CSE 587 – Data Intensive Computing

Project Report

Section C - Fall 2024

Project Title: Leveraging Health Data to Predict Infant Survival and Wellbeing from Pregnancy to Early Childhood

Team Members:

Team Members	UB Number	Email ID
Anchal Daga	50609480	anchalda@buffalo.edu
Keerthana Vangala	50604773	kvangala@buffalo.edu
Grace Evangelene Avula Lael	50595809	graceeva@buffalo.edu
Sharanya Nallapeddi	50593866	snallape@buffalo.edu

Table of Contents

I. PROBLEM STATEMENT:

The most important public health problem of global importance is the protection and well-being of infants from pregnancy into early childhood. Maternal health status, prenatal and postnatal healthcare access, immunization cover, nutrition programs, socioeconomic gaps are some of the complicated, interwoven issues contributing to persisting mother and infant outcomes gaps despite advancement on many health fronts. Low birth weight, for example, high teen pregnancy rates and a lack of medical resources in rural locations lead to a variety of avoidable problems at labour and early development.

Health systems struggle to determine high-risk locations and effectively allocate resources to address these issues. The lack of reliable methods to evaluate and comprehend the enormous volumes of health data available exacerbates this. This leads to vulnerable people being at risk since chances to address avoidable causes of infant death and poor health outcomes are lost.

The aim is to develop predictive models that identify major factors affecting the survival and health of an infant using health data from acknowledged, reliable sources such as https://databank.worldbank.org/. Actionable results on targeted intervention by politicians and healthcare professionals due to these data analysis and modelling will be enabled through this project. Predictive tools might be used, to guide resource allocation, prioritize immunization efforts, and identify areas where improved healthcare facilities are needed. In the end, this approach will reinforce data-driven strategies toward better outcomes for newborns and their families in many places.

II. ABSTRACT:

This project is based on the use of health data to understand and predict the factors that have an influence on infant survival and well-being from pregnancy into early childhood. By analysing variables such as maternal age, access to prenatal care, vaccination rates, low birth weights, and teen pregnancies, patterns may be determined that affect health outcomes. Using valid data from sources such as the CDC, the project hopes to create predictive models that can determine hotspots. These insights will aid healthcare providers and policymakers in resource distribution, such as

vaccines and maternal support services, to help reduce preventable infant deaths and improve overall care in vulnerable communities.

The aim of this initiative is to develop a framework that will support equitable health care through early risk identification and efficient distribution of available resources. The project will target the need for ensuring that critical interventions-such as vaccination drives or prenatal programs-reach only those areas that are in need. This data-driven approach empowers healthcare systems to make informed decisions and contributes to reducing disparities, enhancing maternal and infant health, and building stronger, healthier communities.

III. INTRODUCTION:

- Data Acquisition and Preprocessing with Exploratory Analysis:

Every successful data-driven project starts with doing the homework, which is by gathering the right data, refining it to remove inconsistencies, and exploring it in depth to find meaningful patterns. In our study, we extracted extensive maternal health data from the https://databank.worldbank.org/. website, a rich repository of information. We filtered and combined only the relevant pieces to get a dataset that would closely match our research objectives.

Data cleaning entailed a lot: elimination of duplicate records, exclusions based on data irrelevant to our studies, and inconsistency such as negative values, among others, in relation to entries that had passed the due dates. In some instances, data was rearranged, categories standardized, and sometimes new features created, improving both the richness and relevance of this data set. This therefore clears up, condenses, and prepares the data for analysis in the most appropriate manner.

EDA played an important role in the understanding of the data. Using visualizations and statistical summaries, we showed trends, relationships, and some anomalies that could exist within the data. This was an important step for determining which variables were most influential, identifying any anomalies, and building the remaining steps of our project.

By the end of this phase, we had raw data that was transformed into a structured and insight-oriented foundation, thus leaving all the ground for precise and impact-heavy analysis in the further phases of our research.

- Machine Learning and Statistical Analysis:

  This phase is all about finding complicated patterns and making predictions based on machine learning and statistical approaches. Statistical modelling provides a clear procedure to analyse data, come up with relationships, and forecast results of an event. Basic to advanced methods, including linear regression, logistic regression, can get useful insights from the given data and estimate uncertainty, with great accuracy, out of raw numbers and convert them into something resourceful.

  Statistical models, in the context of maternal and infant health, point to those factors that influence survival and well-being. These models reveal the trends in aspects like maternal nutrition, access to healthcare, and socioeconomic conditions, hence allowing us to predict and better understand the outcomes. For instance, a classification model might estimate the survival chances of an infant based on access to healthcare, whereas regression models could highlight the effect of prenatal care on birth weight. These findings would support healthcare professionals in developing targeted strategies and thus contribute to better early childhood health and more informed public health initiatives.

- Building a data product:

  The data product of this project aims to develop an interactive platform that will enable the analysis and derivation of conclusions from data related to the health status of mothers and babies. This will be a platform that, by combining a strong tech stack comprising Streamlit for user experience, MySQL for data administration, and Python libraries for analytics, will be enabled to create, read, edit, and remove data; plot the data dynamically; and show patterns in critical health indicators. This phase, with an emphasis on accessibility and actionable information, arms healthcare professionals and policymakers with the tools they need to address inequities and promote well-informed decision-making.

IV. DATA PREPARATION AND EXPLORATORY ANALYSIS:

- Data collection:

  Any data analysis work shall begin with data gathering, which should be relevant and reliable. The website https://databank.worldbank.org/ has provided a very vast library of maternal health data across numerous demographic, socio-economic, and healthcare-related characteristics; hence, this study utilizes an extensive dataset derived from this website. Due to its large size, we began carefully looking through what was within the dataset to see which were best suited for our current project. This meant excluding all superfluous extra points and focusing on characteristics, such as maternal health, infant survival, and availability of medical resources that would contribute significantly to our analysis.

  After identifying the pertinent data, we downloaded it and included it into our workspace, making sure it was properly structured and ready for further processing. Thus, a clean and targeted dataset was created that would serve as the foundation for the ensuing cleaning, analysis, and modelling stages. We ensured that our study was informed by the best available data to answer the specific research questions through the investment of time in rigorous data filtering and integration.

- Data Pre-Processing:

  Steps in Data Pre-processing:

  1. Locating the Source of Data: We first searched for a reliable website that would provide us with adequate data regarding maternal and infant health factors. We chose the data bank of World Bank website. Since this was a very informative source, giving a wide variety of data, we used it as the basis for our study.
  2. Choosing Appropriate Information: Since the dataset was huge, we did a critical clean-up to see what exactly would be best for our project. Among the maternal health indicators that were key factors, infant survival rates, socio-economic influences, and health access, we set aside data unrelated to our research.

3. Downloading the Data: We identified the relevant sections and downloaded the dataset in a format that was easily processable and analysable, so that it would be ready for further refinement.
4. Removal of Irrelevant Details: Data was cleaned by erasing data that was not serving the research purpose. This included eliminating the regions or categories that are not under study, outdated data, or those irrelevant factors that one wanted to study.
5. Refining of the Dataset: Third and last, refinement of the dataset was to make it practical; therefore, structuring clearly without any missing or inconsistent values and everything in such a format that makes all the next processes meaningful.

By breaking it down into these steps, we took an overwhelming amount of raw data and turned it into a very organized and focused dataset that was perfectly aligned with the objectives of our study.

- Data Cleaning:

For the dataset to be valid, relevant, and ready for analysis, a proper cleaning process needed to be in place. This stage is purposed to fine-tune the data set by dealing with inconsistencies and removing irrelevant material for higher quality to elicit useful information.

1. Removal of Duplicate Entries: Duplicate records were removed to avoid biased analyses and to have the accurate results. For example, repeated entries of the same region might give overweight to certain data. We applied drop_duplicates () for this purpose.
2. Removing Irrelevant Columns: This helped to reduce unnecessary data processing. Columns like "Year Code" that added little value to our study were deleted. We have used the drop () function with the objective of retaining only the most relevant features to our aims.
3. Row Filtering: Rows containing global data, high-income aggregates, or blank entries were excluded as they were unnecessary. This step helped in providing specific and actionable insights in the dataset.

4. Handling Negative Values: The negative values that do not make any sense in this context were identified and removed. Only positive values were kept for numeric fields by using filtering conditions in pandas.
5. Categorical Data Conversion: Categorical variables, such as region names, were converted into numerical labels to increase their compatibility with machine learning models without losing the integrity of the original information.
6. Transforming Year Data: The "Year" column was converted into a datetime format using pd.to_datetime(), which allows for more advanced filtering and time-based analyses.
7. Renaming Columns: Ambiguous column names, such as [SP.DYN.CBRT.IN], were replaced with clear and descriptive labels. Spaces in names were replaced with underscores to maintain consistency and ease of use.
8. Feature Engineering: New features were created to enrich the dataset. Examples include:
    a. Health Infrastructure Index: A composite measure of hospital beds, nurses, and physicians.
    b. Immunization Coverage Score: Aggregated from key vaccination percentages.
    c. Maternal and Infant Health Risk Index: Combined mortality rates.
    d. Maternal Healthcare Access Index: Measured access to prenatal and postnatal services.
    e. Birth-Death Ratio: Studied demographic trends.
    f. Life Expectancy Variability: Highlighted gender disparities in life expectancy.
9. Outlier Management: Outliers were then detected using the IQR method, and the median replacement and removal of the values were done to maintain the results unbiased and robust.

10. Rounding Numeric Values: Rounding the numeric fields up to two decimal places has improved readability without compromising any precision and consistency.

These steps combined cleaned a raw dataset into a refined, structured base for analysis, supporting reliable insights and actionable outcomes.

- Exploratory Data Analysis:

EDA is one of the crucial steps in data analysis. The nature of the data is analysed by looking for patterns, connections, and insights of importance from the data sets. This method helps the researchers to identify anomalies or outliers, deepen the understanding of the structure of data, and identify which variables are relevant for further study. Decision on data preparation, hypothesis testing, and the selection of suitable techniques of analysis will greatly rely on this step.

EDA makes use of statistical summaries, visual aids, and comparing methods to deeply investigate the data. In this way, significant patterns and links in the data can also be revealed, and issues that need attention can be illustrated, for instance, holes in numbers or discrepancies. By forming a clearer vision of the dataset and its potential, therefore, EDA lays a good foundation for the exploitation of data in problem solving.

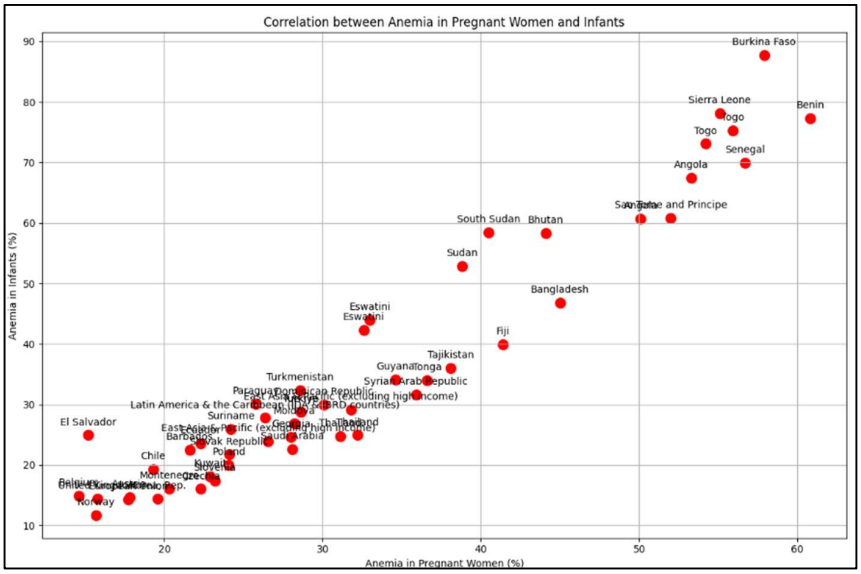
- EDA by Anchal Daga – 50609480:

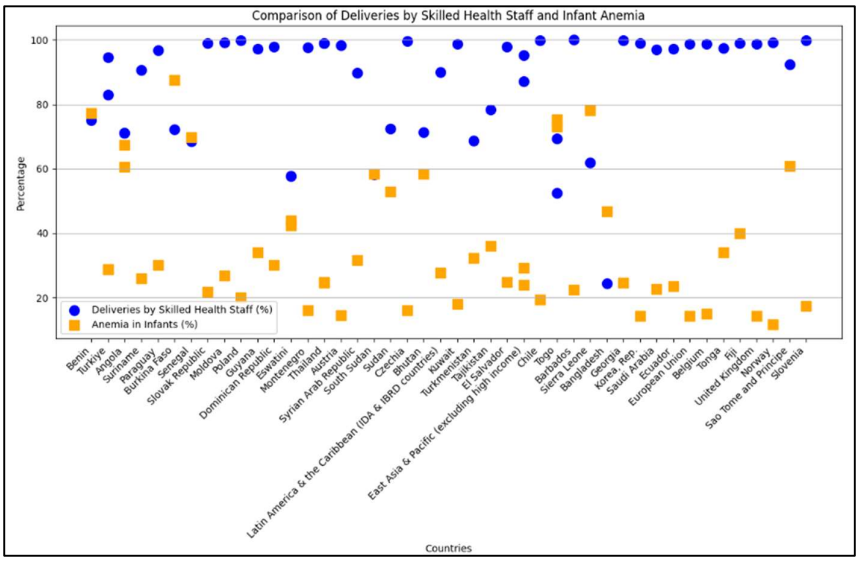
1. Question 1: Does the incidence of anaemia in pregnant women connect with the prevalence of anaemia in newborns, and how is this relationship impacted by maternal nutrition programs or availability to qualified medical personnel?
    - Hypothesis 1: A higher prevalence of anaemia among pregnant women will correlate with an increased prevalence of anaemia in infants.
    - Hypothesis 2: In a region, those with greater percentages of deliveries attended by experienced health workers would display lower rates of anaemia in infants.

- Implementation of Hypothesis 1:

  The relationship of anaemia in pregnant women and anaemia in infants was visualized from the dataset by selecting a sample of 50 countries. A scatter plot was plotted to plot the prevalence of anaemia among pregnant women against prevalence of anaemia in infants. Country names were added as annotations on the data points for context. This visualization allowed the testing of whether a higher prevalence of anaemia in pregnant women correlates with a higher prevalence in infants, and any relationship is far easier to visually inspect.

- Visualizations:





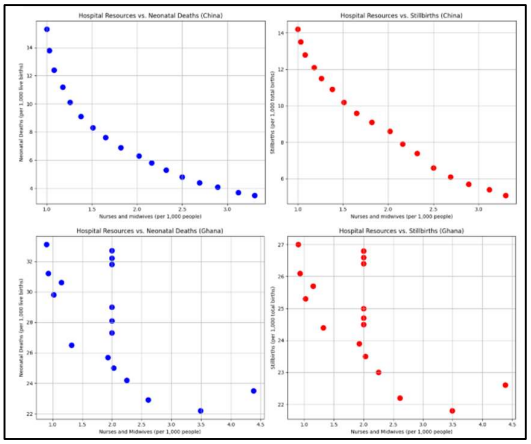
- Observation: The visualizations indicate a clear alignment, hence there might be a correlation between the prevalence of anaemia in pregnant women and infants across the selected countries.

2. Question 2: What is the relationship between hospital infrastructure and the rates of stillbirths and neonatal deaths in different nations?

   - Hypothesis 1: In China, higher hospital resources will be associated with lower rates of neonatal deaths and stillbirths.
   - Hypothesis 2: In Ghana, higher hospital resources will be associated with lower rates of neonatal deaths and stillbirths.
   - Implementation of Hypothesis 1:

     Data for this analysis were filtered from the dataset for China in testing the hypothesis that higher hospital resources in the country result in lower rates of neonatal deaths and stillbirths. Specific focus was placed on the number of nurses and midwives per 1,000 people because of their potential to impact neonatal and stillbirth rates. First, scatter plots were made for the relationships between numbers of nurses and midwives versus neonatal death rates and stillbirth rates, respectively. One plotted the associations of neonatal deaths with hospital resources, while a second showed the ones with stillbirths. By analysing these visualizations, one would be able to see if increasing the number of healthcare professionals per 1,000 people had been associated with lower rates of neonatal and stillbirths, thereby supporting this hypothesis.

   - Visualizations:



- Observation: These visualizations show the possibility of an inverse correlation of neonatal death and stillbirth rates with the level of hospital resources across China, hence confirming that high availability of healthcare professionals can potentially improve birth outcomes.

- EDA by Keerthana Vangala – 50604773:

1. Question 1: What is the association between immunization coverage (HepB, Polio, Measles) and infant mortality?

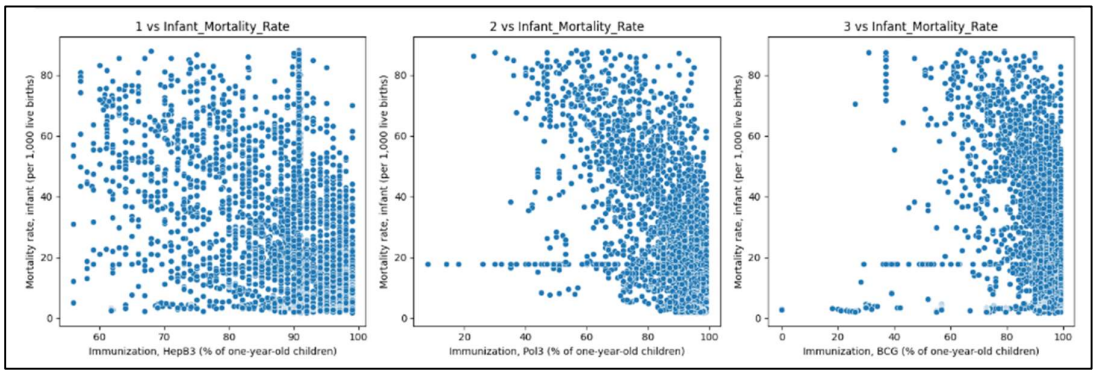
   - Hypothesis 1: Higher immunization rates for HepB, polio, and measles are linked with lower infant mortality rates. If this is correct it means that vaccination given in the 1st one year to a baby are protecting them against the respective diseases.

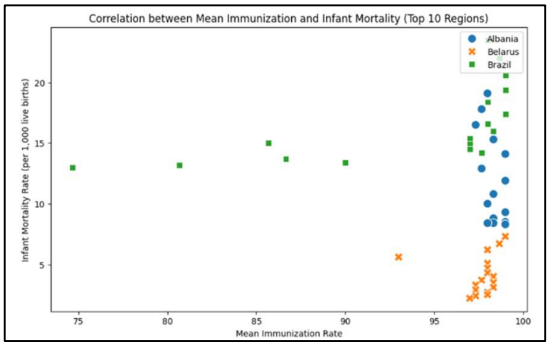
   - Hypothesis 2: Regions with higher immunization records will tend to have lower infant mortality rate.

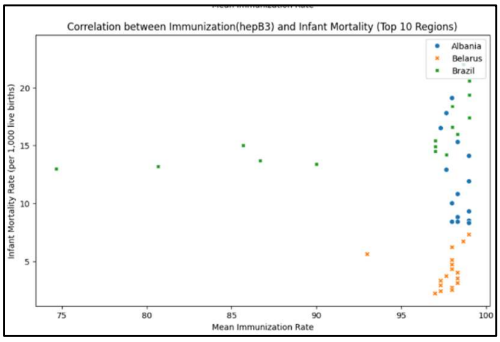
   - Implementation of Hypothesis 1:

     A scatter plot was used to visualize the relationship between immunization rates HepB, Polio, and Measles, and infant mortality for a sample of 50 countries. The immunization rates of each vaccine HepB, Polio, and Measles were plotted against the infant mortality rate on the y-axis. Country names were added as annotations on the data points to give context to the countries that represented the data points. This visualization allowed for a test of whether higher vaccination rates are associated with lower infant mortality rates, enabling an easier visual inspection of this relationship.

   - Visualizations:



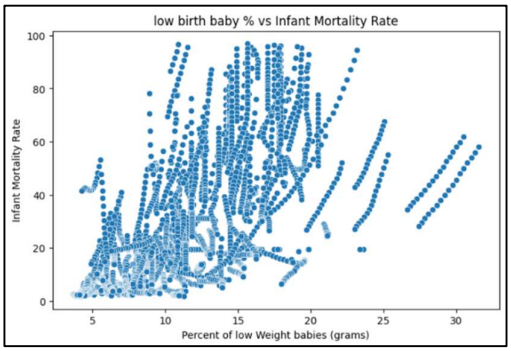


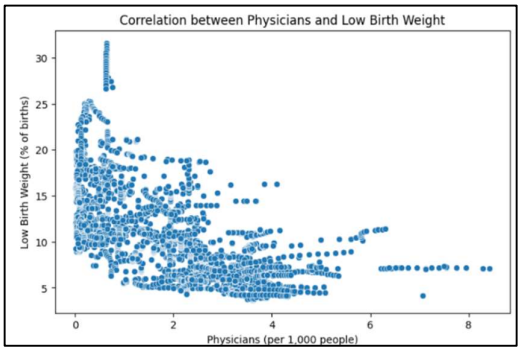


- Observation: In these visualizations, the hypothesis that associates higher rates of immunization with lower infant mortality rates across selected regions is very strongly supported.

2. Question 2. How does low birth weight correlate outcomes in infants' health and infant mortality?
   - Hypothesis 1: Compared to normal birth weight, low birth weight is linked to a higher risk of infant mortality.
   - Hypothesis 2: Is there any correlation between babies with lower weight and attendants taking care of the baby?
   - Implementation of Hypothesis 1:

     A scatter plot showing the percentage of low-birth-weight babies versus the infant mortality rate was constructed. This is because the study will try to ascertain whether low birth weight is associated with increased risks of infant mortality. For Hypothesis 2, an investigation was conducted on the relationship between the number of physicians per 1,000 people and the percent of low-birth-weight babies to examine if a higher availability of medical care is associated with lower low birth weight rates. The scatter plots for both hypotheses are included, showing the above relationships and enabling a clear evaluation of possible links between birth weight, medical care, and infant mortality.





	Physicians (per 1,000 people)	Low-birthweight babies (% of births)
Physicians (per 1,000 people)	1.000000	-0.631968
Low-birthweight babies (% of births)	-0.631968	1.000000

- Observation: The visualizations support the hypotheses by clearly showing a positive correlation between low birth weight and higher infant mortality, and a potential negative correlation between the number of physicians and the prevalence of low birth weight.

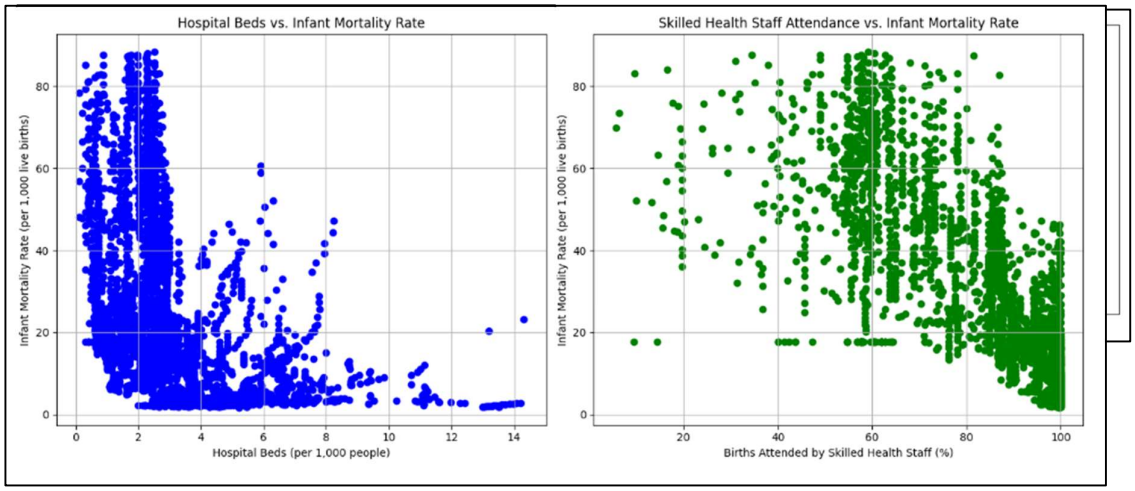
- EDA by Grace Evangelene Avula Lael – 50595809:

1. Question 1: Does the availability of adequate healthcare resources reduce maternal and infant mortality rates?
   - Hypothesis 1: More physicians and other trained healthcare professionals, like nurses and midwives, are on hand in those countries that realize lower comparative maternal mortality rates.
   - Hypothesis 2: Access to delivery with competent health personnel and sufficient hospital bed capacity can markedly reduce the rates of both maternal and infant mortalities.
   - Implementation of Hypothesis 1:

     The relationship between the number of healthcare professionals-physicians and nurses/midwives-and maternal mortality rates across different regions was analysed. Scatter plots are created to visually represent these relationships, with the x-axis representing the number of healthcare professionals per 1,000 people and the y-axis showing the

maternal mortality ratio, per 100,000 live births. The plots allow the testing of whether higher numbers of healthcare workers are associated with lower levels of maternal mortality. The code should provide a visualization that makes the testing of this hypothesis clear.

- Visualizations:



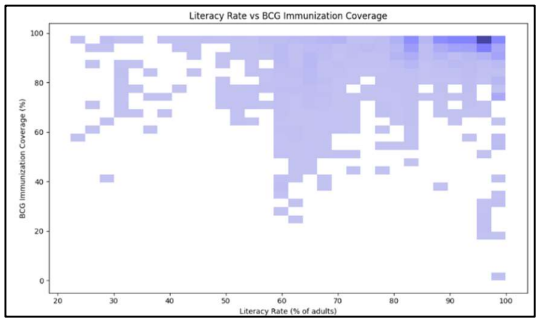
- Observation: The visualizations confirm the hypothesis by showing that regions with more physicians and nurses/midwives tend to have lower maternal mortality ratios.

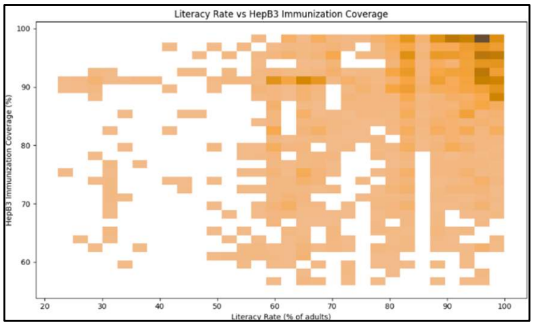
2. Question 2. How does women literacy rate affect healthcare during pregnancy and the child's life?
   - Hypothesis 1: The more literate a community is, the pickier the mothers are in having their children receive vaccinations to improve immunization levels.
   - Hypothesis 2: Higher levels of female literacy are expected to be associated with lower participation in high-risk behaviours, such as tobacco use, during pregnancy.
   - Implementation of Hypothesis 1:

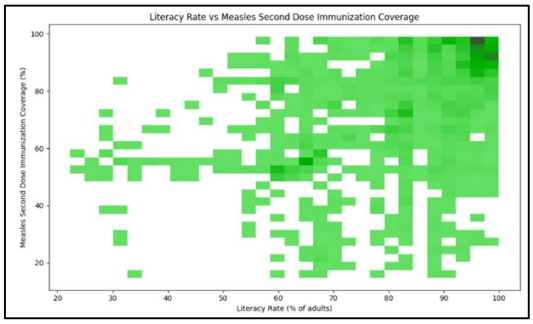
     This explores the association between literacy rates and the immunization coverage pertaining to different types of vaccinations: BCG, HepB3, and Measles second dose. In this, for each kind of vaccination, a histogram is drawn that compares the adult population's

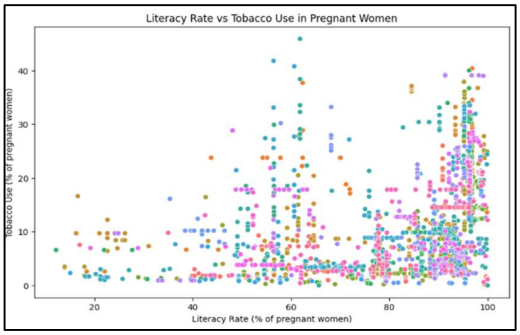
literacy rate to the immunization coverage for one-year-old children. The histograms allow us to see if higher vaccination rates are found in areas where the literacy rates are also higher. The visualizations allow clear testing of the hypothesis through the trends that show the association of literacy rates with improved immunization coverage.

- Visualizations:









- Observation: The visualizations confirm the hypothesis by showing that higher literacy rates are associated with higher immunization coverage for BCG, HepB3, and Measles.

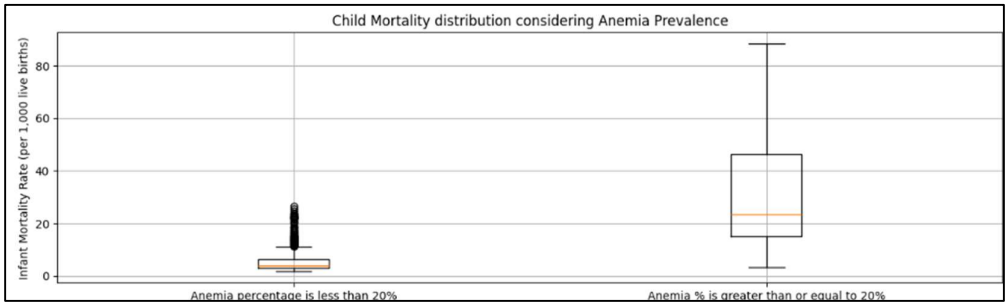
- EDA by Sharanya Nallapeddi – 50593866:

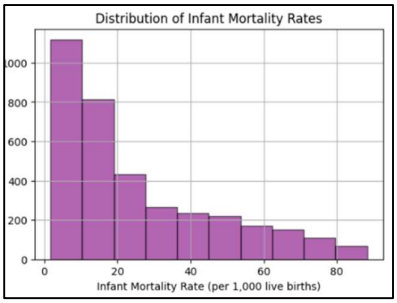
1. Question 1: Is there a correlation between the prevalence of anaemia in children & the number of child deaths?
   - Hypothesis 1: The likelihood of child mortality increases with the prevalence of anaemia in children
   - Hypothesis 2: The correlation between infant anaemia prevalence & infant mortality is substantially lower in areas with high levels of socioeconomic development & maternal education.

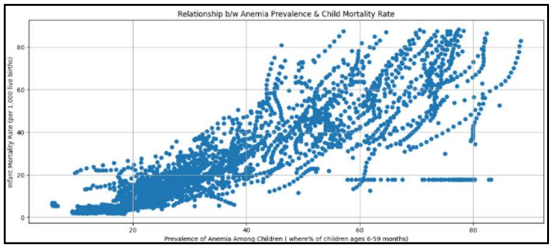
- Implementation of Hypothesis 1:

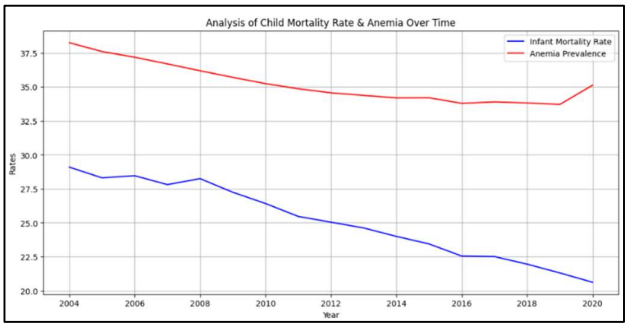
The code now addresses the hypothesis of analysing the prevalence of anaemia in children versus the infant mortality rate. It does this through various visualizations, which include boxplots that distinguish between infant mortality rates based on the prevalence of anaemia, scatter plots that explore the correlation between the rates of anaemia and infant mortality, and a heat map showing the correlation of multiple factors, including anaemia and mortality. Further, a time series plot is used for analysing the trend in anaemia prevalence and infant mortality rates. These, together, show whether the higher the prevalence of anaemia, the higher the rate of infant mortality.

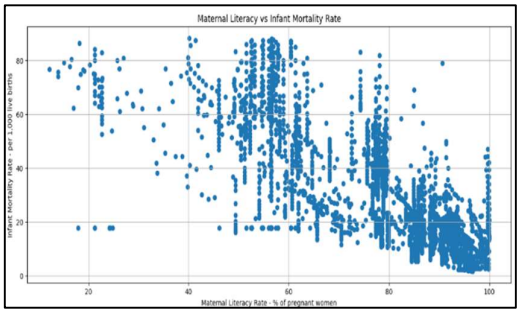
- Visualizations:

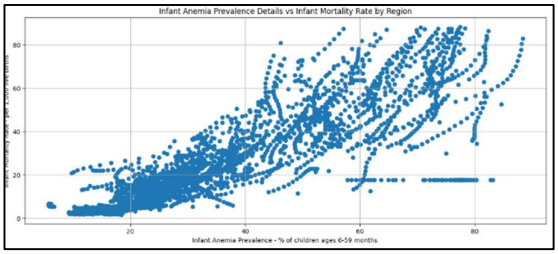


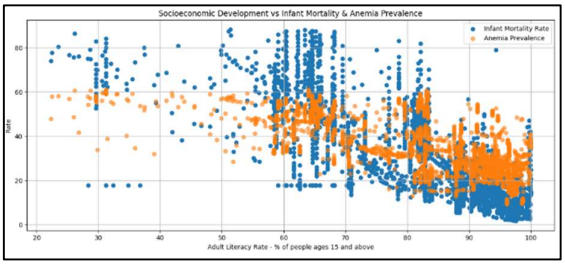


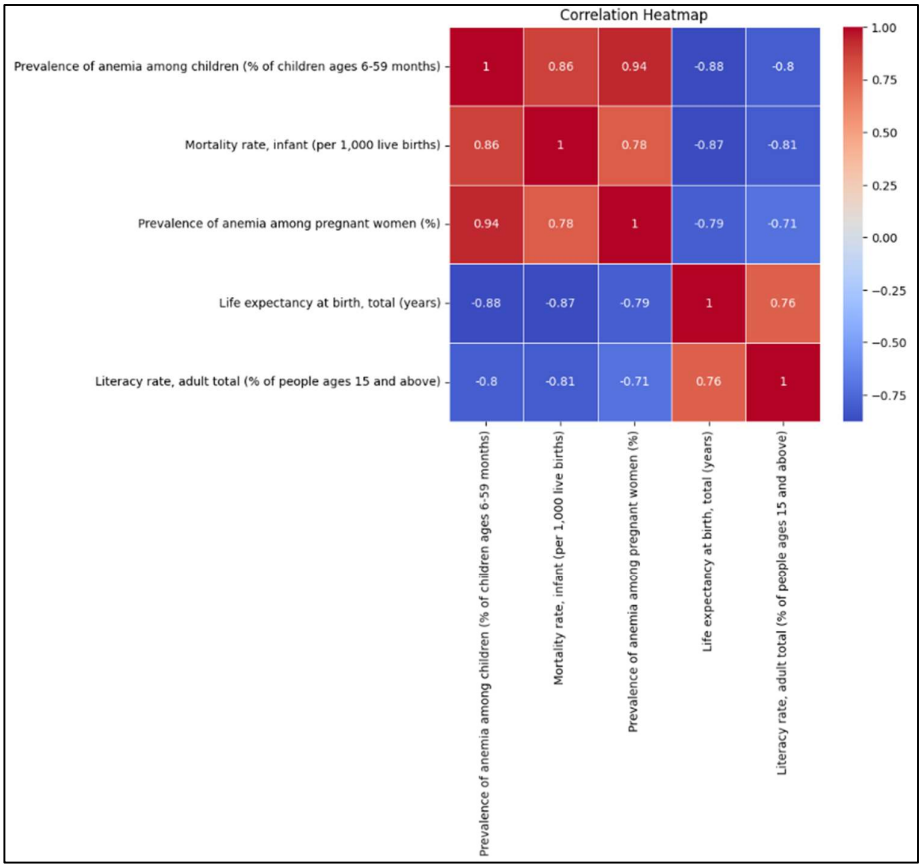












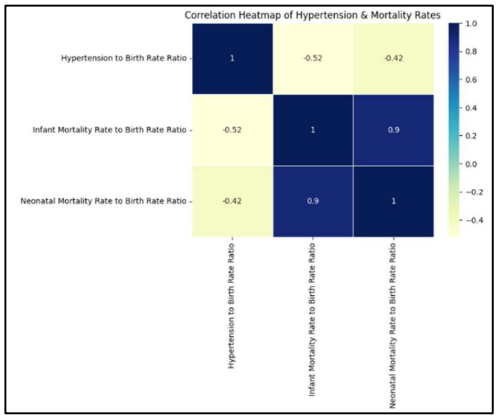
- Observation: The visualizations support the hypothesis by showing a clear correlation between higher anaemia prevalence and higher infant mortality rates.

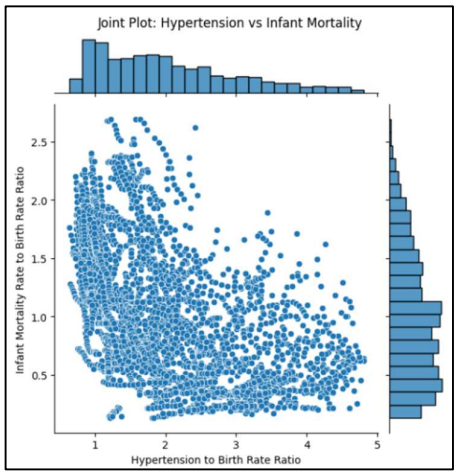
2. Question 2. How does maternal hypertension correlate with low birthweight & infant mortality?
   - Hypothesis 1: Increased maternal hypertension is strongly related with higher rates of low birthweight & infant mortality rates.
   - Hypothesis 2: Exposure to maternal healthcare services serves as a mediating factor in the correlation between low birthweight/infant mortality & maternal hypertension.

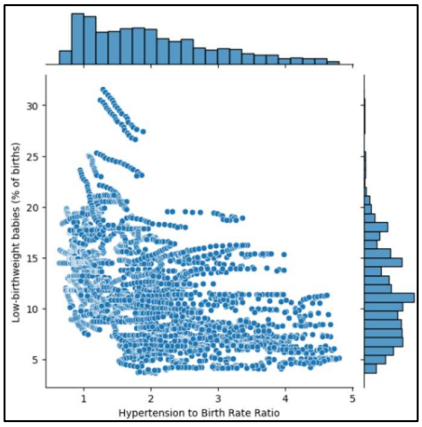
- Implementation of Hypothesis 1:

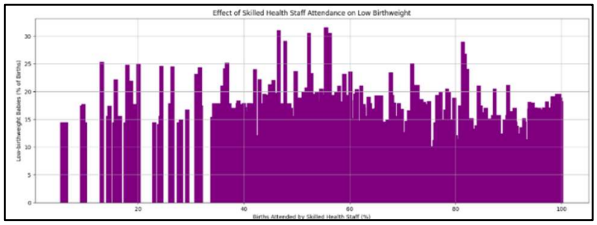
  This code approaches the hypothesis through an analysis of maternal hypertension and its correlation with low birthweight and infant mortality rates. Visualizations are used, for instance: heatmaps to provide a general overview of the relation between the levels of hypertension and the rates of death, scatter plots of joint distribution between hypertension, infant mortality, and low birthweight, and regression plots examining the relation between the level of hypertension with hospital bed availability and skilled attendance at delivery. All these visualizations together help in proving the hypothesis, as they indicate that with increased rates of maternal hypertension, the chances of low birthweight and infant mortality will also be higher, mainly in case of consideration of healthcare factors such as hospital beds and skilled health staff.

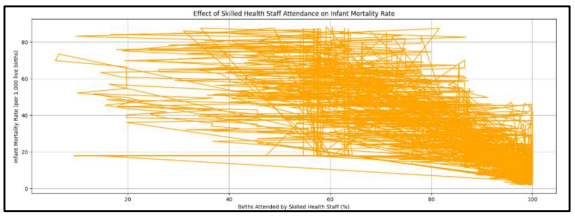
- Visualisations:

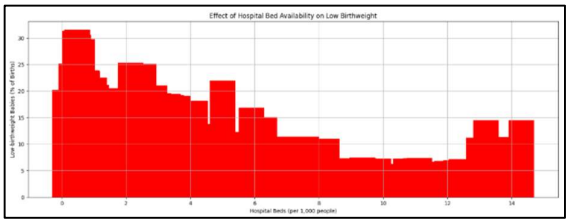


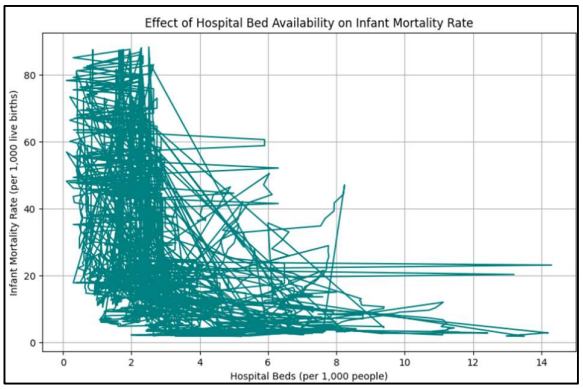


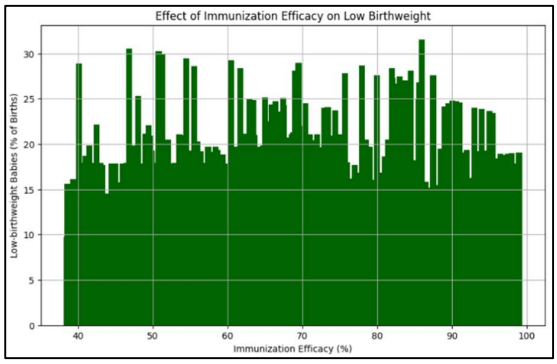


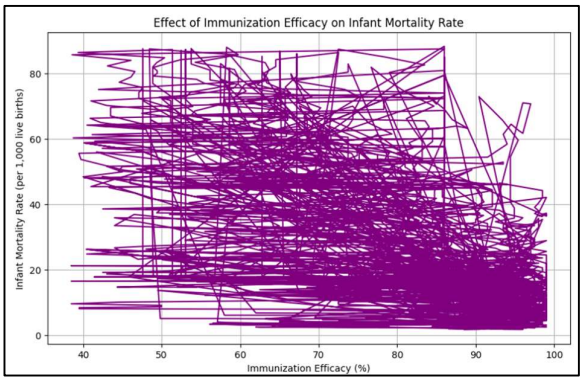


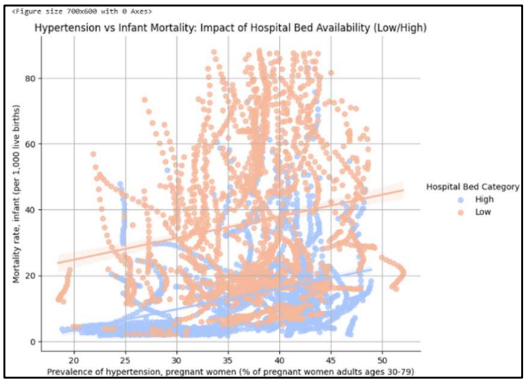


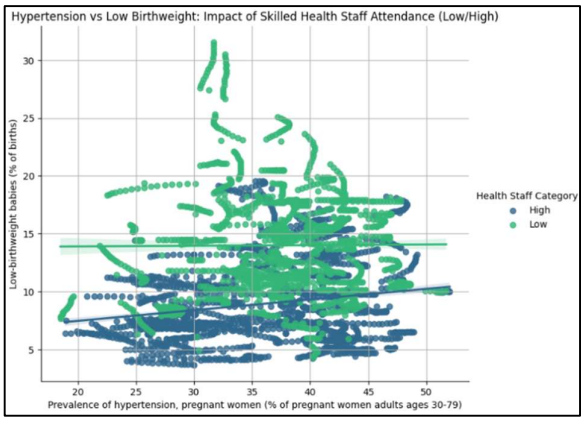












- Observation: The visualizations validate the hypothesis by showing a clear positive correlation between maternal hypertension and both low birthweight and infant mortality rates.

V. STATISTICAL MODELLING:

Statistical modelling is a mathematical framework employed in the analysis of data to find patterns or reach conclusions or predictions for a given population or system. It involves the creation of models using observable data to determine the interaction between different features. These range from simple linear regression to complex algorithms involving logistic regression and machine learning approaches. It can also provide some useful insights from unstructured data, testing the hypothesis in an organized manner by using statistical modelling, which offers a way to measure uncertainty and test the importance of associations.

In predicting infant survival and well-being, statistical modelling helps in identifying the important variables that affect health outcomes. Models can identify trends in maternal health, medical interventions, and socioeconomic circumstances affecting infant survival rates by analysing health data from pregnancy through early childhood. These understandings allow medical professionals to anticipate dangers, create focused interventions, and enhance early childhood outcomes. For example, classification models can predict the likelihood of survival depending on availability to qualified healthcare, and regression models assess the impact of prenatal nutrition on birth weight. Thus, statistical modelling will help improve decision-making in public health by translating complicated datasets into usable knowledge.

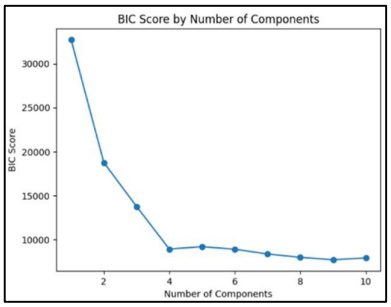
- Modelling by Anchal Daga – 50609480:

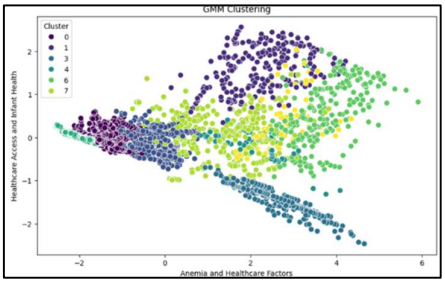
1. Question 1: Does the incidence of anaemia in pregnant women connect with the prevalence of anaemia in newborns, and how is this relationship impacted by maternal nutrition programs or availability to qualified medical personnel?
    - Hypothesis 1: A higher prevalence of anaemia among pregnant women will correlate with an increased prevalence of anaemia in infants.
    - Hypothesis 2: In a region, those with greater percentages of deliveries attended by experienced health workers would display lower rates of anaemia in infants.
    - Implementation of Hypothesis 1:

      The code addresses this hypothesis through a GMM by clustering countries based on health and anaemia-related features. The main

assumption in the hypothesis is that there exists a certain degree of relationship between the proportion of anaemia among pregnant women and infants. This is explored by standardizing the code into relevant features (prevalence of anaemia among pregnant women, prevalence among children, and other healthcare factors) to bring them to a comparable scale, and then applies GMM, a probabilistic clustering algorithm that assumes data is generated from a mixture of several Gaussian distributions. GMM was selected due to its ability to detect complex patterns in the data, which are not strictly linear, and handle overlapped clusters. It works by fitting the data on a series of Gaussian distributions and returns the optimal number of clusters using the Bayesian Information Criterion-criterion that penalizes overfitting. After fitting the model, it assigns each data point, that is, a country to a specific cluster, showing the pattern of anaemia prevalence.

- Visualisations:






```
Cluster Means:
         Prevalence of anemia among pregnant women (%)  \
Cluster
0                                             25.143453
1                                             40.927632
2                                             34.561793
3                                             49.163354
4                                             55.203654
5                                             19.027208
6                                             53.453761
7                                             38.329182
8                                             41.170000

         Prevalence of anemia among children (% of children ages 6-59 months)
Cluster
0                                                23.936876
1                                                54.139430
2                                                33.773879
3                                                66.316738
4                                                67.865385
5                                                14.906003
6                                                69.652569
7                                                44.785383
8                                                46.842540

         Births attended by skilled health staff (% of total)  \
Cluster
0                                             96.926383
1                                             78.783114
2                                             92.241559
3                                             57.980244
4                                             85.741538
5                                             99.131769
6                                             59.716009
7                                             72.086227
8                                             40.847937

         Mortality rate, infant (per 1,000 live births)  \
Cluster
0                                             12.680484
1                                             44.440439
2                                             22.456725
3                                             61.334756
4                                             35.412308
5                                              4.048427
6                                             58.879679
7                                             37.980042
8                                             45.591429

         Maternal mortality ratio (modeled estimate, per 100,000 live births)
Cluster
0                                                41.022244
1                                               432.250000
2                                                76.456140
3                                                64.000000
4                                               181.461538
5                                                 7.305374
6                                               382.298165
7                                               189.686016
8                                               338.158730
```

```
T-test between Cluster 3 and Cluster 4 for Infant Anemia: t-statistic=-1.0158231969836977, p-value=0.31036333425939816
Means of the components:
[[-0.55620335 -0.57480036  0.6171185  -0.57290489 -0.49515754]
 [ 0.78883995  0.9751734  -0.4145241   0.90439302  2.3007859 ]
 [ 0.16299592 -0.08883814  0.36271522 -0.14136417 -0.24239416]
 [ 1.34006446  1.58116197 -1.57331234  1.66990436 -0.32917133]
 [ 1.81730689  1.65737348 -0.02738578  0.46803343  0.51441578]
 [-1.05026846 -1.03859428  0.74159514 -0.97250494 -0.73697579]
 [ 1.6584507   1.71896557 -1.47032902  1.51620019  1.91773475]
 [ 0.46386357  0.45542071 -0.73435805  0.56768952  0.54605695]
 [ 0.65871308  0.58721851 -2.4718915   0.95344949  1.6438753 ]]
Covariances of the components:
[[[ 1.35645775e-01  5.08447026e-02 -8.34631894e-03  1.02874117e-02
   -7.50827295e-03]
  [ 5.08447026e-02  5.20568839e-02 -9.37263317e-03  2.52653574e-02
    3.84221289e-03]
  [-8.34631894e-03 -9.37263317e-03  4.45370578e-02 -1.67629669e-02
   -1.42557136e-02]
  [ 1.02874117e-02  2.52653574e-02 -1.67629669e-02  4.45910209e-02
    2.24774571e-02]
  [-7.50827295e-03  3.84221289e-03 -1.42557136e-02  2.24774571e-02
    3.44389193e-02]]

 [[ 4.38350773e-01  2.41875584e-01  4.79677272e-02  7.84888562e-02
    1.85427504e-01]
  [ 2.41875584e-01  2.08102669e-01 -1.91094980e-02  9.81832042e-02
    9.16174520e-02]
  [ 4.79677272e-02 -1.91094980e-02  2.85840009e-01 -1.39862808e-01
   -9.07016338e-02]
  [ 7.84888562e-02  9.81832042e-02 -1.39862808e-01  2.03791764e-01
    1.63118608e-01]
  [ 1.85427504e-01  9.16174520e-02 -9.07016338e-02  1.63118608e-01
    4.68563055e-01]]

 [[ 2.75890379e-01  1.03231273e-01  7.05733763e-02 -3.73857280e-03
    5.33515871e-03]
  [ 1.03231273e-01  1.06797992e-01  1.12340755e-02  1.43655758e-02
    2.89908284e-03]
  [ 7.05733763e-02  1.12340755e-02  1.12212805e-01  4.21503013e-04
   -9.43338018e-03]
  [-3.73857280e-03  1.43655758e-02  4.21503013e-04  9.65547347e-02
    3.80522462e-03]
  [ 5.33515871e-03  2.89908284e-03 -9.43338018e-03  3.80522462e-03
    6.68751259e-02]]

 [[ 4.04313582e-01  2.98029587e-01  1.05383860e-01 -4.36010313e-02
   -8.93976555e-19]
  [ 2.98029587e-01  2.79196842e-01  5.64030958e-02 -4.00100638e-02
   -6.48446379e-19]
  [ 1.05383860e-01  5.64030958e-02  6.65666849e-01 -4.45804806e-02
   -4.46332828e-18]
  [-4.36010313e-02 -4.00100638e-02 -4.45804806e-02  8.83881159e-01
   -1.00610149e-18]
  [-8.93976555e-19 -6.48446379e-19 -4.46332828e-18 -1.00610149e-18
    1.00000000e-06]]

 [[ 4.12732329e-02  8.49667551e-02 -7.20026456e-02  8.34906327e-02
    6.63754749e-03]
  [ 8.49667551e-02  2.13951553e-01 -1.51525757e-01  1.35146673e-01
   -2.09933187e-02]
  [-7.20026456e-02 -1.51525757e-01  5.32032575e-01 -1.41009682e-01
```

```
[ 1.05383860e-01  5.64030958e-02  6.65666849e-01 -4.45804806e-02
 -4.46332828e-18]
[-4.36010313e-02 -4.00100638e-02 -4.45804806e-02  8.83881159e-01
 -1.00610149e-18]
[-8.93976555e-19 -6.48446379e-19 -4.46332828e-18 -1.00610149e-18
  1.00000000e-06]]

[[ 4.12732329e-02  8.49667551e-02 -7.20026456e-02  8.34906327e-02
  6.63754749e-03]
 [ 8.49667551e-02  2.13951553e-01 -1.51525757e-01  1.35146673e-01
 -2.09933187e-02]
 [-7.20026456e-02 -1.51525757e-01  5.32032575e-01 -1.41009682e-01
 -1.18123277e-02]
 [ 8.34906327e-02  1.35146673e-01 -1.41009682e-01  2.35646493e-01
  4.54012080e-02]
 [ 6.63754749e-03 -2.09933187e-02 -1.18123277e-02  4.54012080e-02
  4.76762956e-02]]

[[ 1.22316387e-01  6.90036376e-02  7.07752384e-03  7.94589177e-03
 -1.06320715e-03]
 [ 6.90036376e-02  4.54862251e-02  2.90447678e-03  6.19440900e-03
 -6.73197695e-04]
 [ 7.07752384e-03  2.90447678e-03  2.13591894e-03  2.91450579e-04
 -1.29213550e-05]
 [ 7.94589177e-03  6.19440900e-03  2.91450579e-04  4.69386606e-03
  8.48552353e-04]
 [-1.06320715e-03 -6.73197695e-04 -1.29213550e-05  8.48552353e-04
  6.04547542e-04]]

[[ 1.07815264e-01  1.39912816e-01  5.32182830e-02  1.78300043e-02
  1.57297804e-01]
 [ 1.39912816e-01  2.27246422e-01  1.06313439e-01  6.77027669e-02
  1.97274202e-01]
 [ 5.32182830e-02  1.06313439e-01  3.93070774e-01 -5.03059143e-03
  3.50216810e-02]
 [ 1.78300043e-02  6.77027669e-02 -5.03059143e-03  5.50805285e-01
  8.22023919e-02]
 [ 1.57297804e-01  1.97274202e-01  3.50216810e-02  8.22023919e-02
  7.07380624e-01]]

[[ 4.39557250e-01  3.22598633e-01 -1.15251032e-01  1.43491160e-01
  1.64064009e-01]
 [ 3.22598633e-01  3.71300300e-01 -1.29109501e-01  2.47258008e-01
  2.08226331e-01]
 [-1.15251032e-01 -1.29109501e-01  5.81636084e-01 -1.35149203e-01
 -9.38501746e-02]
 [ 1.43491160e-01  2.47258008e-01 -1.35149203e-01  3.32357344e-01
  2.20505260e-01]
 [ 1.64064009e-01  2.08226331e-01 -9.38501746e-02  2.20505260e-01
  3.61771292e-01]]

[[ 3.44020772e-01 -4.20712932e-02  5.75725280e-02 -6.43789924e-02
 -2.05296442e-01]
 [-4.20712932e-02  3.12448724e-02 -9.79687721e-02  9.52527872e-03
  8.28724307e-02]
 [ 5.75725280e-02 -9.79687721e-02  8.42128975e-01  6.05721489e-02
 -1.63038667e-01]
 [-6.43789924e-02  9.52527872e-03  6.05721489e-02  2.67388609e-01
  3.15333584e-01]
 [-2.05296442e-01  8.28724307e-02 -1.63038667e-01  3.15333584e-01
  6.59174224e-01]]]
```

- Observation: The output provides the cluster means, which give the prevalence of anaemia among both pregnant women and children for each cluster, and the covariance between the features. The t-test comparison of the two clusters, Cluster 3 and Cluster 4, indicates no significant difference in the prevalence of anaemia among children in these clusters, with a p-value of 0.31, hence the relationship between anaemia in pregnant women and infants may not be that strong or consistent across all clusters as initially hypothesized. Yet the observed trend in cluster means support the hypothesis to some degree; for example, high prevalence of anaemia in pregnant women parallels high prevalence in children, even if deeper analysis would perhaps have to be done to draw these conclusions.

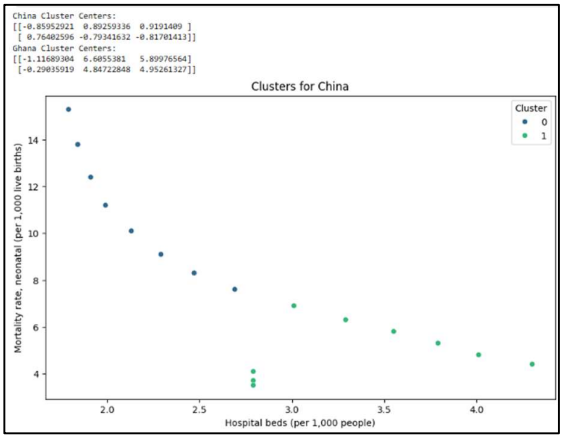
2. Question 2: What is the relationship between hospital infrastructure and the rates of stillbirths and neonatal deaths in different nations?

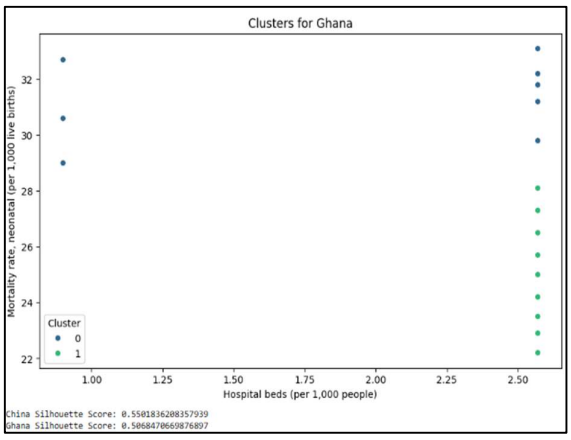
- Hypothesis 1: In China, higher hospital resources will be associated with lower rates of neonatal deaths and stillbirths.
- Hypothesis 2: In Ghana, higher hospital resources will be associated with lower rates of neonatal deaths and stillbirths.
- <u>Implementation of Hypothesis 1:</u>

  K-means clustering is one of the most common unsupervised machine learning algorithms applied to testing the hypothesis, in view of hospital resources and neonatal outcomes in China and Ghana. The hypothesis was that there will be a lower rate of neonatal deaths and stillbirths associated with increased hospital resources. K-Means groups the data into clusters of similar hospital bed availability and mortality rates. Standardization puts the features on the same scale to ensure clustering is not affected by different units of measurement. Then, the algorithm assigns data points or regions to clusters in a way that minimizes the variance within each group. The silhouette score is a measure of cluster quality for all datasets, where higher scores mean better-defined clusters.

  The algorithm works in this case, where the K-Means method can detect a pattern in the relationship between hospital resources and health outcomes, without any a priori knowledge about the number of clusters or exactly what that relationship is.

- <u>Visualisations:</u>





- Observation: The cluster centres and silhouette scores that result-0.55 for China and 0.51 for Ghana-indicate that the two countries do indeed show distinct groupings based on hospital resources and health outcomes, which serves as indirect support for the hypothesis. These results, however, are partially supported by the moderate silhouette scores, suggesting that the clustering may not be perfectly definitive, and further analysis or refinement of the model might be necessary to fully confirm the hypothesis.

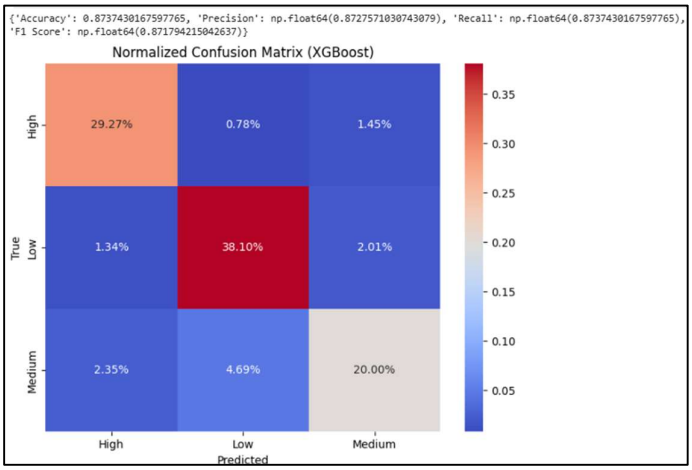
- Modelling by Keerthana Vangala – 50604773:

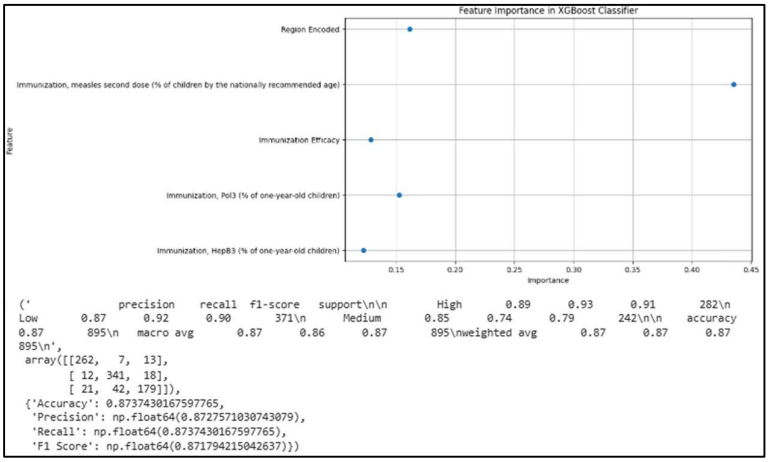
  1.Question 1: What is the association between immunization coverage and infant mortality?

  - Hypothesis 1: Higher immunization rates for HepB, Polio, and Measles are linked with lower infant mortality rates for specific regions. If this is correct it means that vaccination given in the 1st one year to a baby are protecting them against the respective diseases.
  - Hypothesis 2: Different regions with higher immunization have lower Infant mortality rates.
  - Implementation of Hypothesis 1:

    This code will address the hypothesis by using a gradient boosting algorithm, XGBoost, to analyse the relationship between the immunization rates (HepB, Polio, and Measles) and infant mortality rates in different regions. The hypothesis here is that high immunization rates are related to low infant mortality rates, and the code should predict and categorize the infant mortality rates based on the efficacy of the immunization. First, the data is pre-processed to categorize infant mortality into three levels: Low, Medium, and High, with thresholds set for the mortality rate values. Then, the features of immunization rates are used to train the XGBoost classifier, which builds an ensemble of decision trees to make predictions about the mortality category.

- Visualisations:





- Observation: XGBoost works well here, since the problem includes complex relationships and non-linearities within the data, hence befitting for this classification task. It uses 75% of the data to train the model and 25% to test it. It will check the accuracy, precision, recall, and F1 score. Indeed, it performs very well since it has an accuracy of 87.37%, precision of 87.28%, and recall of 87.37%. It also performs well in both high and low mortality categories while having a medium category with slightly lower results, according to the confusion matrix. According to the feature importance plot, immunization rates have much to say about the predictions made by the model, which suggests that the higher the immunization rates, the lower the rate of infant mortality.

2.Question 2: How do different regions vary based on the infant mortality rates?

- Hypothesis 1: Analysis Infant mortality rates in different regions based on various factors like low birth weights, number of infant deaths and still births to see where the mortality rates are high.
- Hypothesis 2: Does the Lower birth weight have increased dependency on Mortality rates of specific region?
- Implementation of Hypothesis 1:

  The code addresses the hypothesis by analysing the factors that influence infant mortality rates across different regions, such as low birth weights, the number of infant deaths, stillbirths, and mortality rate ratios. The approach used in this is a Decision Tree Classifier, which is a supervised learning algorithm that splits the data into branches based on feature values to classify regions into different mortality categories. The model is trained on features like "Low-birthweight babies," "Mortality rate, infant," and "Number of stillbirths," for which the region name is predicted. This algorithm works in a way that it finds patterns and the relationship between input features and output categories (regions) to effectively classify regions with high or low infant mortality rates. Accuracy, precision, recall, the F1 score, and ROC curves are used to evaluate the model, reflecting the performance of the classifier.

- Visualisations:

```
Dcision Tree classifier
Accuracy 0.8932135728542914
Precision 0.9102030066850426
Recall 0.8932135728542914
F1 Score 0.8896647093855617
```

accuracy			0.89	1002
macro avg	0.90	0.90	0.89	1002
weighted avg	0.91	0.89	0.89	1002

Classification Report:

	precision	recall	f1-score	support
Afghanistan	1.00	1.00	1.00	5
Africa Eastern and Southern	1.00	0.20	0.33	5
Africa Western and Central	0.67	1.00	0.80	8
Albania	1.00	0.86	0.92	7
Algeria	0.80	1.00	0.89	4
Angola	1.00	1.00	1.00	5
Antigua and Barbuda	1.00	1.00	1.00	3
Arab World	1.00	1.00	1.00	4
Argentina	1.00	0.75	0.86	4
Armenia	1.00	0.62	0.77	8
Australia	1.00	1.00	1.00	8
Austria	1.00	1.00	1.00	6
Azerbaijan	1.00	1.00	1.00	7
Bahamas, The	1.00	1.00	1.00	5
Bahrain	1.00	1.00	1.00	8
Bangladesh	1.00	1.00	1.00	6
Barbados	0.88	1.00	0.93	7
Belarus	0.60	0.75	0.67	4
Belgium	1.00	1.00	1.00	7
Belize	1.00	0.50	0.67	4
Benin	1.00	1.00	1.00	4
Bhutan	1.00	1.00	1.00	5
Bolivia	0.67	1.00	0.80	2
Bosnia and Herzegovina	1.00	1.00	1.00	6
Botswana	1.00	0.67	0.80	6
Brazil	1.00	0.86	0.92	7
Brunei Darussalam	1.00	1.00	1.00	6
Bulgaria	0.50	1.00	0.67	5
Burkina Faso	1.00	1.00	1.00	4
Burundi	1.00	1.00	1.00	5
Cabo Verde	1.00	1.00	1.00	4
Cambodia	1.00	1.00	1.00	6
Cameroon	0.89	1.00	0.94	8
Canada	1.00	1.00	1.00	6
Caribbean small states	0.67	1.00	0.80	6
Central African Republic	1.00	1.00	1.00	4
Central Europe and the Baltics	1.00	1.00	1.00	2
Chad	0.80	1.00	0.89	4
Chile	1.00	0.80	0.89	5
China	1.00	1.00	1.00	5
Colombia	1.00	1.00	1.00	5
Comoros	1.00	1.00	1.00	4
Congo, Dem. Rep.	1.00	1.00	1.00	4
Congo, Rep.	1.00	1.00	1.00	5
Costa Rica	1.00	1.00	1.00	5
Cote d'Ivoire	1.00	1.00	1.00	4
Croatia	1.00	1.00	1.00	6
Cuba	1.00	1.00	1.00	8
Cyprus	1.00	1.00	1.00	5
Czechia	1.00	1.00	1.00	3
Denmark	1.00	1.00	1.00	7
Djibouti	1.00	1.00	1.00	3
Dominican Republic	1.00	1.00	1.00	5
East Asia & Pacific	0.00	0.00	0.00	3
East Asia & Pacific (IDA & IBRD countries)	0.00	0.00	0.00	3
East Asia & Pacific (excluding high income)	0.00	0.00	0.00	7
Ecuador	1.00	0.75	0.86	4
Egypt, Arab Rep.	1.00	1.00	1.00	3
El Salvador	1.00	1.00	1.00	5
Equatorial Guinea	1.00	1.00	1.00	5
Eritrea	1.00	1.00	1.00	7
Estonia	0.75	1.00	0.86	6
Eswatini	1.00	1.00	1.00	8
Ethiopia	1.00	1.00	1.00	4
Euro area	1.00	0.83	0.91	6
Europe & Central Asia	1.00	1.00	1.00	5
Europe & Central Asia (IDA & IBRD countries)	0.80	1.00	0.89	4
Europe & Central Asia (excluding high income)	0.80	0.80	0.80	5
European Union	0.83	1.00	0.91	5
Fiji	1.00	1.00	1.00	2
Finland	1.00	1.00	1.00	4
France	1.00	1.00	1.00	5
Gabon	1.00	1.00	1.00	5
Gambia, The	1.00	1.00	1.00	5
Georgia	1.00	0.75	0.86	4
Germany	0.60	0.75	0.67	4
Ghana	1.00	1.00	1.00	2
Greece	1.00	1.00	1.00	2
Grenada	1.00	1.00	1.00	4
Guatemala	1.00	1.00	1.00	5
Guinea	1.00	1.00	1.00	3
Guinea-Bissau	0.67	1.00	0.80	4
Guyana	1.00	1.00	1.00	4
Haiti	1.00	1.00	1.00	6
Honduras	1.00	1.00	1.00	4
Hungary	1.00	1.00	1.00	8
Iceland	1.00	1.00	1.00	2
India	1.00	1.00	1.00	7
Indonesia	1.00	1.00	1.00	5
Iran, Islamic Rep.	1.00	1.00	1.00	2
Iraq	1.00	0.67	0.80	6
Ireland	0.67	0.50	0.57	4
Israel	1.00	1.00	1.00	8
Italy	0.83	1.00	0.91	5
Jamaica	1.00	0.40	0.57	10
Japan	1.00	1.00	1.00	2
Jordan	1.00	1.00	1.00	6
Kazakhstan	1.00	1.00	1.00	6
Kenya	0.60	1.00	0.75	3
Kiribati	1.00	1.00	1.00	3
Korea, Dem. People's Rep.	0.50	1.00	0.67	4
Korea, Rep.	0.67	0.50	0.57	4
Kuwait	1.00	0.38	0.55	8
Kyrgyz Republic	1.00	1.00	1.00	5
Lao PDR	1.00	1.00	1.00	4
Latin America & Caribbean	0.25	0.33	0.29	3
Latin America & Caribbean (excluding high income)	0.50	0.50	0.50	4
Latin America & the Caribbean (IDA & IBRD countries)	0.25	0.20	0.22	5
Latvia	0.00	0.00	0.00	5
Lebanon	1.00	0.57	0.73	7
Lesotho	1.00	0.75	0.86	4
Liberia	1.00	1.00	1.00	5
Libya	1.00	1.00	1.00	4
Lithuania	0.60	0.75	0.67	4
Luxembourg	1.00	1.00	1.00	7
Madagascar	1.00	1.00	1.00	4
Malawi	1.00	1.00	1.00	2
Malaysia	1.00	1.00	1.00	4
Maldives	1.00	0.67	0.80	9
Mali	1.00	1.00	1.00	3
Malta	1.00	1.00	1.00	2
Mauritania	1.00	1.00	1.00	4
Mauritius	1.00	1.00	1.00	4
Mexico	1.00	1.00	1.00	4
Micronesia, Fed. Sts.	0.78	1.00	0.88	7
Middle East & North Africa	1.00	1.00	1.00	3
Middle East & North Africa (IDA & IBRD countries)	1.00	1.00	1.00	7
Middle East & North Africa (excluding high income)	1.00	1.00	1.00	1
Middle income	1.00	1.00	1.00	4
Moldova	0.67	1.00	0.80	4
Mongolia	0.83	1.00	0.91	5
Montenegro	1.00	1.00	1.00	6
Morocco	1.00	1.00	1.00	4
Mozambique	1.00	1.00	1.00	5
Myanmar	1.00	0.67	0.80	3
Namibia	1.00	1.00	1.00	6
Nepal	1.00	1.00	1.00	5
Netherlands	1.00	1.00	1.00	4
New Zealand	0.00	0.00	0.00	1
Nicaragua	0.80	0.80	0.80	5
Niger	1.00	1.00	1.00	5
Nigeria	1.00	1.00	1.00	3
North America	1.00	1.00	1.00	1
North Macedonia	1.00	0.83	0.91	6
Norway	1.00	1.00	1.00	2
Solomon Islands	1.00	1.00	1.00	3
Somalia	0.67	1.00	0.80	6
South Africa	1.00	0.50	0.67	8
South Asia	1.00	1.00	1.00	5
South Sudan	1.00	1.00	1.00	6
Spain	1.00	1.00	1.00	3
Sri Lanka	1.00	0.80	0.89	5
St. Lucia	1.00	1.00	1.00	3
St. Vincent and the Grenadines	1.00	1.00	1.00	3
Sub-Saharan Africa	0.00	0.00	0.00	4
Sub-Saharan Africa (IDA & IBRD countries)	0.00	0.00	0.00	2
Sub-Saharan Africa (excluding high income)	0.00	0.00	0.00	6
Sudan	1.00	1.00	1.00	5
Suriname	1.00	1.00	1.00	3
Sweden	1.00	1.00	1.00	7
Switzerland	1.00	1.00	1.00	3
Syrian Arab Republic	1.00	1.00	1.00	6
Tajikistan	1.00	0.50	0.67	4
Tanzania	1.00	1.00	1.00	3
Thailand	1.00	1.00	1.00	7
Timor-Leste	1.00	1.00	1.00	5
Togo	1.00	1.00	1.00	1
Tonga	1.00	1.00	1.00	7
Trinidad and Tobago	1.00	1.00	1.00	3
Tunisia	1.00	1.00	1.00	1
Turkiye	1.00	1.00	1.00	5
Turkmenistan	1.00	1.00	1.00	3
Uganda	1.00	1.00	1.00	6
Ukraine	1.00	0.83	0.91	6
United Arab Emirates	1.00	1.00	1.00	7
United Kingdom	0.80	0.80	0.80	5
United States	1.00	1.00	1.00	3
Uruguay	0.67	1.00	0.80	6
Uzbekistan	1.00	1.00	1.00	8
Vanuatu	1.00	1.00	1.00	2
Venezuela, RB	0.60	1.00	0.75	3
Viet Nam	1.00	1.00	1.00	1
West Bank and Gaza	1.00	1.00	1.00	4
Yemen, Rep.	1.00	1.00	1.00	6
Zambia	1.00	1.00	1.00	2
Zimbabwe	1.00	1.00	1.00	4

- Observation: The results show an accuracy of 89.3%, a precision of 91%, and a recall of 89.3%, indicating that this model is effectively able to identify the region based on infant mortality rates. This has, therefore, validated the hypothesis through the selected features being truly representative in identifying regional mortality rate differences. Moreover, the feature importance plot shows the most influencing factors, such as low birthweight or death rates of infants, on the mortality prediction.

- Modelling by Grace Evangelene Avula Lael – 50595809:

  1. Question 1: How do socio-economic factors, health infrastructure, and disease prevalence together influences the life expectancy in different regions?
     - Hypothesis 1: Medical resources, literacy rates, and prevalence of diseases and their respective mortality rates of a given region directly impact the life expectancy by either being positive or negative in nature
     - Hypothesis 2: Life expectancy will be higher in regions with better health infrastructure, higher literacy rates, and lesser prevalence of diseases than in regions with poor infrastructure and greater disease burdens.
     - Implementation of Hypothesis 1:

       The code uses the hypothesis through machine learning models that can predict life expectancy based on many socio-economic and health care factors, including medical resources, literary rates, prevalence of diseases, mortality rates, among others. It considers the use of two algorithms: Convolutional Neural Networks and K-Nearest Neighbours. It captures the complex patterns in data, especially in those cases when relationships between features can be non-linear, by using the CNN model. KNN is used as a simple, distance-based regression method to make predictions by relying on similar instances in training data. Both models have been trained on scaled data for comparable features and are evaluated on their performance in terms of Mean Squared Error. Results showing the expected and predicted life expectancy, along with a classification of the key features as "High" or "Low" depending on their impact on the same life expectancy.

     - Visualisations:

Sample Predictions with Expected and Predicted Outputs:

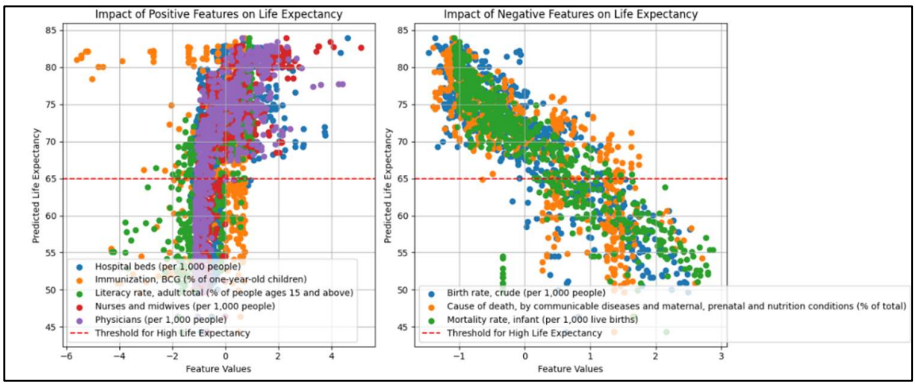
	Expected Output	Predicted Output Life Expectancy	\
0	73.57	72.574608	High
1	69.68	74.499817	High
2	56.34	55.858917	Low
3	65.73	63.812386	Low
4	72.64	76.166718	High
5	76.44	79.387421	High
6	70.94	69.373581	High
7	52.97	58.005642	Low
8	76.13	73.806969	High
9	67.21	66.594955	High

	Positive Features (High/Low)	Negative Features (High/Low)
0	High	Low
1	High	Low
2	Low	High
3	Low	High
4	High	Low
5	Low	Low
6	Low	Low
7	Low	High
8	High	Low
9	Low	Low

Mean Squared Error for Life Expectancy Prediction with KNN: 2.8784142346368706

	Expected Output	Predicted Output Life Expectancy	\
0	[73.57]	73.030	High
1	[69.68]	69.600	High
2	[56.34]	55.304	Low
3	[65.73]	66.124	High
4	[72.64]	73.166	High
5	[76.44]	76.824	High
6	[70.94]	70.928	High
7	[52.97]	54.542	Low
8	[76.13]	75.640	High
9	[67.21]	67.466	High

	Positive Features (High/Low)	Negative Features (High/Low)
0	High	Low
1	High	Low
2	Low	High
3	Low	High
4	High	Low
5	Low	Low
6	Low	Low
7	Low	High
8	High	Low
9	Low	Low



- Observation: This shows that the values of forecasted life expectancy are reasonably within what one would expect from such data; that is, the considered features affect the life expectancy of a particular country within its socio-economic and health-care space. High "Hospital beds" and "Literacy rate", for instance, are in fact associated with "High" life expectancy predictions; the same way "Mortality rate" and "Birth rate" factors result in life expectancy with lower expectations, supporting this hypothesis by affecting it either positively or negatively.

2. Question 2: How do birth rates, literacy rates, availability of medical resources, and health conditions interact with each other in affecting the maternal mortality rate in various regions?

- Hypothesis 1: Maternal mortality rate may be forecasted by the levels of Birth rates, Literacy rate, medical resources and health conditions.
- Hypothesis 2: Maternal mortality rates are influenced by birth rates, literacy rates, medical resources, and health conditions, with higher mortality in less literate regions and places where access to healthcare facilities is poor.
- Implementation of Hypothesis 1:

  The code approaches the hypothesis by using two machine learning models to predict maternal mortality rates based on birth rates, literacy rates, medical resources, and health conditions. First, there is the model Multilayer Perceptron (MLP), which is a feedforward artificial neural network that is trained using the input features from the dataset to predict maternal mortality rates. The MLP model is a fully connected network with ReLU activation, hence very good at picking complex nonlinear relationships between input features and the target variable. It was trained with MSE as the loss, appropriate for regression tasks. The

second model is KNN, a simpler algorithm that predicts the target variable by averaging the results of the nearest neighbours in the feature space. In both models, a dataset of multiple features is used for the estimation of maternal mortality rate. These algorithms work well here because they can capture patterns in data with varying levels of complexity, such as how birth rates or literacy levels correlate with mortality rates.

- Visualisations:

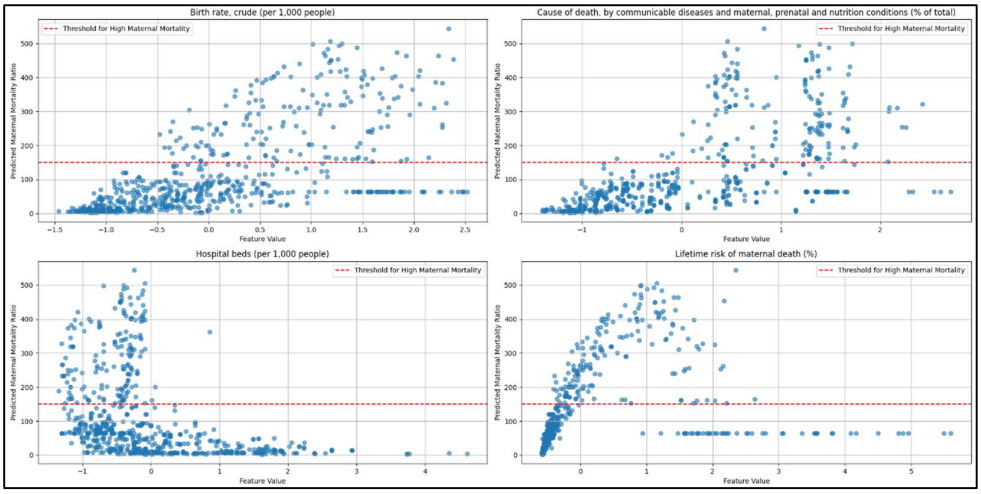
Sample Predictions with Expected and Predicted Outputs:

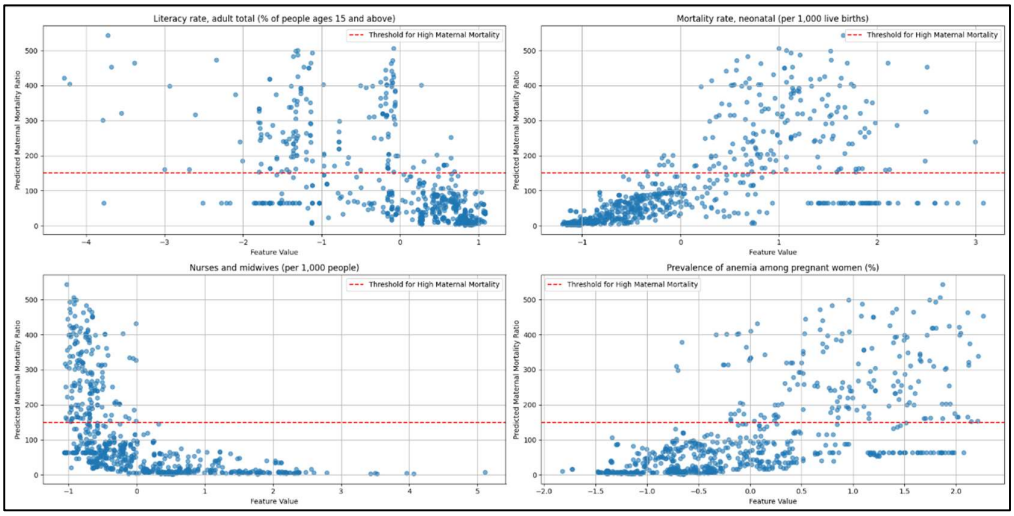
	Expected Output	Predicted Output	Maternal Mortality Ratio	Features
0	103.0	107.913345	Low	Low
1	14.0	11.465513	Low	High
2	64.0	49.122501	Low	High
3	327.0	282.577820	High	High
4	13.0	7.602503	Low	Low
5	6.0	5.443392	Low	Low
6	139.0	143.283615	Low	Low
7	435.0	484.738647	High	High
8	70.0	58.446960	Low	Low
9	39.0	42.703011	Low	Low

Mean Squared Error for Maternal Mortality Ratio Prediction with KNN: 2871.471787709497

Sample Predictions with Expected and Predicted Outputs:

	Expected Output	Predicted Output	Maternal Mortality Ratio	Features
0	103.0	116.6	Low	Low
1	14.0	14.2	Low	High
2	64.0	64.0	Low	High
3	327.0	345.2	High	High
4	13.0	19.8	Low	Low
5	6.0	6.6	Low	Low
6	139.0	138.6	Low	Low
7	435.0	386.2	High	High
8	70.0	69.8	Low	Low
9	39.0	35.8	Low	Low





- Observation: These results, including the root mean square error and predicted outputs of both models, are sufficient to tell how the models estimate maternal mortality. For instance, from the prediction, one gets an instance where the model predicted a maternal mortality ratio with its expected output. The features that make for high or low mortalities have been identified, therefore confirming the hypothesis. The fact that higher maternal mortality ratios are predicted for areas with poor health or low medical care reinforces the hypothesis that maternal mortality is related to birth rates, literacy rates, and access to healthcare.

- Modelling by Sharanya Nallapeddi – 50593866:

1. Question 1: Is there a correlation between the prevalence of anaemia in children & the number of child deaths?

   - Hypothesis 1: The likelihood of child mortality increases with the prevalence of anaemia in children.

   - Hypothesis 2: The correlation between infant anaemia prevalence & infant mortality is substantially lower in areas with high levels of socioeconomic development & maternal education.

   - Implementation of Hypothesis 1:

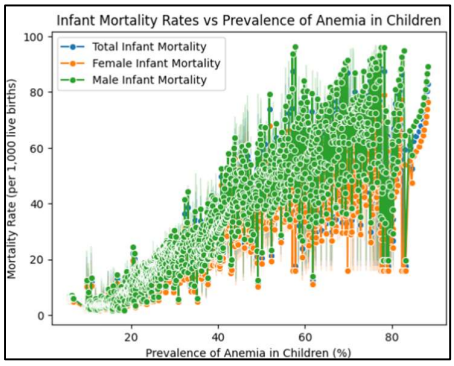
     The code considers two hypotheses by using machine learning algorithms: Logistic Regression and Support Vector Machines (SVM). In the Hypothesis 1, which says that with the increase in the prevalence of anaemia among children, the likelihood of infant mortality also increases, the model used will be a Logistic Regression. This model predicts the

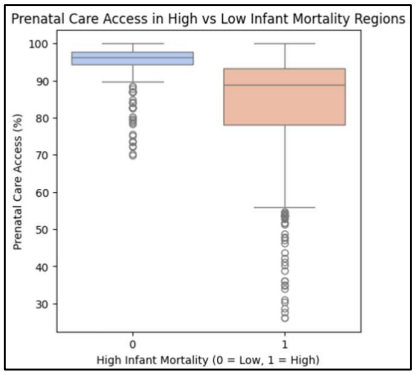
binary outcome-high or low infant mortality-based on the anaemia prevalence, male and female infant mortality rates. The Logistic Regression model will learn the relationship between features and targets during training-that is, anaemia prevalence versus mortality rates, high or low mortality-and predict mortality categories on a test set. For Hypothesis 2, which states that the correlation between anaemia and infant mortality is lower in areas with higher socioeconomic development and maternal education, a similar approach is followed where SVM uses a linear kernel to separate high and low mortality cases based on anaemia prevalence, prenatal care, literacy rates, and other health-related factors. The SVM works by finding a hyperplane that maximizes the margin between the two classes: high vs. low infant mortality. Further, both models will be trained, evaluated, and tested for accuracy, precision, recall, and F1-score.

- <u>Visualisations:</u>

```
0.9962756052141527
[[552   2]
 [  2 518]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       554
           1       1.00      1.00      1.00       520

    accuracy                           1.00      1074
   macro avg       1.00      1.00      1.00      1074
weighted avg       1.00      1.00      1.00      1074
```

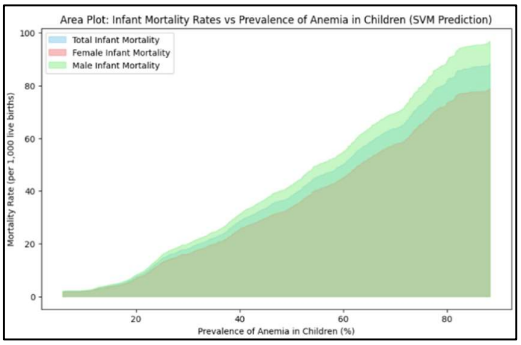
```
0.9962756052141527
[[553   1]
 [  3 517]]
              precision    recall  f1-score   support

           0       0.99      1.00      1.00       554
           1       1.00      0.99      1.00       520

    accuracy                           1.00      1074
   macro avg       1.00      1.00      1.00      1074
weighted avg       1.00      1.00      1.00      1074
```

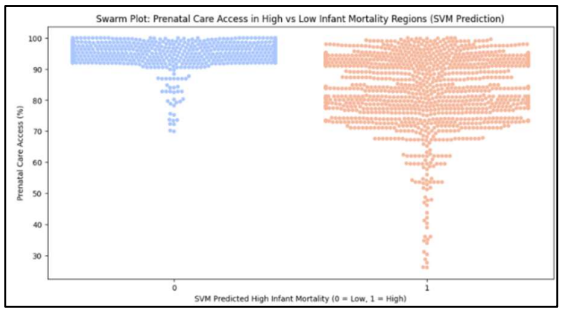





```
0.9963666852990497
[[1775    5]
 [   8 1790]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1780
           1       1.00      1.00      1.00      1798

    accuracy                           1.00      3578
   macro avg       1.00      1.00      1.00      3578
weighted avg       1.00      1.00      1.00      3578
```

```
0.995248742314142
[[1776    4]
 [  13 1785]]
              precision    recall  f1-score   support

           0       0.99      1.00      1.00      1780
           1       1.00      0.99      1.00      1798

    accuracy                           1.00      3578
   macro avg       1.00      1.00      1.00      3578
weighted avg       1.00      1.00      1.00      3578
```





- Observation: The output verifies the hypotheses, with high accuracy, precision, and recall from both the Logistic Regression and the SVM models, which asserts the relationship between anaemia prevalence and infant mortality. The strong predictive performance in Hypothesis 1 confirms the fact that the higher the prevalence of anaemia, the greater the likelihood of child mortality. The models further point out, for Hypothesis 2, the moderating role of socioeconomic factors such as access to prenatal care and maternal literacy. In areas where healthcare and education are relatively better, the mortality rates are lower despite the prevalence of anaemia. These findings confirm both hypotheses by demonstrating how anaemia and socioeconomic conditions together impact infant mortality.

2. Question 2: How does maternal hypertension correlate with low birthweight & infant mortality?

- Hypothesis 1: Increased maternal hypertension is strongly related with higher rates of low birthweight & infant mortality rates.
- Hypothesis 2: Exposure to maternal healthcare services serves as a mediating factor in the correlation between low birthweight/infant mortality & maternal hypertension.
- Implementation of Hypothesis 1:

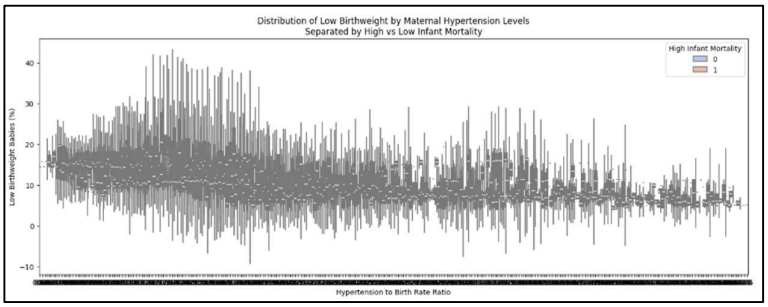
  The hypotheses are approached with the use of Logistic Regression and Support Vector Machine algorithms that estimate the association between maternal hypertension and low birthweight with infant mortality, and the mediating role of maternal healthcare services. Logistic Regression is used as a probabilistic model to estimate the likelihood of high infant mortality given certain input features such as hypertension prevalence, low birthweight rates, and healthcare factors. SVM finds the

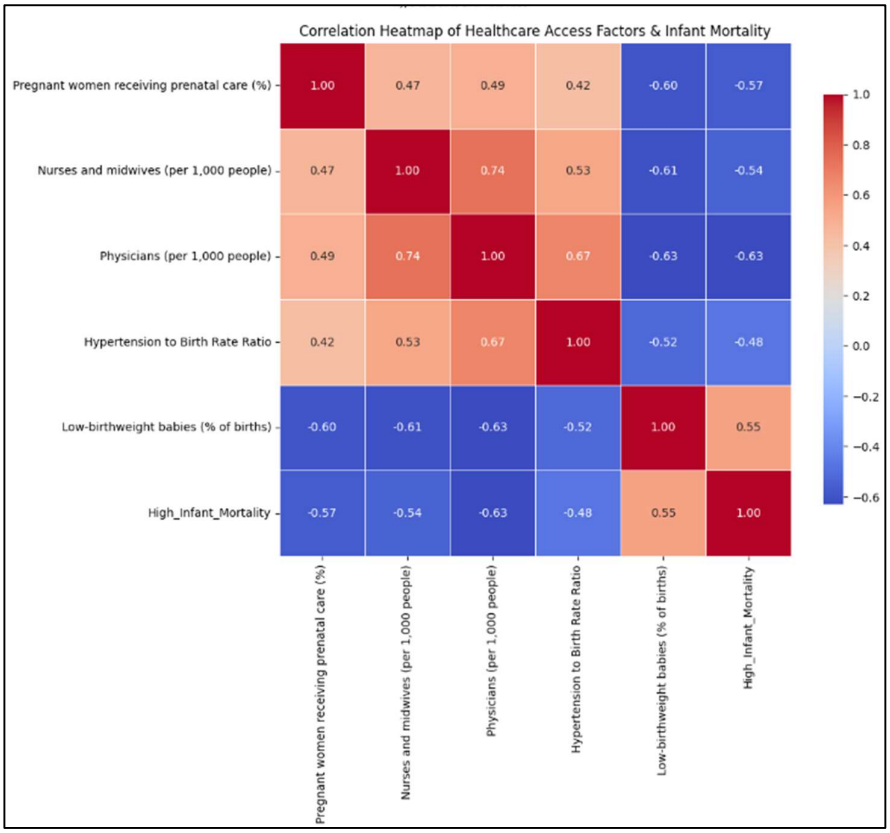
hyperplane that best separates the data in a transformed feature space to classify high and low infant mortality, using linear kernels for simplicity and interpretability. These algorithms are suitable because Logistic Regression provides direct probabilities for classification, while SVM effectively handles high-dimensional data and ensures robust separation.

- Visualisations:

```
0.8528864059590316
[[484  65]
 [ 93 432]]
              precision    recall  f1-score   support

           0       0.84      0.88      0.86       549
           1       0.87      0.82      0.85       525

    accuracy                           0.85      1074
   macro avg       0.85      0.85      0.85      1074
weighted avg       0.85      0.85      0.85      1074
```


```
0.8528864059590316
[[484  65]
 [ 93 432]]
              precision    recall  f1-score   support

           0       0.84      0.88      0.86       549
           1       0.87      0.82      0.85       525

    accuracy                           0.85      1074
   macro avg       0.85      0.85      0.85      1074
weighted avg       0.85      0.85      0.85      1074
```



- Observation: These performance metrics-accuracy above 85% for both hypotheses-serve to validate both hypotheses. In fact, the results clearly show that maternal hypertension is highly associated with higher rates of low birth weight and infant mortality, and the use of maternal health care serves to alleviate these concerns, thus acting as an intervening variable.

VI. BUILDING DATA PRODUCT:

- Purpose:

  Our mission is to arm stakeholders, communities, and policy thinkers with the ability to act on maternal and infant health through comprehensive data sets that can answer critical health disparities. The online tool is designed to allow users to explore complex interactions among maternal health, education, and health infrastructure, and their consequences for infant and maternal outcomes. We transform raw data into meaningful analysis in the quest to make informed decisions that inspire change in equitable and accessible healthcare practices around the world.

  We have analysed data on the immunization coverage, prevalence of anaemia, infrastructure in hospitals, maternal mortality, and birth weight trends. In doing so, this research will try to answer key questions such as the reduction of anaemia in pregnant women and the newborn due to maternal nutrition programs, the status of immunization relating to infant mortality, and how neonatal deaths and stillbirths are related to available resources in hospitals. It will, therefore, be through such responses that patterns are found out and disparities highlighted, hence appropriate evidence-based interventions for different regions and various challenges.

  The objective of our platform goes beyond data analysis; it goes to advocacy and empowerment. Policymakers and healthcare providers will engage through an intuitive interface in dynamic trends and predictive models that form the core of policy programs in maternal health education, targeted immunization, and infrastructure investments. As low birthweight and maternal anaemia are being addressed, so do the lives of the mothers and infants improve-the foundation for healthier communities that create a legacy of sustained good health. We can bridge the gaps in care, inspire action, and fuel progress in global maternal and infant health-together.

- Tech Stack: -

  Our project is powered by a modern and efficient tech stack aimed at seamless functionality, an intuitive user experience, and solid analytical capabilities. Each component of our tech stack is explained below.

  1. Streamlit: Streamlit is a powerful open-source framework for building interactive web applications in data science and machine learning projects.
     a. Purpose: Streamlit is the frontend of our application, which allows users to interact smoothly with visualizations, predictive models, and data filters.
     b. Characteristics:
        i. Fast prototyping and deployment of web applications based on data.
        ii. Tailor your data with custom widgets that support real-time user input-sliders, dropdowns, buttons, and more.
        iii. Straightforward integration with Python scripts, making it a developer-friendly tool for data scientists.
     c. Benefits towards Our Project:
        i. It allows users to explore datasets, run predictions, and visualize health insights without requiring any coding expertise.
        ii. The graphs and charts are of a dynamic nature; this enhances the user experience.
  2. MySQL: MySQL is a very popular relational database management system that will store, organize, and manage our data.
     a. Purpose: MySQL will be used as the back-end database, storing extensive data sets related to maternal and infant health indicators.
     b. Characteristics:
        i. Efficient storing and retrieving of data ensure high performance for big data.
        ii. SQL allows easy querying and updating and managing data.

iii. It has good security features for sensitive health data.

c. Benefits towards Our Project:

i. Handles the complexity of multi-dimensional data: regional, temporal, and categorical variables.

ii. Permits integration with Python to enable smooth data access and processing.

iii. Guarantees data integrity and reliability that are so crucial for health-related analyses.

3. Python Libraries: We made use of various Python libraries to facilitate data manipulation, machine learning, visualization, and interaction with the database. Following is the list of libraries utilized along with their respective functions:

a. Streamlit:

i. Provides the interface and application logic for deploying our platform.

ii. Enables direct, real-time user interactions with widgets for model inputs, data exploration, and filter customization.

b. Pandas:

i. Used for data manipulation and analysis.

ii. Allows cleaning, transforming, and integrating data from MySQL into an analysable format.

c. mysql.connector:

i. Establishes a direct connection between the Python application and the MySQL database.

ii. Allows for efficient querying and updating to keep the database and web application synchronized.

d. sklearn (Scikit-learn):

i. RandomForestRegressor: This is used in predictive modeling to analyze the relationship between health indicators and outcomes.

ii. train_test_split: Splits the dataset into training and testing subsets for model evaluation.

iii. mean_squared_error: This measures the accuracy of predictions by quantifying errors between observed and predicted values.

e. plotly.express:

i. A powerful library to create interactive visualizations.

ii. Develops dynamic graphs and charts to represent trends in maternal and infant health data, including mortality rates, immunization coverage, and the distribution of hospital resources

- Integration of Various tech stack:

1. Database Layer: Here is where the health datasets will reside in an orderly fashion in the database MySQL, along with an efficient access mechanism.
2. Backend Logic, Python: Fetches data from the database, applies data processing and machine learning models, and prepares insights for visualization.
3. Frontend Layer: Deploys interactive visualizations along with a prediction tool that allows widgets for data exploration using the Streamlit library.

This complete tech stack will ensure that our platform is functional, scalable, and user-friendly for many stakeholders involved in maternal and infant health research.

- Development:

Our development process of maternal and infant health analytics rests on two major factors, namely, setting up Streamlit to handle the user interface and integrating MySQL into the system for handling backend databases. These together put up a seamless system for exploration, analysis, and visualization of data.

1. Streamlit Setup:

Streamlit constitutes one of the important ingredients in our platform. It comes with a robust, easy-to-use framework for developing a web-based interactive application. As it was tailored with data science projects in mind, working with it allows for easily turning a Python script into a fully functional web interface. With Streamlit, we can provide an

interactive user-friendly front end for exploring health data, apply various filters, and display visualized insights with dynamic charts and graphs.

The application is a Python script in which we are going to define the layout, widgets, and data visualizations. Streamlit allows the developer to include real-time interaction, where a user can input parameters of a year, region, or demographic data and immediately see the results. It's very flexible, allowing for everything from complex models to interactively visualizing results, and still having a very accessible interface. This will ensure that users, whether policymakers or healthcare professionals, can draw actionable insights without technical expertise

2. Database Integration with MySQL:

   MySQL is a robust and efficient relational database management system, forming the backbone of our backend. In MySQL, these key indicators comprise birth rates, maternal mortality ratios, immunization coverage, and neonatal health metrics, making for a big dataset in a centralized repository. The database is structured to enable quick and efficient querying, hence interaction with the frontend application.

   MySQL is integrated into our platform with the help of Python's `mysql-connector`, which secures and accelerates the communication between the application and the database. This makes it possible for the platform to fetch real-time data, process, and represent it. Moreover, this database is prepared for big volumes, keeping the main concern about data integrity and security to the fore when handling health information.

   This is made possible by integrating Streamlit with MySQL so that the platform can dynamically access the dataset for querying by the user, visualizing trends, and analysing health metrics. This will ensure seamless integration in which data is consistent and reliable, forming a foundation for more advanced features like predictive modelling or comparative analysis.

Put together, Streamlit and MySQL form one harmonious ecosystem that bridges the gap from data storage to user interaction. This development process will make sure the platform remains accessible, scalable, and is

able to deliver meaningful insights for improvements in maternal and infant health outcomes.

- Creating UI:

The website is built thoughtfully around three important features: Home, Table, and Models, which ensure an intuitive and engaging experience for users. The Home section gives a quick view of what the whole project is, what it aspires to, what it solves, and what it opens to the world. In the Table section, the user explores the data interactively by creating, reading, updating, and deleting data and filtering the data to find specific trends and visualize important aspects of the dataset. Finally, the Models section goes deep into the analytic components of the project, where hypotheses and research questions are presented with data visualization and insights. It allows a structured approach from understanding the purpose of the project, through the data engagement, and into the analytical outcomes.

Each page has following components:

1. Table Page: This page is used to implement crud operations as well as filtering mechanisms.
    a. Crud operations implemented: -
        i. Lookup: Has been used to display either a full data set or filtered one. An option to filter any entry related to year, region, and category is available that reduces the search scope to needed data.
        ii. Add Entry (Create): Allows the user to add a new entry into the database. Pre-submission validation of data type, whether it is an integer, float, or string, for data integrity. Automatically inserts new data into the database.
        iii. Modify Entry (Update): Allows the user to edit existing entries by specifying unique identifiers
        iv. Remove Entry (Delete): Facilitates deletion of an already existing entry using its unique identifiers. Confirms whether data exists before removal from the database.
    b. Filtering mechanisms implemented:
        i. Region Based filter: The filter is used to retrieve data based on specific countries.

ii. Year based filter: The filter is used to retrieve data based on particular year.

iii. Health resources-based filter: This displays information surrounding the following data:

- Births attended by skilled health staff
- Hospital beds per 1000 people
- BCG HepB3 Immunization
- Pol3 Immunization
- Newborns Protected Against Tetanus
- Nurses and midwives per 1000 people
- Physicians per 1000 people
- Pregnant women receiving prenatal care
- Vitamin A supplementation coverage

iv. Maternal Data based filter: This displays information surrounding the following data:

- Life expectancy at birth female years
- Literacy rate Pregnant Women
- Maternal mortality ratio
- Pregnant women receiving prenatal care percent
- Prevalence of anemia among pregnant women percent
- Prevalence of current tobacco use pregnant women
- Prevalence of hypertension pregnant women

v. Infant health-based filter: This displays information surrounding the following data:

- Birth rate crude per 1000 people
- Births attended by skilled health staff percent of total
- Mortality rate infant per 1000 live births
- Mortality rate neonatal per 1000 live births

- Number of infant deaths
- Stillbirth rate per 1000 total births
- Neonatal Mortality Rate to Birth Rate Ratio
- Vitamin A supplementation coverage rate

c. Error Handling: The application provides for robust error handling in maintaining data integrity and not crashing on a bug encounter:

2. Model Page: The Model Page of the UI is structured to facilitate hypothesis-driven analysis, allowing users to select specific columns from a dataset dynamically and visualize the results using machine learning models and 3D visualizations.

   a. Analysis: The Analysis section is the backbone of the Model page, where the user explores various hypotheses related to a selected question. Each hypothesis is tied to specific columns in the dataset, which are dynamically chosen by the user.

      The key elements of this process include:

      i. Question Selection: Users select a question from predefined options. There are 4 questions for the user to choose from.

      ii. Hypothesis Selection: For each question, there are two hypotheses to explore.

      iii. Column Selection: Users must choose at least three columns related to the hypothesis to proceed with the analysis.

      iv. Machine Learning Model Execution: Once the required columns are selected, the data is processed, and features are separated from the target variable. A machine learning model is trained and tested on the data to generate insights. After the training, the model then predicts outcomes and evaluates its performance using other metrics.

      v. Key Insights: Observations and conclusions are drawn based on model outputs

   b. Visualization: The visualization section is where the selected data and analysis come to life through detailed visualizations, particularly using advanced 3D plots. The plots displayed are dynamically linked

to the user's selected columns, making the visualizations tailored and actionable.

- Questions and hypothesis implemented: -

1. Question 1: How do healthcare system factors, such as the number of healthcare professionals, hospital bed capacity, and literacy among healthcare workers, influence maternal and general mortality rates?
    a. Hypothesis 1: Increased healthcare professionals linked to lower maternal mortality rates.
    b. Hypothesis 2: Hospital bed capacity linked to mortality rates.
    c. Literacy among healthcare workers affects outcomes.
2. Question 2: How do factors such as skilled health staff attendance during births and maternal health conditions (e.g., anaemia, hypertension, or tobacco use) influence neonatal and infant mortality rates?
    a. Hypothesis 1: Regions with higher percentages of births attended by skilled health staff will have significantly lower neonatal and infant mortality rates.
    b. Hypothesis 2: Mothers with higher prevalence of anaemia, hypertension, or tobacco use are more likely to experience higher rates of neonatal mortality and stillbirths.
3. Question 3: How does the combination of immunization coverage (BCG, HepB3, Pol3), literacy rates, and low birthweight contribute to variations in neonatal and infant health outcomes (such as neonatal mortality and stillbirth rates) across different regions?
    a. Hypothesis 1: Higher immunization coverage (BCG, HepB3, Pol3) is associated with lower neonatal mortality and stillbirth rates in regions with higher literacy rates, suggesting that better health education and access to vaccines contribute to improved neonatal health outcomes.
    b. Hypothesis 2: Regions with lower prevalence of low birthweight babies will exhibit lower neonatal mortality and stillbirth rates, indicating that maternal nutrition and healthcare interventions to

prevent low birthweight play a critical role in improving infant health outcomes.

4. Question 4: How do crude birth and death rates, life expectancy, and low-birthweight rates correlate with infant mortality and stillbirth rates across different regions and years?

    a. Hypothesis 1: Birth Rate Dynamics and Mortality Regions with extremely high or low crude birth rates show higher infant mortality and stillbirth rates, indicating population stress or resource constraints.

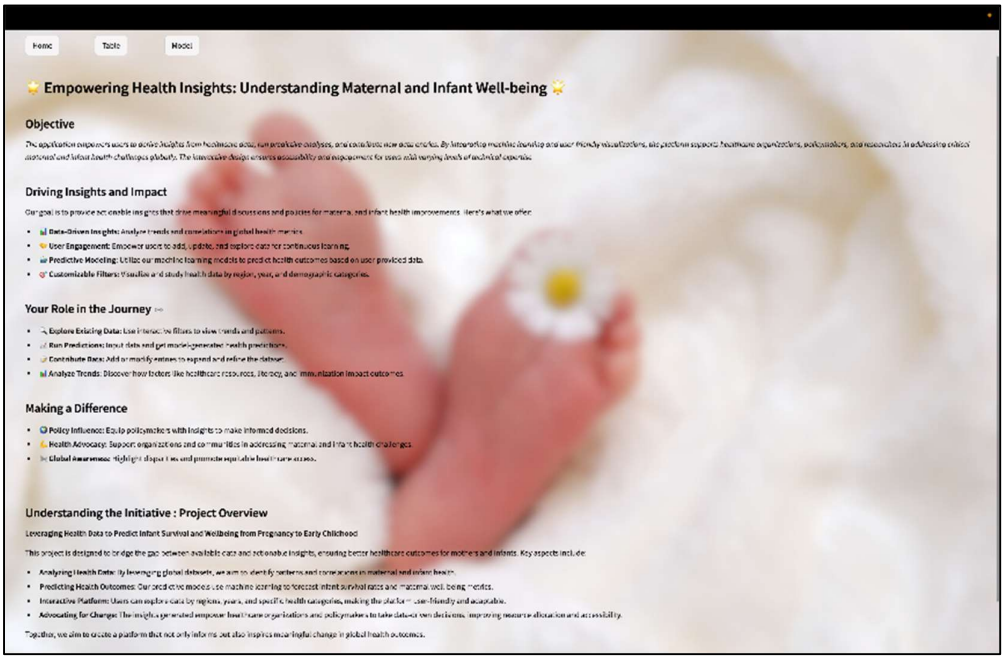
    b. Hypothesis 2: Yearly Trends: Over time, regions show a decline in infant mortality rates and neonatal mortality-to-birth rate ratios, correlating with improvements in life expectancy and a reduction in low-birthweight babies.

- Models implemented: -

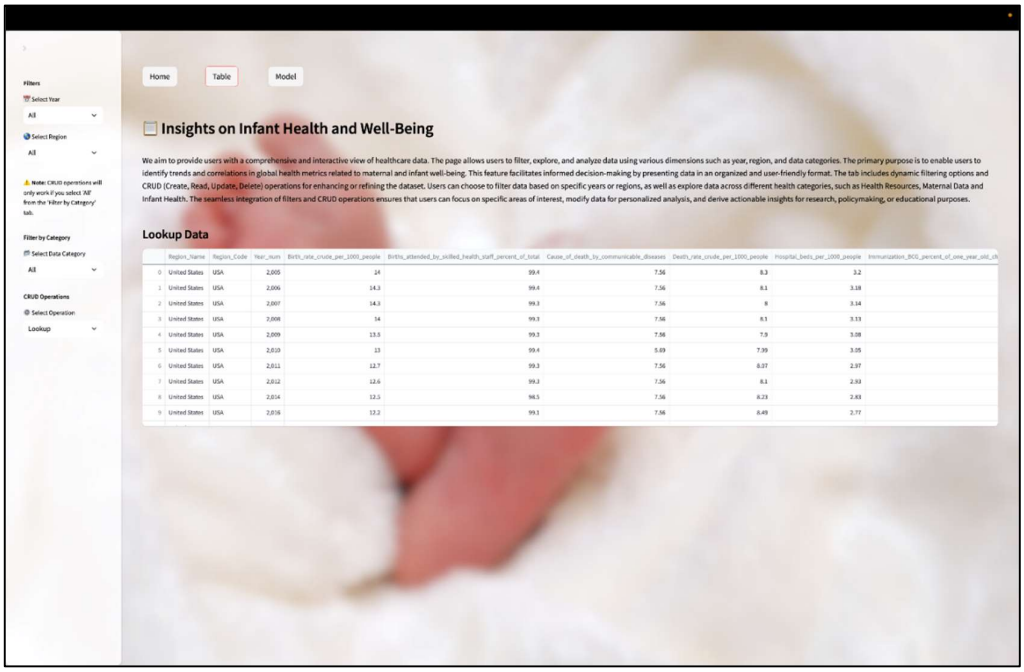
1. Support Vector Machine
2. Decision Tree Algorithm
3. Polynomial Regression
4. LightGBM Algorithm
5. XG Boost
6. Random Forest Regressor

VII. TESTING:

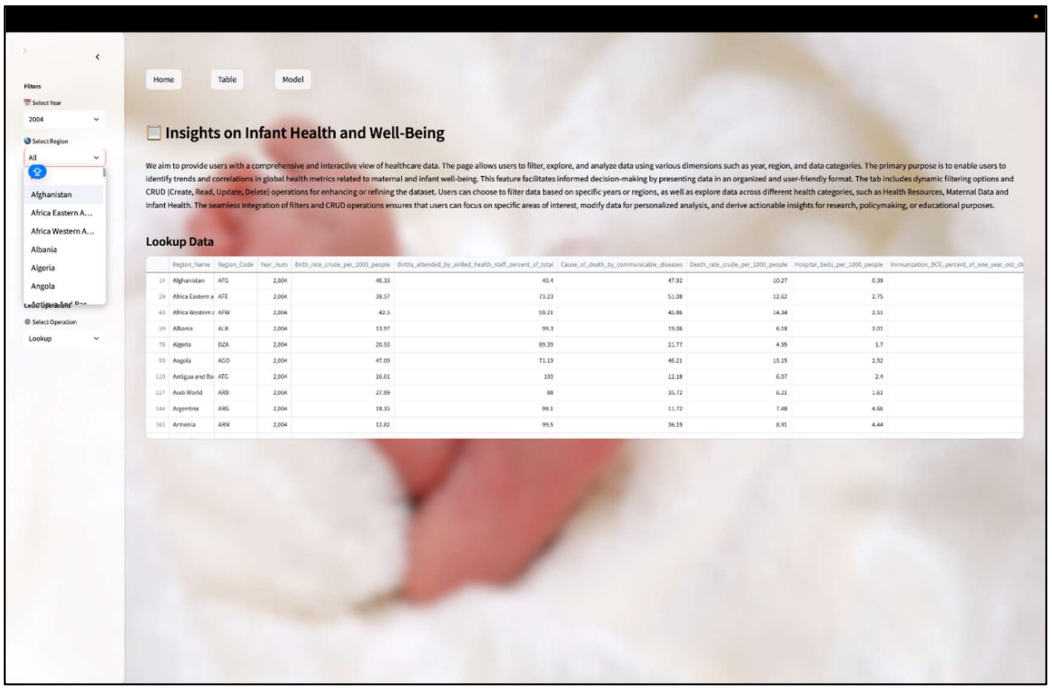
- User Interface:

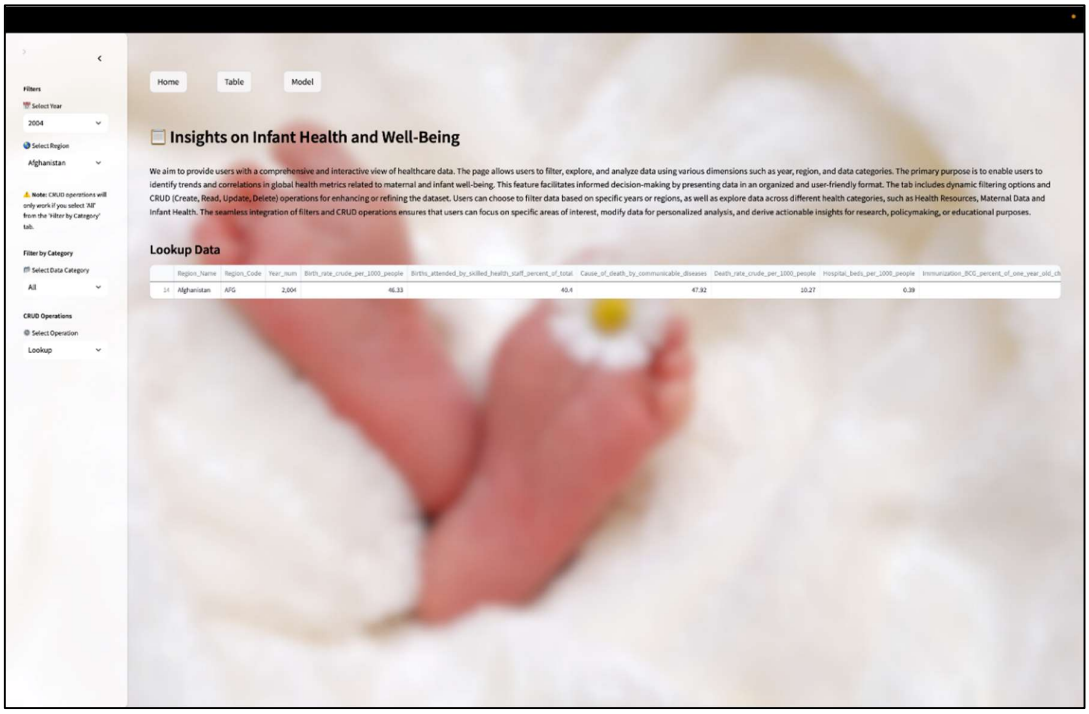


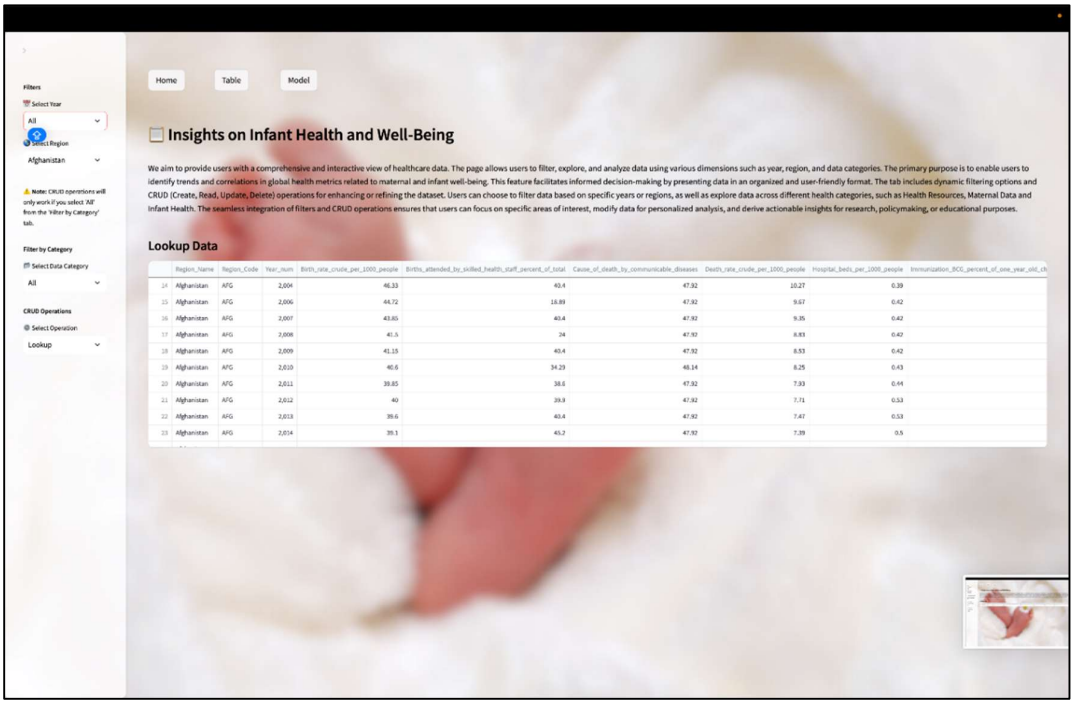
- Table Page:



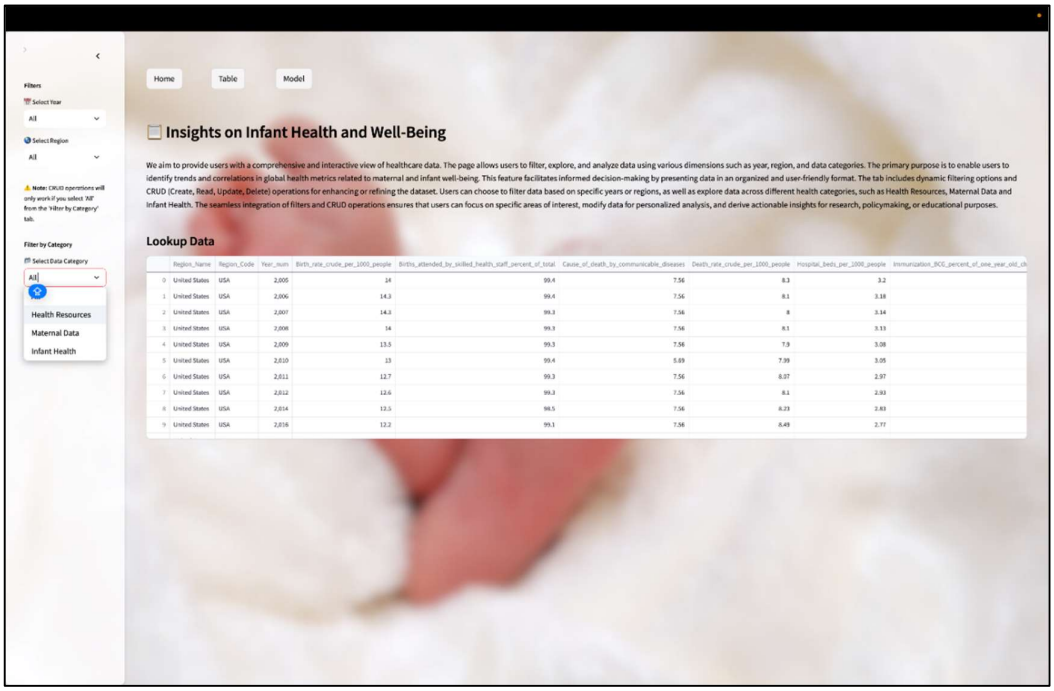
- Region and Year based filtering:

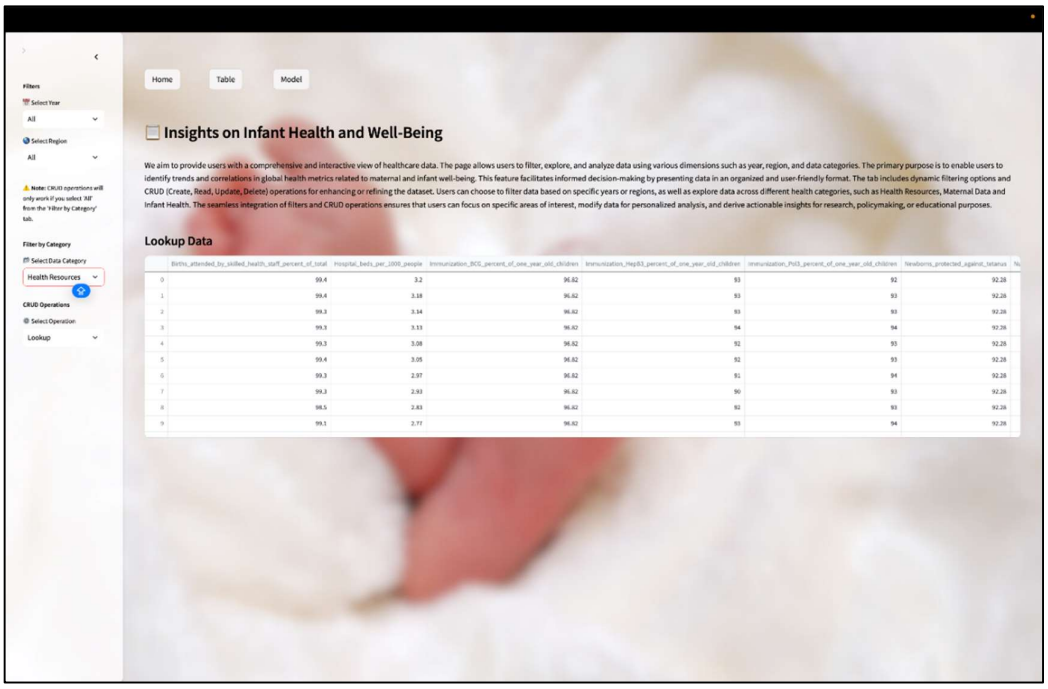


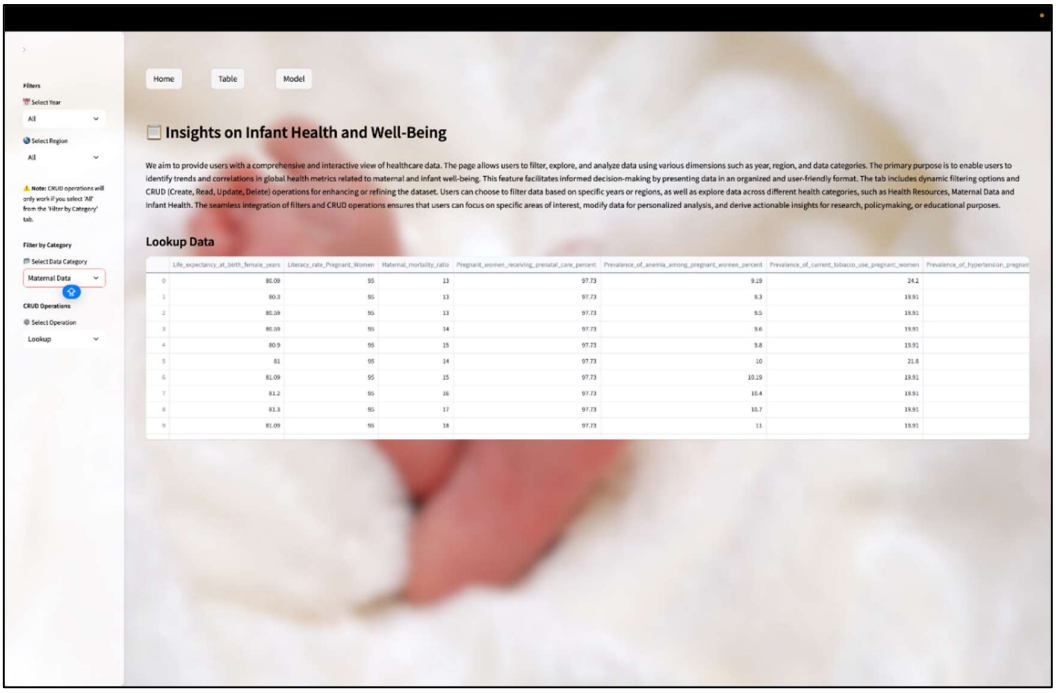


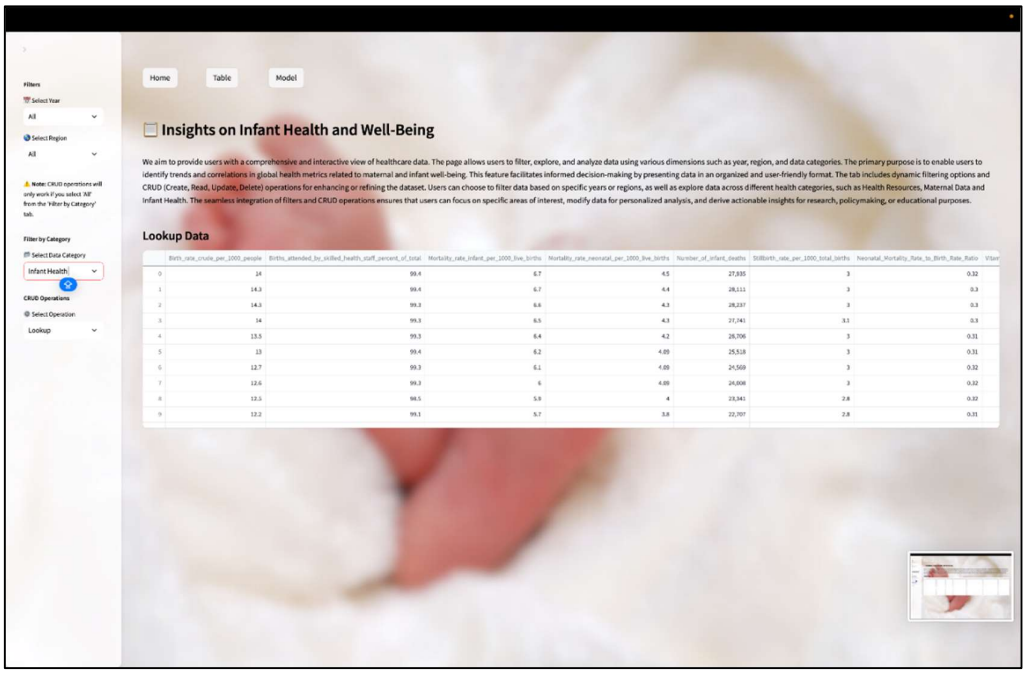


- <u>Category based filtering:</u>



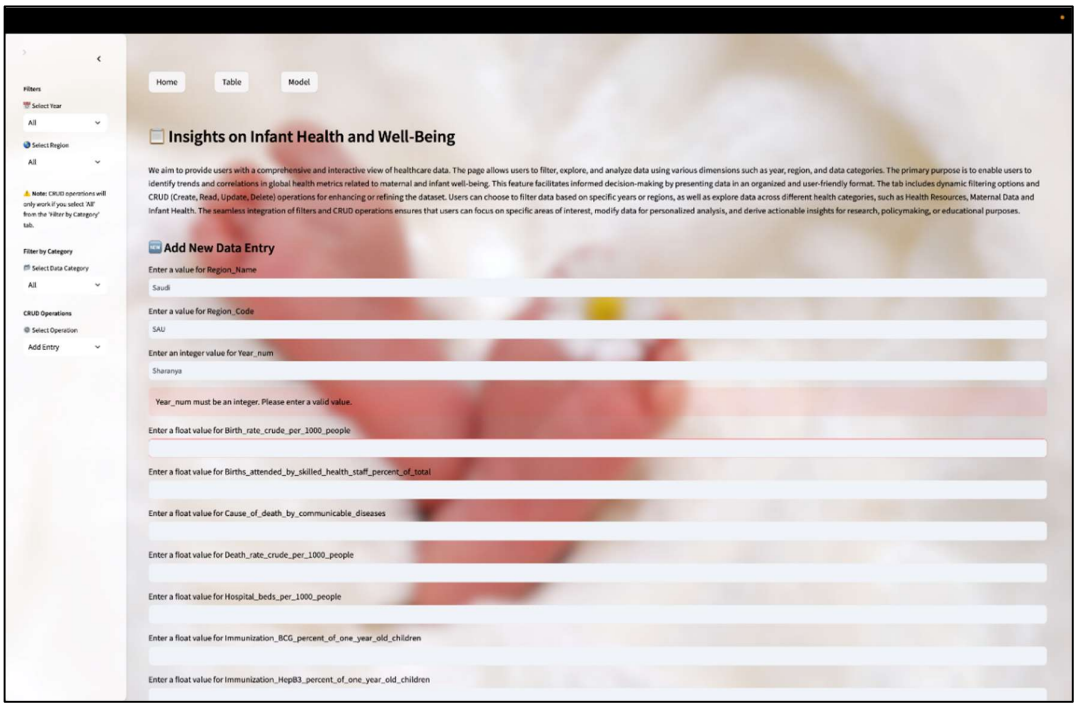

Filters
Select Year
All
Select Region
All
Note: CRUD operations will only work if you select 'All' from the 'Filter by Category' tab.
Filter by Category
Select Data Category
Health Resources
CRUD Operations
Select Operation
Lookup
Home
Table
Model
Insights on Infant Health and Well-Being
We aim to provide users with a comprehensive and interactive view of healthcare data. The page allows users to filter, explore, and analyze data using various dimensions such as year, region, and data categories. The primary purpose is to enable users to identify trends and correlations in global health metrics related to maternal and infant well-being. This feature facilitates informed decision-making by presenting data in an organized and user-friendly format. The tab includes dynamic filtering options and CRUD (Create, Read, Update, Delete) operations for enhancing or refining the dataset. Users can choose to filter data based on specific years or regions, as well as explore data across different health categories, such as Health Resources, Maternal Data and Infant Health. The seamless integration of filters and CRUD operations ensures that users can focus on specific areas of interest, modify data for personalized analysis, and derive actionable insights for research, policymaking, or educational purposes.
Lookup Data
Births_attended_by_skilled_health_staff_percent_of_total
Hospital_beds_per_1000_people
Immunization_BCG_percent_of_one_year_old_children
Immunization_Hep03_percent_of_one_year_old_children
Immunization_Pol3_percent_of_one_year_old_children
Newborns_protected_against_tetanus
0 99.4 3.2 96.82 93 92 92.28
1 99.4 3.18 96.82 93 93 92.28
2 99.3 3.14 96.82 93 93 92.28
3 99.3 3.13 96.82 94 94 92.28
4 99.3 3.08 96.82 92 93 92.28
5 99.4 3.05 96.82 92 93 92.28
6 99.3 2.97 96.82 91 94 92.28
7 99.3 2.93 96.82 90 93 92.28
8 98.5 2.83 96.82 92 93 92.28
9 99.1 2.77 96.82 93 94 92.28

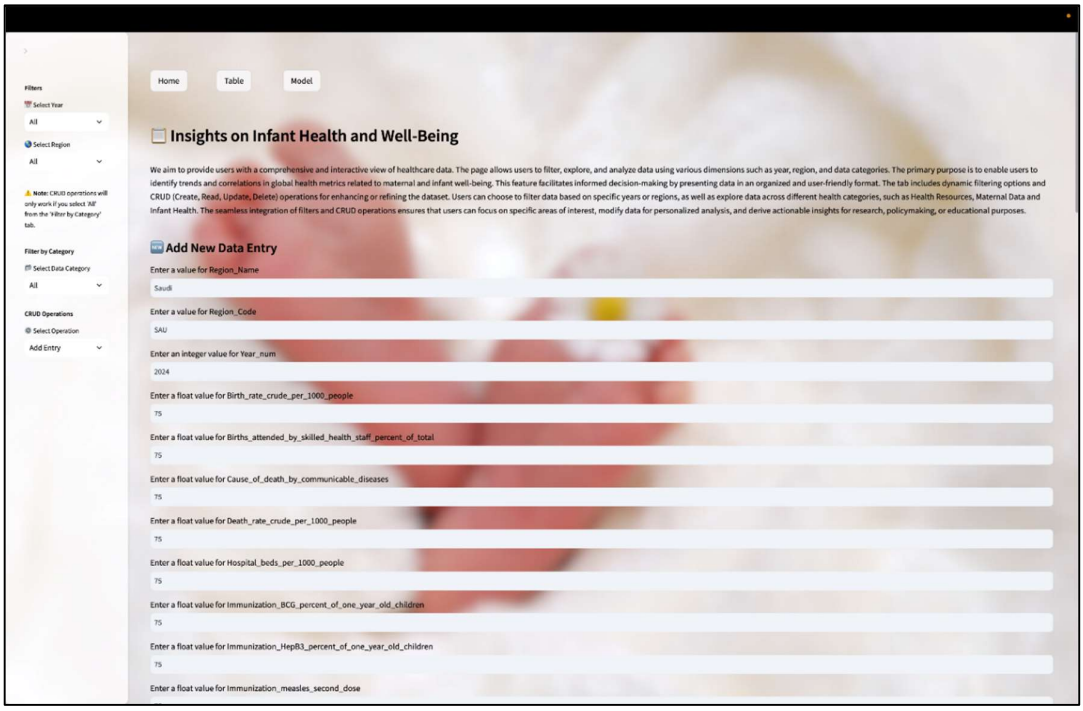

Filters
Select Year
All
Select Region
All
Note: CRUD operations will only work if you select 'All' from the 'Filter by Category' tab.
Filter by Category
Select Data Category
Maternal Data
CRUD Operations
Select Operation
Lookup
Home
Table
Model
Insights on Infant Health and Well-Being
We aim to provide users with a comprehensive and interactive view of healthcare data. The page allows users to filter, explore, and analyze data using various dimensions such as year, region, and data categories. The primary purpose is to enable users to identify trends and correlations in global health metrics related to maternal and infant well-being. This feature facilitates informed decision-making by presenting data in an organized and user-friendly format. The tab includes dynamic filtering options and CRUD (Create, Read, Update, Delete) operations for enhancing or refining the dataset. Users can choose to filter data based on specific years or regions, as well as explore data across different health categories, such as Health Resources, Maternal Data and Infant Health. The seamless integration of filters and CRUD operations ensures that users can focus on specific areas of interest, modify data for personalized analysis, and derive actionable insights for research, policymaking, or educational purposes.
Lookup Data
Life_expectancy_at_birth_female_years
Literacy_rate_Pregnant_Women
Maternal_mortality_ratio
Pregnant_women_receiving_prenatal_care_percent
Prevalence_of_anemia_among_pregnant_women_percent
Prevalence_of_current_tobacco_use_pregnant_women
Prevalence_of_hypertension_pregnan
0 80.09 95 13 97.73 9.19 24.2
1 80.3 95 13 97.73 9.3 19.91
2 80.38 95 13 97.73 9.5 19.91
3 80.39 95 14 97.73 9.6 19.91
4 80.9 95 15 97.73 9.8 19.91
5 81 95 14 97.73 10 21.8
6 81.09 95 15 97.73 10.19 19.91
7 81.2 95 16 97.73 10.4 19.91
8 81.3 95 17 97.73 10.7 19.91
9 81.09 95 18 97.73 11 19.91

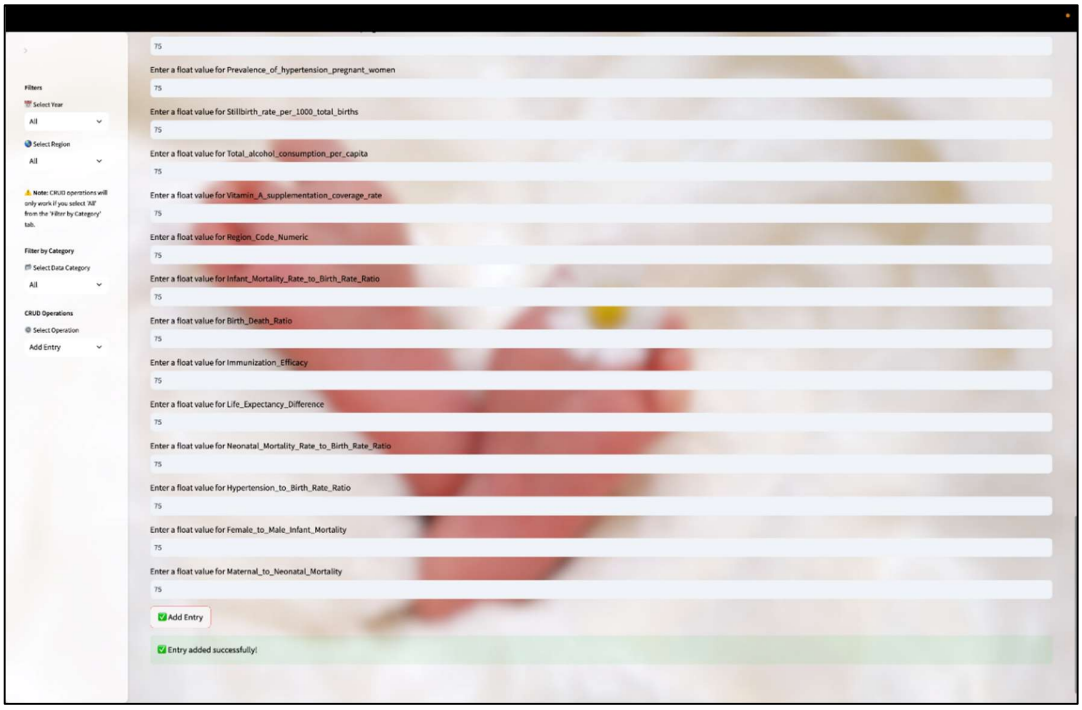


- <u>Crud Operations:</u>

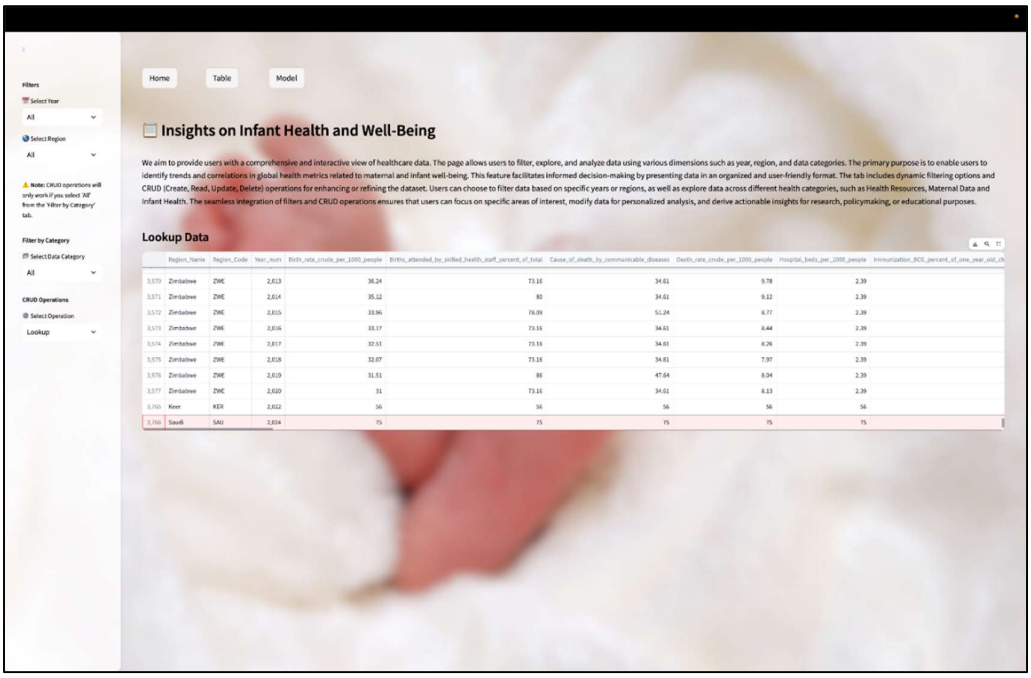
1. Adding data: The fields are bound to data type error handling mechanism when the user enters the data.

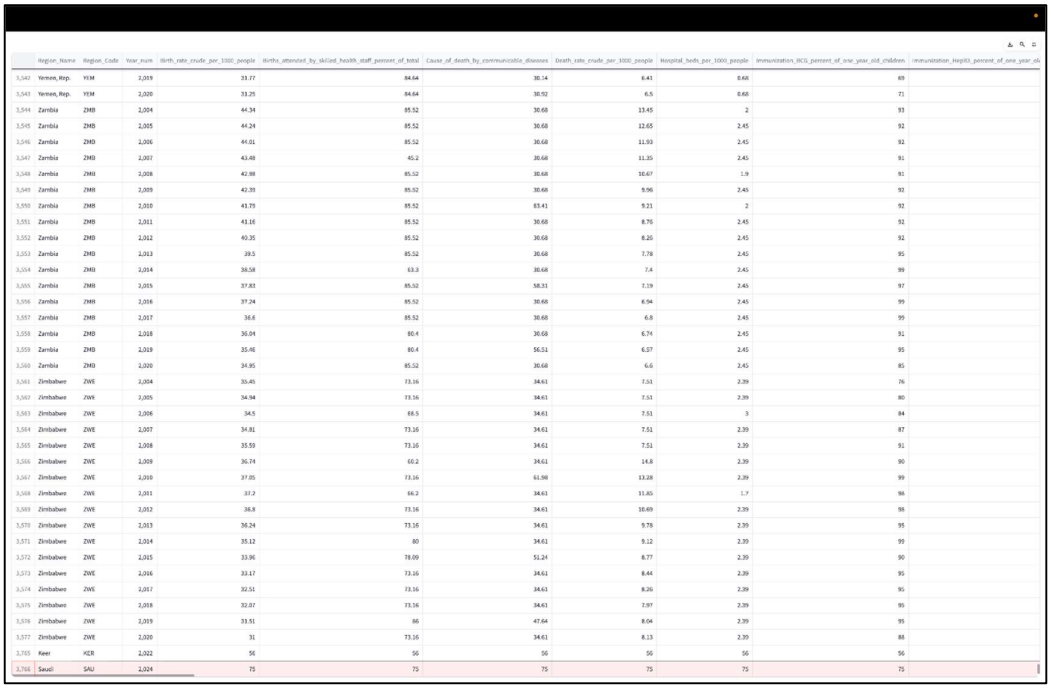



Filters
Select Year
All
Select Region
All
Note: CRUD operations will only work if you select 'All' from the 'Filter by Category' tab.
Filter by Category
Select Data Category
All
CRUD Operations
Select Operation
Add Entry
Home
Table
Model
Insights on Infant Health and Well-Being
We aim to provide users with a comprehensive and interactive view of healthcare data. The page allows users to filter, explore, and analyze data using various dimensions such as year, region, and data categories. The primary purpose is to enable users to identify trends and correlations in global health metrics related to maternal and infant well-being. This feature facilitates informed decision-making by presenting data in an organized and user-friendly format. The tab includes dynamic filtering options and CRUD (Create, Read, Update, Delete) operations for enhancing or refining the dataset. Users can choose to filter data based on specific years or regions, as well as explore data across different health categories, such as Health Resources, Maternal Data and Infant Health. The seamless integration of filters and CRUD operations ensures that users can focus on specific areas of interest, modify data for personalized analysis, and derive actionable insights for research, policymaking, or educational purposes.
Add New Data Entry
Enter a value for Region_Name
Saudi
Enter a value for Region_Code
SAU
Enter an integer value for Year_num
2024
Enter a float value for Birth_rate_crude_per_1000_people
Enter a float value for Births_attended_by_skilled_health_staff_percent_of_total
Enter a float value for Cause_of_death_by_communicable_diseases
Enter a float value for Death_rate_crude_per_1000_people
Enter a float value for Hospital_beds_per_1000_people
Enter a float value for Immunization_BCG_percent_of_one_year_old_children
Enter a float value for Immunization_HepB3_percent_of_one_year_old_children
Enter a float value for Immunization_measles_second_dose

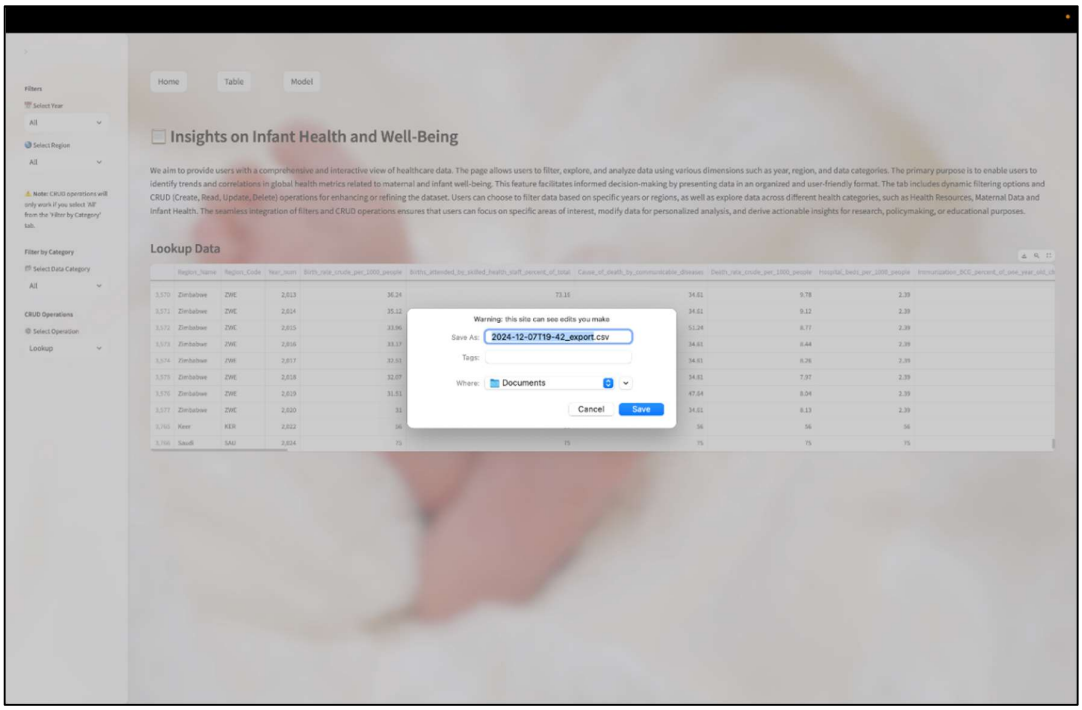

Filters
Select Year
All
Select Region
All
Note: CRUD operations will only work if you select 'All' from the 'Filter by Category' tab.
Filter by Category
Select Data Category
All
CRUD Operations
Select Operation
Add Entry
Enter a float value for Prevalence_of_hypertension_pregnant_women
Enter a float value for Stillbirth_rate_per_1000_total_births
Enter a float value for Total_alcohol_consumption_per_capita
Enter a float value for Vitamin_A_supplementation_coverage_rate
Enter a float value for Region_Code_Numeric
Enter a float value for Infant_Mortality_Rate_to_Birth_Rate_Ratio
Enter a float value for Birth_Death_Ratio
Enter a float value for Immunization_Efficacy
Enter a float value for Life_Expectancy_Difference
Enter a float value for Neonatal_Mortality_Rate_to_Birth_Rate_Ratio
Enter a float value for Hypertension_to_Birth_Rate_Ratio
Enter a float value for Female_to_Male_Infant_Mortality
Enter a float value for Maternal_to_Neonatal_Mortality
Add Entry
Entry added successfully!

- Data successfully added and displayed in Database:

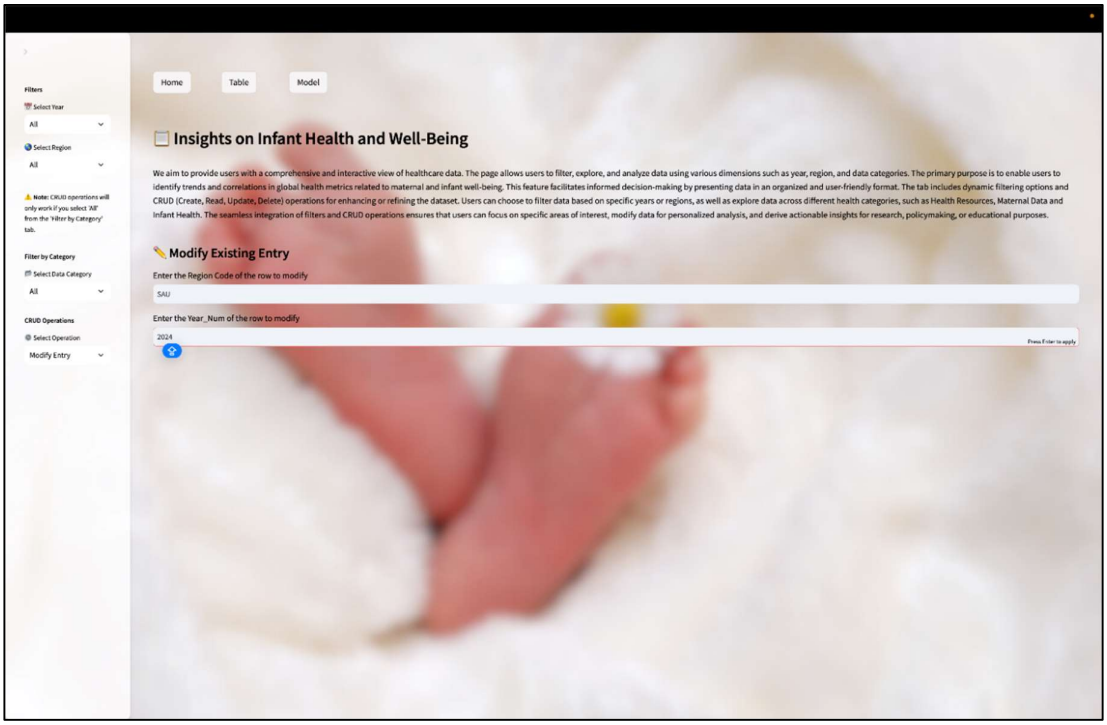




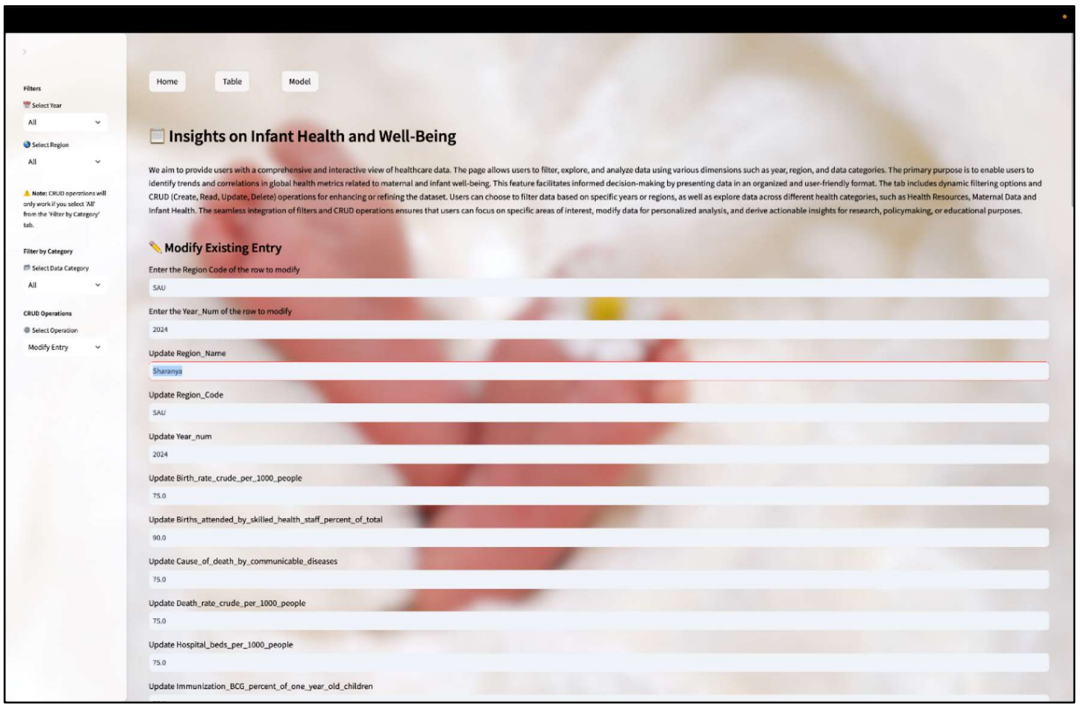
2. Downloading the filtered data as a csv:

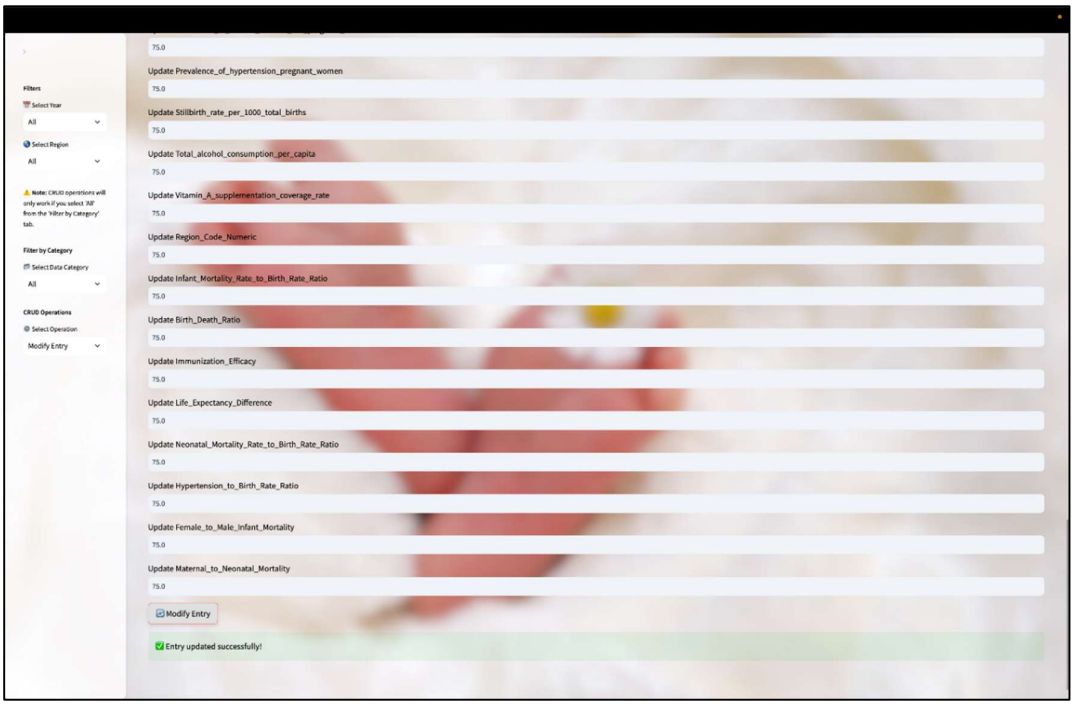


3. Modifying data: We first fetch the data using the primary key, which is a combination of the region code and the year number.

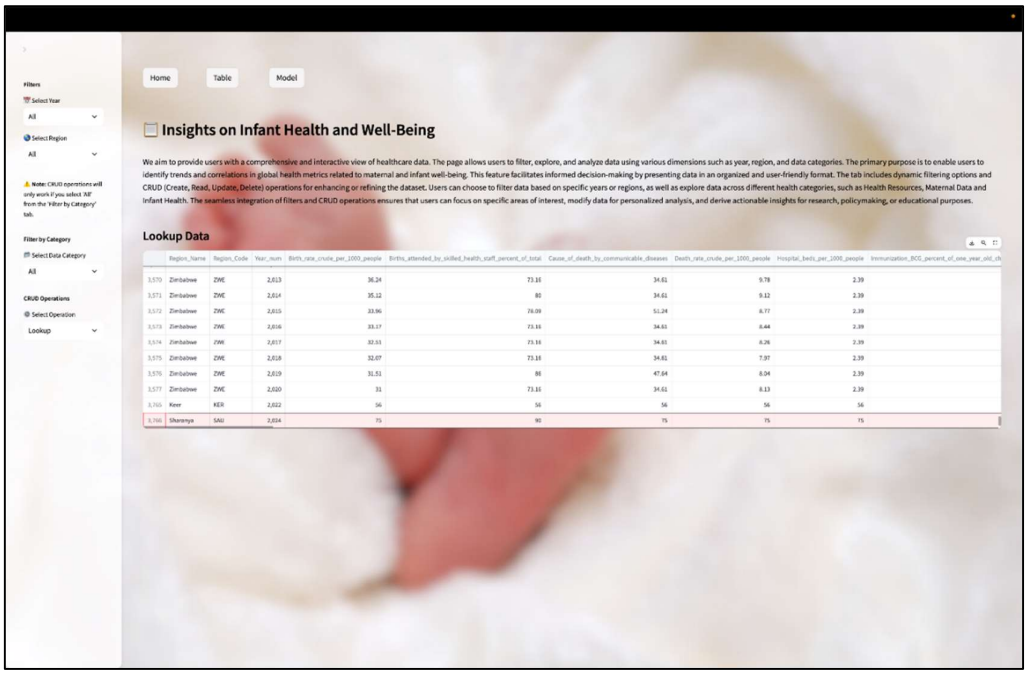


- After data is retrieved, the data can now be modified

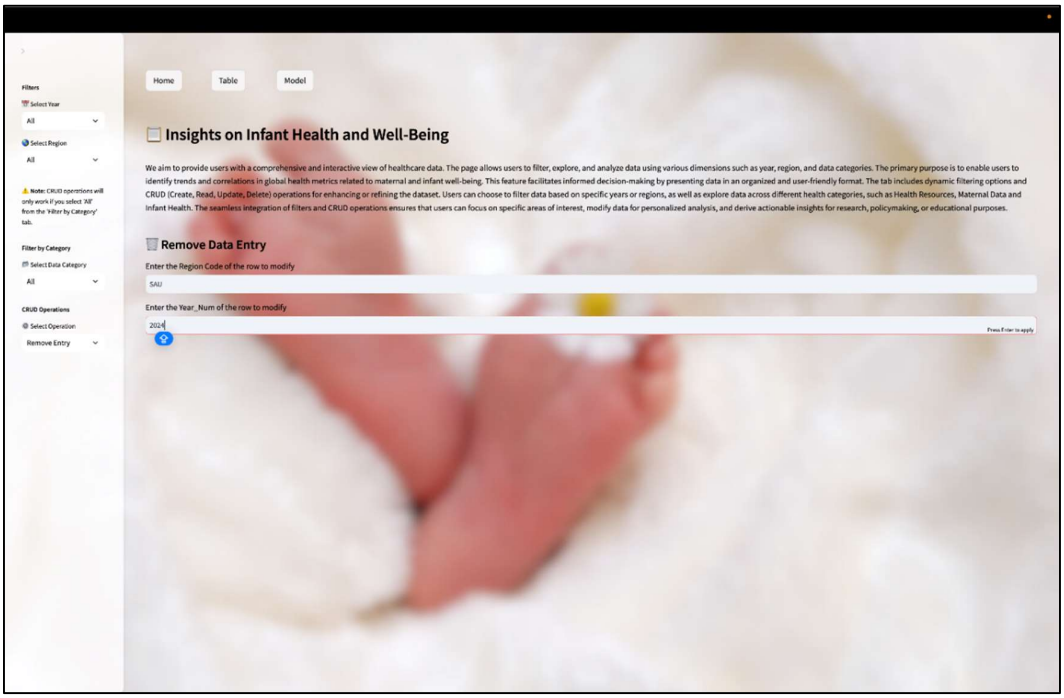


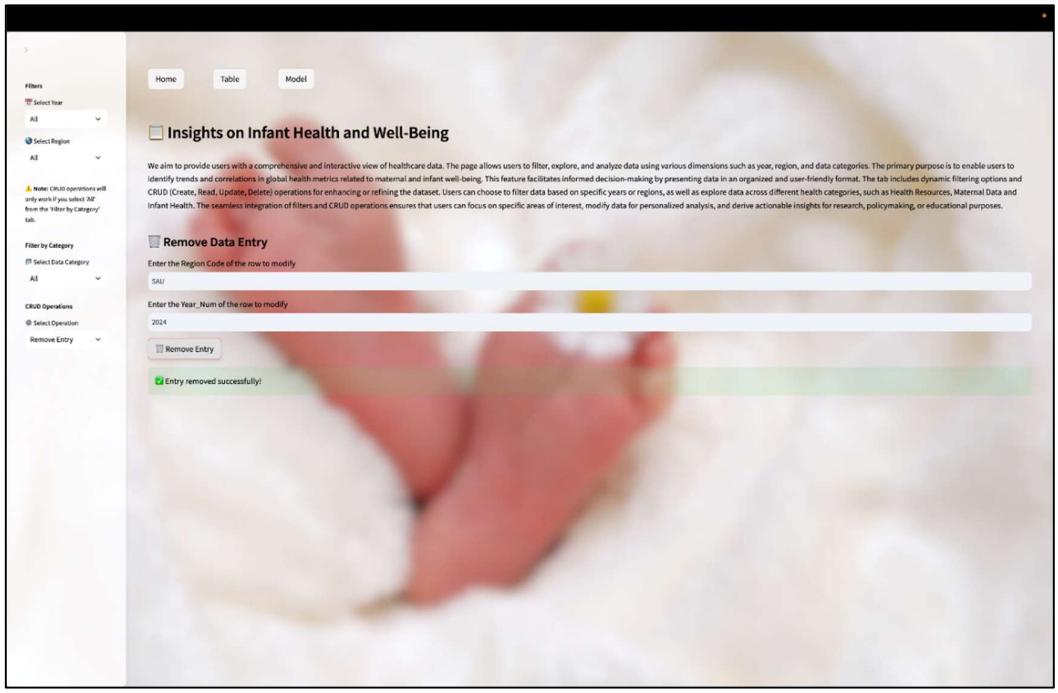


- Modified data is now changed, and entry is updated:

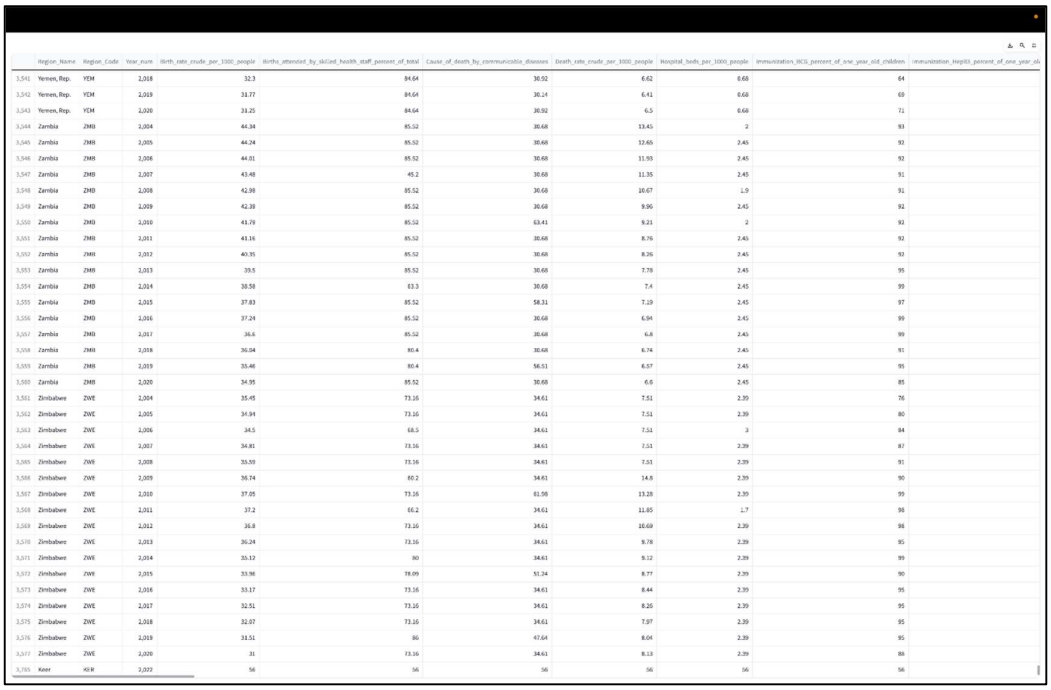


4. <u>Remove data entries:</u> We fetch the data using primary key, and the retrieved data can then be deleted.



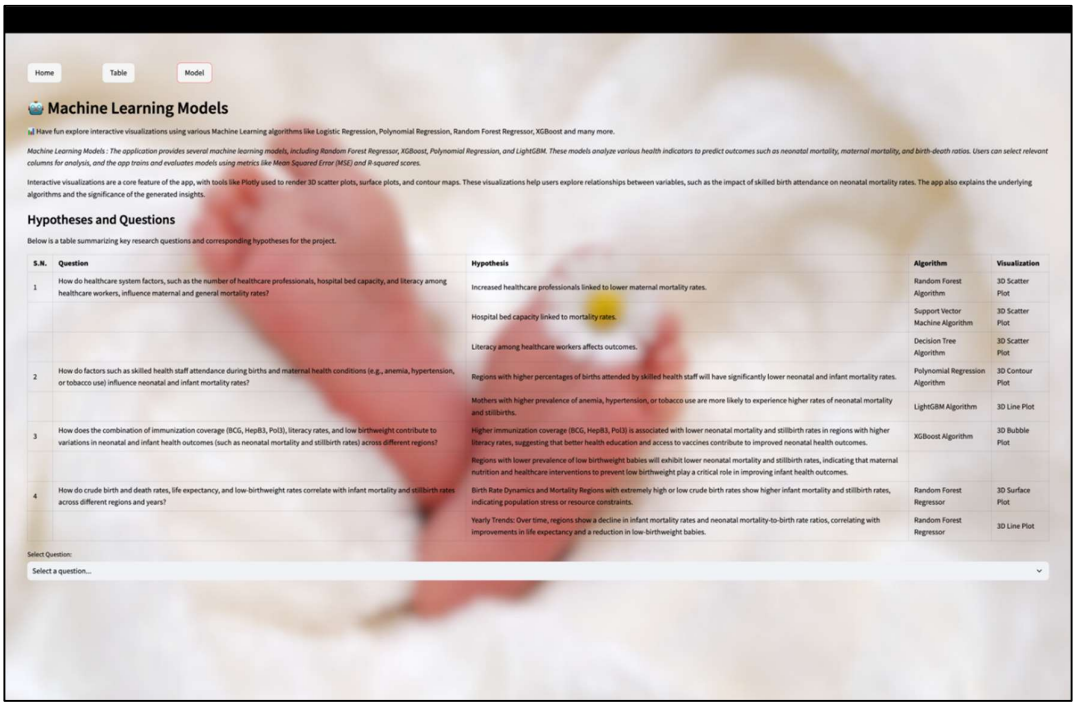


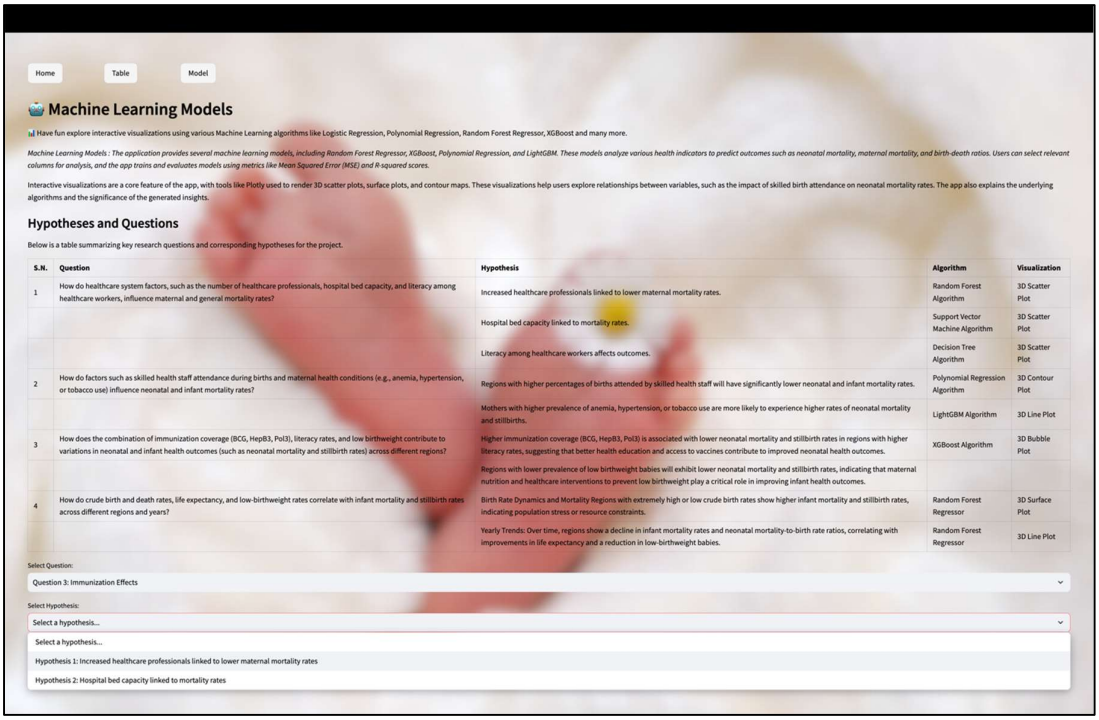
- Deleted data no longer exists in the Database:

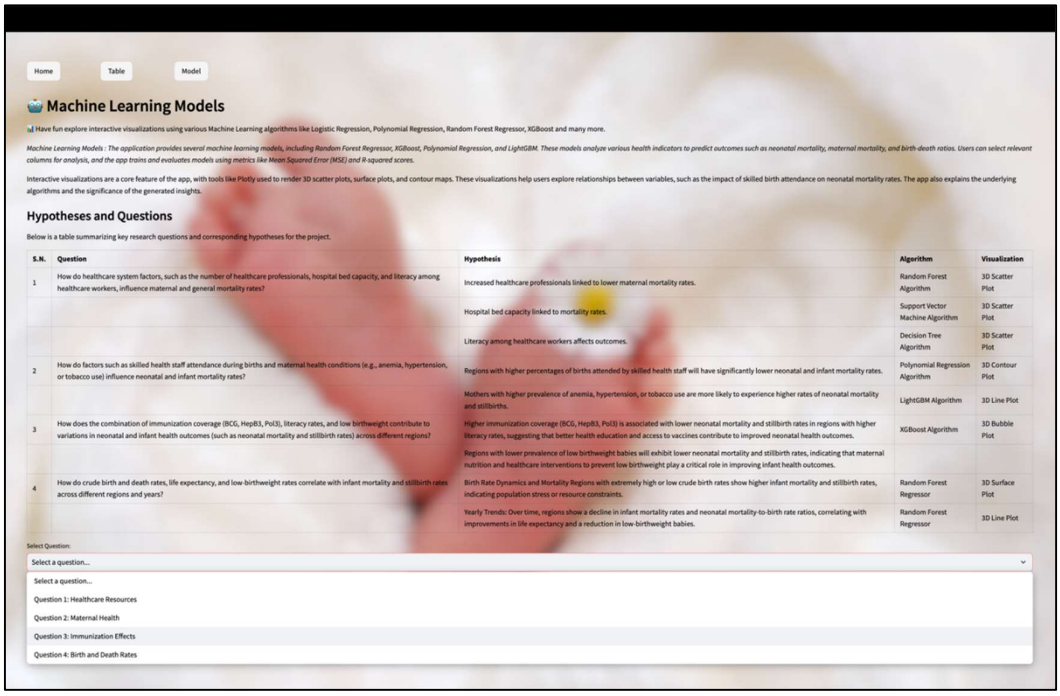


	Region_Name	Region_Code	Year_num	Birth_rate_crude_per_1000_people	Births_attended_by_skilled_health_staff_percent_of_total	Cause_of_death_by_communicable_diseases	Death_rate_crude_per_1000_people	Hospital_beds_per_1000_people	Immunization_BCG_percent_of_one_year_old_children	Immunization_Hepb3_percent_of_one_year_ol
3,541	Yemen, Rep.	YEM	2,018	32.3	84.64	30.92	6.62	0.68	64	
3,542	Yemen, Rep.	YEM	2,019	31.77	84.64	30.14	6.41	0.68	69	
3,543	Yemen, Rep.	YEM	2,020	31.25	84.64	30.92	6.5	0.68	71	
3,544	Zambia	ZMB	2,004	44.34	85.52	30.68	13.45	2	93	
3,545	Zambia	ZMB	2,005	44.24	85.52	30.68	12.65	2.45	92	
3,546	Zambia	ZMB	2,006	44.01	85.52	30.68	11.93	2.45	92	
3,547	Zambia	ZMB	2,007	43.48	45.2	30.68	11.35	2.45	91	
3,548	Zambia	ZMB	2,008	42.98	85.52	30.68	10.67	1.9	91	
3,549	Zambia	ZMB	2,009	42.39	85.52	30.68	9.96	2.45	92	
3,550	Zambia	ZMB	2,010	41.79	85.52	63.41	9.21	2	92	
3,551	Zambia	ZMB	2,011	41.16	85.52	30.68	8.76	2.45	92	
3,552	Zambia	ZMB	2,012	40.35	85.52	30.68	8.26	2.45	93	
3,553	Zambia	ZMB	2,013	39.5	85.52	30.68	7.78	2.45	95	
3,554	Zambia	ZMB	2,014	38.58	83.3	30.68	7.4	2.45	99	
3,555	Zambia	ZMB	2,015	37.83	85.52	58.31	7.19	2.45	97	
3,556	Zambia	ZMB	2,016	37.24	85.52	30.68	6.94	2.45	99	
3,557	Zambia	ZMB	2,017	36.6	85.52	30.68	6.8	2.45	99	
3,558	Zambia	ZMB	2,018	36.04	80.4	30.68	6.74	2.45	91	
3,559	Zambia	ZMB	2,019	35.46	80.4	56.51	6.57	2.45	95	
3,560	Zambia	ZMB	2,020	34.95	85.52	30.68	6.6	2.45	85	
3,561	Zimbabwe	ZWE	2,004	35.45	73.16	34.61	7.51	2.39	76	
3,562	Zimbabwe	ZWE	2,005	34.94	73.16	34.61	7.51	2.39	80	
3,563	Zimbabwe	ZWE	2,006	34.5	68.5	34.61	7.51	3	84	
3,564	Zimbabwe	ZWE	2,007	34.81	73.16	34.61	7.51	2.39	87	
3,565	Zimbabwe	ZWE	2,008	35.59	73.16	34.61	7.51	2.39	91	
3,566	Zimbabwe	ZWE	2,009	36.74	80.2	34.61	14.8	2.39	90	
3,567	Zimbabwe	ZWE	2,010	37.05	73.16	61.98	13.28	2.39	99	
3,568	Zimbabwe	ZWE	2,011	37.2	66.2	34.61	11.85	1.7	98	
3,569	Zimbabwe	ZWE	2,012	36.8	73.16	34.61	10.69	2.39	98	
3,570	Zimbabwe	ZWE	2,013	36.24	73.16	34.61	9.78	2.39	95	
3,571	Zimbabwe	ZWE	2,014	35.12	80	34.61	9.12	2.39	99	
3,572	Zimbabwe	ZWE	2,015	33.96	78.09	51.24	8.77	2.39	90	
3,573	Zimbabwe	ZWE	2,016	33.17	73.16	34.61	8.44	2.39	95	
3,574	Zimbabwe	ZWE	2,017	32.51	73.16	34.61	8.25	2.39	95	
3,575	Zimbabwe	ZWE	2,018	32.07	73.16	34.61	7.97	2.39	95	
3,576	Zimbabwe	ZWE	2,019	31.51	86	47.64	8.04	2.39	95	
3,577	Zimbabwe	ZWE	2,020	31	73.16	34.61	8.13	2.39	88	
3,785	Keer	KER	2,022	56	56	56	56	56	56	

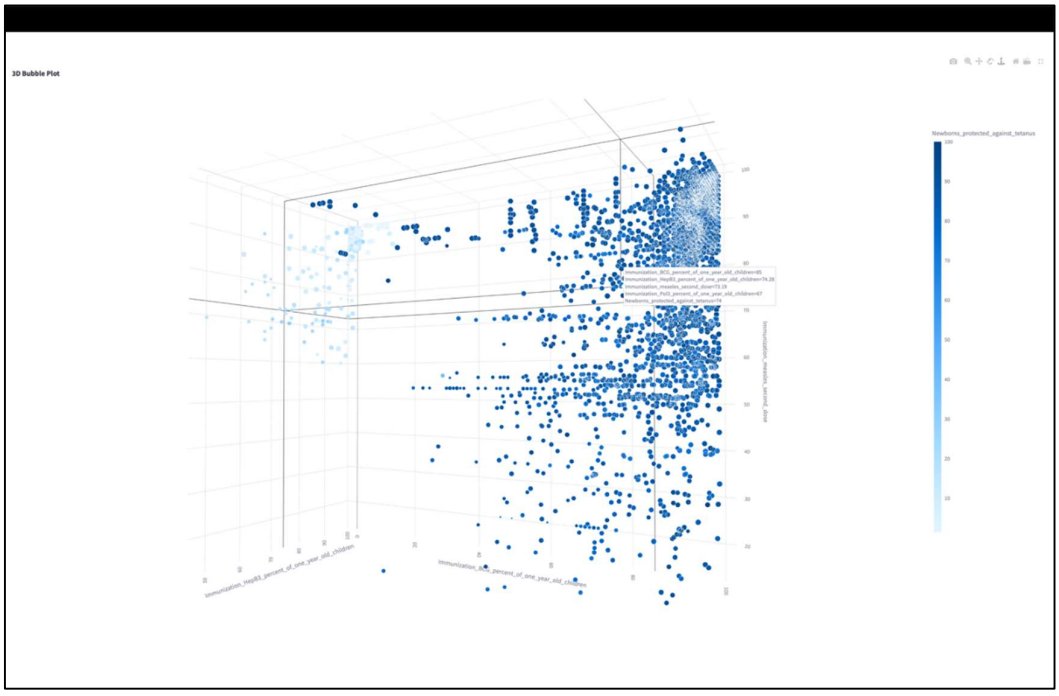
- Modelling:
  - Selecting a question and corresponding hypothesis:







- Selecting relevant columns: Using these columns, the model and visualization will be generated.



VIII. CONCLUSION:

The completion of this project marks a significant step toward leveraging data-driven insights to improve maternal and infant health outcomes. Through meticulous data acquisition, cleaning, and analysis, we have successfully identified critical factors that influence infant survival and maternal well-being. Our work has demonstrated the potential of statistical modelling and machine learning to uncover hidden patterns, predict high-risk areas, and suggest targeted interventions. It converts raw datasets into actionable knowledge and therefore underscores the role of technology and data analytics in addressing even complex public health challenges.

The predictive models in this study show how, from logistic regression to state-of-the-art neural networks, the vital role of variables related to healthcare infrastructure, immunization coverage, maternal literacy, and general socioeconomic conditions prevails. These insights make a basic understanding on which policymakers, healthcare providers, and leaders of the community can direct resources with a focus on health priorities and develop strategies addressing the disparities in maternal and infant care.

The easy-to-use data product developed in this work lowers the barrier between sophisticated analytics and actionable insights. We provided an interactive platform where; by combining data visualization, hypothesis testing, and predictive modelling, users could explore, analyse, and act on key findings. This tool not only enhances decision-making but also fosters greater accessibility and collaboration in tackling public health issues.

This project demonstrates the transformative potential of health data in improving outcomes for vulnerable populations. There is certainly room for future enhancements, such as integrating real-time data and refining models for broader generalizability, but our findings provide a solid foundation for continued research and intervention. By focusing on equitable healthcare delivery and leveraging data-driven strategies, this initiative aspires to contribute to a future where maternal and infant health disparities are significantly reduced.

IX. REFERENCES:

1. Dataset

   https://databank.worldbank.org/source/health-nutrition-and-population-statistics#

2. Streamlit

   https://docs.streamlit.io/

3. MySQL

   https://dev.mysql.com/doc/

4. Gaussian Mixture Models (GMM)

   https://scikitlearn.org/stable/modules/mixture.html

5. XGBoost

   https://xgboost.readthedocs.io/en/stable/

6. Decision Tree Classifier

   https://scikitlearn.org/dev/modules/generated/sklearn.tree.DecisionTreeClassifier.html

7. Convolutional Neural Networks (CNN)

   https://www.kaggle.com/code/kanncaa1/convolutional-neural-network-cnn-tutorial

8. K-Nearest Neighbours (KNN)

   https://scikitlearn.org/1.5/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

9. Logistic Regression

   https://scikitlearn.org/1.5/modules/generated/sklearn.linear_model.LogisticRegression.html

10. Support Vector Machines (SVM)

    https://scikit-learn.org/dev/modules/generated/sklearn.svm.SVC.html

11. Random Forest Regressor

    https://scikitlearn.org/dev/modules/generated/sklearn.ensemble.RandomForestRegressor.html

12. Multilayer Perceptron (MLP)

    https://scikit-learn.org/1.5/modules/neural_networks_supervised.html