor learned deep nonlinear patterns in the data to improve overall prediction performance.

Each regression model was trained using a pipeline combining preprocessing and estimator, with performance evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 on training and testing sets. Predictions were visualized using scatter plots comparing true versus predicted values. Transformed datasets were saved for all models to facilitate reproducibility and further analysis.

5. RESULTS

Unsupervised clustering was applied to the unlabeled health dataset to derive risk categories. Two clustering algorithms, K-Means and Gaussian Mixture Models (GMM), were evaluated using common cluster validation metrics. The evaluation results are presented in Table 4.

Table 3. Results of Clustering Algorithms

Algorithm	Silhouette Score	Davies–Bouldin Index	Calinski–Harabasz Score
K-Means	0.4880	0.6924	102,783.70
GMM	0.4875	0.6936	102,165.04

The Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Score indicate that both algorithms generated reasonable cluster separation, with K-Means slightly outperforming GMM across all metrics. Specifically, K-Means achieved a higher Silhouette Score and Calinski–Harabasz Score, and a lower Davies–Bouldin Index, suggesting more compact and well-separated clusters. Based on these results, K-Means labels were selected as the definitive risk categories for subsequent supervised classification tasks.

Supervised classification models were trained on the labeled dataset derived from K-Means clustering to predict the health risk categories of newborns and adolescents. The performance of seven classification algorithms was evaluated using standard metrics, including Accuracy, Precision, Recall, and F1-Score on both training and test sets. The overall results are summarized in Table 4.

The results indicate excellent overall predictive performance across all models, with most achieving near-perfect training and test metrics. However, slight variations in test accuracy highlight potential overfitting in some models. For instance, K-Nearest Neighbors demonstrated perfect training performance 100% but a comparatively lower test accuracy 92.93%, suggesting overfitting. Support Vector Machine and MLP models exhibited high performance with minor drops between training and testing, indicating good generalization.

Table 4. Results of Classification Algorithms

Model	Train Accuracy (%)	Test Accuracy (%)	Train F1 (%)	Test F1 (%)
Logistic Regression	99.79	99.83	99.79	99.83
Decision Tree	100.00	99.93	100.00	99.93
Random Forest	100.00	99.97	100.00	99.97
Support Vector Machine	99.76	99.22	99.76	99.22
K-Nearest Neighbors	100.00	92.94	100.00	92.93
XGBoost	100.00	99.90	100.00	99.90
MLP	99.82	99.62	99.82	99.62

Random Forest emerged as the best-performing model, achieving a train accuracy of 100% and a test accuracy of 99.97%, with perfect precision, recall, and F1-scores for all risk categories. The optimized hyperparameters for Random Forest were a maximum depth of 20, 1 minimum sample per leaf, and 300 estimators, ensuring both complexity and stability. The confusion matrix for Random Forest (Figure 16) confirms that all samples were correctly classified, demonstrating exceptional predictive reliability.

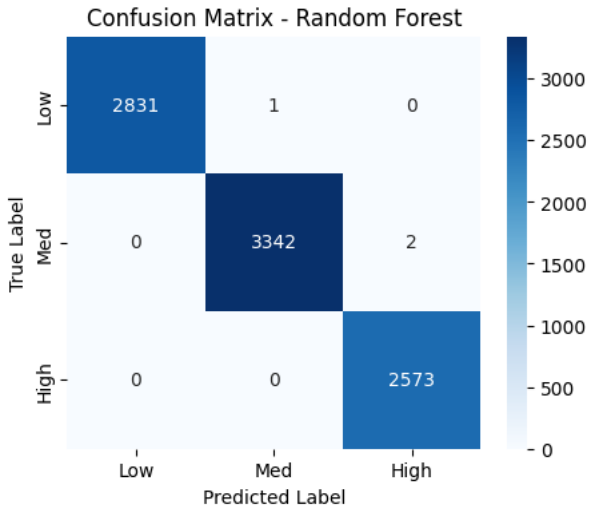

Fig 16. Confusion Matrix for Random Forest

The regression analysis aimed to predict the continuous risk score, Rate_SF_pop, for newborns and adolescents using a suite of nine models. Model performance was evaluated using standard metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score for both training and test sets as shown in Table 5.

Model	Train MAE	Test MAE	Train RMSE	Test RMSE	Train R^2	Test R^2
Linear Regression	0.01	0.01	0.03	0.06	1.00	1.00
Ridge Regression	0.01	0.01	0.03	0.06	1.00	1.00
Lasso Regression	0.01	0.01	0.03	0.06	1.00	1.00
Decision Tree	0.00	0.00	0.00	0.03	1.00	1.00
Random Forest	0.00	0.00	0.01	0.02	1.00	1.00
Support Vector Regressor	0.04	0.04	0.05	0.06	1.00	1.00
K-Nearest Neighbors	0.00	0.07	0.00	0.12	1.00	0.98
XGBoost Regressor	0.00	0.01	0.01	0.02	1.00	1.00
MLP Neural Network	0.02	0.02	0.03	0.04	1.00	1.00

Overall, all models demonstrated high predictive capability, with R^2 scores exceeding 0.996 on the test set. Linear, Ridge, and Lasso regressions performed robustly but exhibited slightly higher test MAE and RMSE compared to tree-based and ensemble models. Decision Tree and Random Forest models achieved near-perfect training R^2, suggesting very low training error; however, the slight discrepancy between train and test RMSE in Decision Tree indicates potential minor overfitting. K-Nearest Neighbors exhibited high training accuracy but poorer generalization on the test set, indicating overfitting to training data.

Among all models, XGBoost Regressor emerged as the best performer, balancing high predictive accuracy

with low generalization error. The optimized hyperparameters for XGBoost were: learning_rate = 0.1, max_depth = 5, n_estimators = 300, reflecting moderate model complexity with controlled overfitting. This model achieved Test R^2 = 0.99974, Test MAE = 0.00521, and Test RMSE = 0.01620, demonstrating robust performance in predicting the multi-factor risk scores across the dataset. The prediction plot in Figure 17 shows that the XGBoost regressor's predictions closely follow the true values along the diagonal, indicating strong predictive accuracy with minimal deviation. Only a few points slightly deviate, suggesting minor prediction errors.

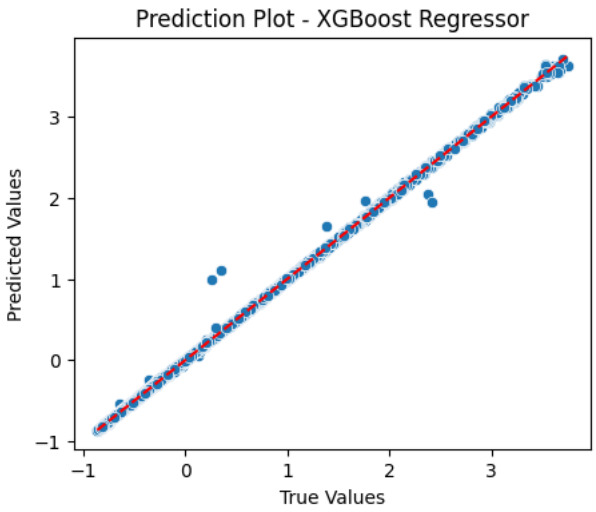

Fig 17. Prediction Plot for XGBoost Regressor