Group Information

Topic / Dataset: Fertility Prediction

Group No: 11

Table 1: Group information

Student No	Student ID	Name	% of Contribution
1	2020-1-60-049	Lamia Haider	
2	2020-1-60-050	Shajia Hossain	
3	2020-1-60-129	Subah Nuzhat	

Table 2: List of datasets (DO NOT DELETE ROW 0)

No	Year: Dataset Short Name	Dataset Full Name	Dataset Link	Found by Student No	Comment
0	2023: ABC	Absolute Branch Dataset	Link1	3	The dataset was downloaded and stored on a shared drive
1		Fertility	Link	Alfe	
2		Fertility Rate By Race	Link	Alfe	
3		Fertility Dataset	Link	Alfe	
4		Fertility Dataset	Link	Alfe	
5		Bangladesh: Fertility rate from 2011 to 2021	Link	Alfe	
6	FD	Fertility Dataset	Link	3	The dataset was downloaded from an online source
7	ASW	Fertility Dataset of Age-specific women in Bangladesh for selected years	Link Link1	3	This dataset was uploaded on my drive

8	CFW	Completed fertility of women between age 40-50	Link Link2	3	The dataset was downloaded from an online source
9	ASWR	Fertility Dataset of Age-specific women in Bangladesh by Residence	Link Link1	3	This dataset was uploaded on my drive
10	PFL	Population, fertility rate, Life expectancy	Link	1	The dataset was downloaded from an online source
11	CW	Fertility rate: children per woman	Link	1	The dataset was downloaded from an online source
12	CWHBR	Fertility rate in each continent and worldwide, from 1950 to 2024	Link	2	The dataset was downloaded from an online source
13	CWLFR	Countries with the lowest fertility rates in 2023	Link	2	The dataset was downloaded from an online source
14	CWHFR	Countries with the highest fertility rates in 2023	Link	2	The dataset was downloaded from an online source
15	FRWC	Countries with the highest birth rate in 2023	Link	2	The dataset was downloaded from an online source
16	AFB	Adolescent fertility rate in Bangladesh from 2011 to 2021	Link	1	The dataset was downloaded from an online source
17	FE	Total fertility rate in Europe in 2023, by country	Link	1	The dataset was downloaded from an online source

Table 3: List of articles that cited the datasets in previous table (DO NOT DELETE ROW 0)

No	Paper title	Journal/conference name	Published Year: Paper Link	Citation count	Paper Cited dataset No. x from Table 2	Found by Student No
0	A brief history of time: an example paper name		2023: Link	1206	0, 5 and 6	2
1	Explainable AI to Predict Male Fertility Using Extreme Gradient Boosting Algorithm with SMOTE	MDPI	2022:Link	7	6	3
2	Increasing population densities predict decreasing fertility rates over time: A 174-nation investigation.	American Psychologist	2021:Link		8	3
3	Methods Protocol for the Human Fertility Collection	Methods Protocol for the HFC 17.06.2020	2015:Link	7	7,9	3
4	Male Female Fertility differential across 17 high-income countries	European Journal of Population (2021) 37:417–441	2018:Link	33	7,9	3
5	Variation in wealth and educational drivers of fertility decline across 45 countries	The Society of Population Ecology	2021:Link	43	8	3
6	Population Policies,	JSTOR	Link	132	8	3

	Family Planning Programs, and Fertility: The Record					
7						
8	Trends in total fertility rate in Ghana by different inequality dimensions from 1993 to 2014	BMC Women's Health	2022:Link	8	12	2
9	Population aging in the context of education. Comparison of selected EU countries	Population and Economics	2023:Link	Unknown	12	2
10	The influence of adolescent age at first union on physical intimate partner violence and fertility in Uganda : a path analysis	South African Journal of Child Health	2018:Link	7	14	2
11	The Health Status of Somalia	Foundations of Global Health	2021:Link	Unknown	14	2
12	Power Logics of Consumers’ Gendered (In)justices: Reading Reproductive Health Interventions through the Transformative Gender Justice Framework	Gender After Gender in Consumer Culture	2020:Link	50	14	2
13	Family-supportive workplace policies and benefits and fertility intentions in South Korea	Community, Work & Family	2020:Link	3	13	2
14	Historical Factors on Declining Fertility Rate		Link	Unknown	13	2

	in Japan					
15	A Systematic Review of Genetics and Reproductive Health Outcomes: Asian Perspective	Reproductive Sciences	2023:Link	1	13	2
16	Long-Run Macroeconomic Consequences of Taiwan's Aging Labor Force: An Analysis of Policy Options	Journal of Policy Modeling	2023:Link	2	13	2
17	SMEs in the 4th Industrial Revolution: Creative Tools to Attract Talents and Shape the Future of Work	IGI Global	2019:Link	Unknown	13	2
18	The World Needs More Asian Leadership in Global Health	Asia Global Institute	2022:Link	Unknown	13	2
19	Female Labour Supply and Fertility in Spain.	Lund University	2016:Link	Unknown	13	2
20	Global trends in total fertility rate and its relation to national wealth, life expectancy and female education	BMC Public Health	2022:Link	18	10	1
21	Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204	GLOBAL HEALTH METRICS	2020:Link	1023	10	1

	countries and territories, 1950–2019: a comprehensive demographic analysis for the Global Burden of Disease Study 2019					
22	Life Expectancy, Fertility, and Women’s Lives: A Life-History Perspective	Cross-Cultural Research	2013:Link	50	10	1
23	The impact of a reduced fertility rate on women's health	BMC Women's Health	2004:Link	25	11	1
25	Fertility Rate	Our World in Data	2014:Link	152	11	1
26	The Fertility Decline in Developing Countries	Scientific American	1993:List	175	11	1
27	Adolescent Fertility Rate for Bangladesh (SPADOTFRTBGD)	FRED ECONOMIC DATA	2023:Link	UNKNOWN	16	1
28	The Gap Between Lifetime Fertility Intentions and Completed Fertility in Europe and the United	Population Research and Policy Review	2019:Link	139	17	1

	States: A Cohort Approach					

Introduction

—--

Instructions section: Do not delete this section. It will be updated and color coded time to time

Word count suggestion: 1000 or more

In this section, write in your own words:

1. General discussion about these datasets:
   a. Why they are important
   b. How to they help us
2. General characteristics of these types of datasets:
   a. Types of data
   b. Quantity of data
3. General application cases / area / domain
4. You can write anything else related to this

—--

Write below:

Review of datasets

—--

Instructions section: Do not delete this section. It will be updated and color coded time to time

In this section, write in your own words:

1. An introductory paragraph (500 words or more)
2. Make a separate section describing each dataset. Do not copy-paste even a single line.
    a. Write characteristics of that dataset
    b. Write the limitation of that dataset
3. Finally, summarize all the datasets in a table.
4. You can write anything else related to this.—--

Write below:

Introductory para (>500 words):

Fertility is the quality of a woman's ability to produce offspring, which is dependent on age, health, and other factors and the total fertility rate often abbreviated as TFR is the average number of children born to a woman during her reproductive years. The Fertility Rate (TFR) dataset is an important repository of statistical information that provides insights into population dynamics and reproductive trends across the world and nations. This dataset indicates some special factors, including age, education, socioeconomic status, and cultural variables, providing a comprehensive perspective on fertility figures. Researchers and policymakers make use of the TFR dataset to recognize shifts in population dynamics, evaluate the impact of social and economic factors on fertility rates, and formulate informed strategies for family planning and public health involvements. By surveying fertility data, demographers can make predictions about future demographic interchange. This information is crucial for governments, businesses, and organizations in arranging for the needs of evolving populations. TFR datasets also allow for the evaluation of the effectiveness of various involvements and programs aimed to improve reproductive health and family planning.

Researchers, medical professionals, and policymakers stand to gain valuable insights from the fertility dataset—an extensive compilation of data on reproductive health and fertility patterns. This carefully curated collection incorporates a diverse array of variables, encompassing factors such as age, lifestyle choices, medical history, and socioeconomic indicators, contributing significantly to our understanding of the intricate dynamics of human reproduction. The dataset, inclusive of both male and female fertility, serves as a crucial resource for comprehensive analyses. Derived from a variety of sources, including clinical research, surveys, and medical records, the fertility dataset ensures privacy through anonymized data. This approach facilitates thorough analysis while respecting confidentiality. As a valuable tool for advancing reproductive science, the dataset enables the exploration of correlations and the identification of trends and risk factors. With its broad temporal scope, it offers insights into evolving fertility patterns over time, fostering evidence-based strategies for reproductive health management. The global fertility rate has seen a decline over the past few decades. Past fertility rates significantly differ from recent fertility rates over defined periods.Many nations have seen decreased fertility rates recently when compared to previous years. This trend has been fueled by a number of factors, including easier access to education, postponed marriages, and higher rates of female employment. A few nations have enacted laws to promote family planning and address the drop in birth rates.Most of the nations had some of the highest fertility rates in the world at the beginning of the 2000s—above five children per woman. Conversely, several European nations, including Spain and Italy, reported lower fertility rates, frequently falling short of the replacement level of 2.1 children per woman required to sustain a stable population.Some countries in the Middle East and South Asia also reported higher fertility rates compared to many developed nations..Over the past few decades, the global fertility rate has significantly decreased, with an average of 5 births per woman aged 15-49 years.

Dataset 6 (Fertility Dataset) :

Semen samples from 100 participants are included in this dataset, which was obtained from the UCI Machine Learning Repository and examined using the WHO 2010 criteria.

Numerous characteristics are included in the collection, such as lifestyle behaviors, environmental factors, health conditions, and sociodemographic data. The classification shows whether the sperm sample is classified as normal or changed. There are 100 cases overall with 10 features. The season, age, history of childhood illnesses, incidence of accidents or severe trauma, history of surgical procedures, episodes of high fevers within the previous year, frequency of alcohol consumption, length of daily sitting, and the final diagnosis are among the characteristics that have been identified.

Dataset 7 (Fertility Dataset of Age-specific women in Bangladesh for selected years) :

This dataset, which spans the years 1986 to 2021, provides information on Bangladesh's overall fertility rate (per woman) as well as age-specific fertility rates (per 1000 women). The Vital Registration System (VRS) of the Bangladesh Bureau of Statistics (B.B.S.) is the source of this dataset, which contains important characteristics including the overall fertility rate and age-specific fertility rates for different age groups, from 15–19 to 45–49.

Although the dataset offers a thorough understanding of fertility dynamics over time, it is important to remember that its application is limited to the designated era, which may limit its relevance to more recent developments. Fertility rate accuracy is dependent on the Vital Registration System's dependability and completeness, and there is a chance that some reports were submitted incorrectly or with insufficient information. Furthermore, the dataset could not contain comprehensive data on environmental factors—like socioeconomic status, medical developments, or legislative changes—that affect fertility rates. Moreover, it doesn't explore the precise causes of fertility rate variations or take into consideration momentous occasions that might have influenced these patterns.

Dataset 8 (Completed fertility of women between age 40-50) :

Important insights into fertility patterns can be gained from the datasets covering women aged 40 to 50, race and Hispanic origin, nativity, educational attainment, labor force status, and region of residence.

The dataset unveils essential traits of women aged 40 to 50, encompassing metrics like the percentage of never-married individuals (14.8%), the rate of childlessness (15.8%), and the childbirth rate (2,002 children per 1,000 women). Additionally, it examines fertility patterns concerning race, Hispanic origin, nativity, educational attainment, labor force status, and region of residence. These statistics offer valuable insights into marital and childbearing statuses, educational achievements, workforce participation, and the impact of geographical location on fertility.

The dataset on fertility trends has limitations, including a lack of detailed breakdowns within racial categories, a lack of insights into socio-economic factors, and insufficient details on educational attainment, labor force status, and region of residence. Despite these, the datasets contribute to understanding fertility trends and suggest areas for further exploration. However, they provide valuable insights into fertility patterns.

Dataset 9 (Fertility Dataset of Age-Specific Women in Bangladesh by Residence) :

The national fertility rate fell from 2.98 to 2.11 between 1998 and 2018, according to the dataset. The dataset provides a comprehensive analysis of changes in fertility over the past 20 years, but it is missing information on important determinants like socioeconomic status and education. Its 2018 endpoint restricts its relevance to current trends and offers little explanation for observed variations.

Total fertility rates in the rural dataset for the same period range from 2.28 in 2009 to 3.0 in 1998. However, generalizability is limited by the exclusion of metropolitan areas, and the dataset is deficient in information on contextual factors such as the accessibility of healthcare in rural areas.On the other hand, the total fertility rate decreased from 2.24 to 1.71 in the urban dataset covering the years 1998 to 2018.

Although it sheds light on urban fertility trends, it ignores rural areas and may thus be missing variances. There is also a lack of specific data on urban elements that affect reproduction, such as career possibilities and lifestyle choices.

Dataset 10 (Population, fertility rate, Life expectancy) :

This analysis focuses on several demographic datasets from various countries, covering the years 1960 to 2013. The datasets include information on the total population for each country during this period.

The dataset contains columns such as "Country Name," "Country Code," "Indicator Name," "Indicator Code," and population counts for each year from 1960 to 2013.For example, For instance, in 1960, Aruba recorded a population of 54,208, Andorra had 13,414, Afghanistan's population stood at 8,994,793, and so forth. Analyze and pinpoint countries displaying consistent population growth or decline. Contrast population sizes among countries in specific regions or continents. Identify nations with the highest and lowest populations for a given year. Compute the percentage change in population for each country over defined periods. Investigate broader global population trends during distinct decades. Determine the countries boasting the highest and lowest populations in the most recent available year.

The dataset only covers the period from 1960 to 2013, limiting the analysis to a specific historical range and aggregated at the country level, providing an overall population figure without detailed demographic breakdowns.

Together, these data shed light on age-specific fertility rates from 1998 to 2018, but interpretation should be done with care because data beyond 2018 are scarce, contributing factors have not been thoroughly examined, and certain geographic areas have been left out of the rural and urban datasets.

Dataset 11 (Fertility rate: children per woman) :

The provided data represents population figures for various countries and regions over the years 1950 and 2021. It includes absolute and relative changes in population for each country, as well as aggregated data for regions and income groups.

The dataset provides information on the population of various countries and regions for the years 1950 and 2021. The data includes the absolute change and relative change in population during this period. For dataset 1, each row represents a specific country or region, and the columns include the population data for the years 1950 and 2021, as well as the absolute change and relative change in population during this period. For dataset 2, groups of countries based on their income levels, distinguishing between high-income countries, low-income countries, upper-middle-income countries, and lower-middle-income countries.It provides population data for the years 1950 and 2021, along with the absolute and relative changes for each income group.

Some limitations of those dataset the regional classification might not capture the full diversity within each region, as countries within the same region can still have significant variations.The dataset does not explore the socio-economic factors contributing to population changes in different income groups.

Dataset 12 (Fertility rate in each continent and worldwide, from 1950 to 2024) :

The dataset holds the fertility rate in each continent and worldwide from the year 1950 to 2024.

The total fertility rate of the world has decreased from around five children per woman in 1950, to 2.3 children per woman in 2023, which defines that women today are preferring fewer than half the number of children that women did 75 years ago. When examining fertility rates on a continental scale, Africa stands out as the sole region surpassing the global average, along with Oceania, as the only regions displaying fertility rates above replacement levels. Prior to the 1980s, African women typically had nearly seven children in their lifetime, and as of 2023, there are some African countries where women of childbearing age can anticipate having five or more children. In contrast, Europe has historically maintained the lowest fertility rate globally throughout the past century, dropping below replacement levels in 1975. Europe's population has grown through a combination of migration and increased life expectancy; however, despite substantial immigration, its population entered a decline in 2021. Furthermore, when the worldwide average dips below the replacement level, (roughly 2.1 children per woman), anticipated to occur in the 2050s, it will result in a sustained global population decrease over the long term.

The drawback of the dataset is that it doesn’t contain the information of the factor that is influencing the change of the fertility rate over the years. The factors can be the substantial decline in infant and child mortality rates, a decrease in occurrences of child marriages, enhanced educational and vocational prospects for women, and improved accessibility and effectiveness of contraception. Since the dataset doesn’t contain these factors it’s hard to assure them.

Dataset 13 (Countries with the lowest fertility rates in 2023) :

The dataset represents the countries with the lowest fertility rates in 2023. It shows the lowest fertility rates of 20 countries. From that, the fertility rate in Taiwan is estimated to be at 1.09 children per woman in the year 2023, making it the lowest fertility rate worldwide.

In this dataset we have discovered only two characteristics which are Countries and the Total Fertility Rate(TFR). We can see in developed countries, fertility rates are usually much lower. Let’s assume the availability of birth controls over the counter and women often prioritize their career before becoming a mother is the reason behind it. Moreover, when there is a decrease in the number of women in their childbearing years, the fertility rate of a nation also decreases. On the contrary, developing countries have a higher fertility rate. It may be because of a deficiency in the availability of birth control and contraception, and women usually pursue a higher education rather than taking care of housework. It's worth noting that Bangladesh falls somewhere in between, not ranking among the countries with the lowest or highest fertility rates. At 2.00 TFR children per woman, the fertility rate in our country has remained slightly below the global average fertility rate of about 2.4 children per woman over the last decade.

The limitation of this dataset is the absence of one information that is the fertility rate of women by income group. If the income rate was present in the dataset we could’ve admitted the fact that the reason we have assumed behind the lower fertility rate is true.

Dataset 14 (Countries with the highest fertility rates in 2023) :

The data presents the estimated highest fertility rates in 2023 for 20 countries. During that year, Niger was projected to have a fertility rate of approximately 6.73 children per woman.

Niger boasts the world's highest fertility rate, nearing 7 children per woman, with Mali following closely in second place. theOn the other hand we can see that countries which have lower fertility rates are developing or developed countries. Consequently, as illustrated earlier, 9 of the 10 countries occupying the highest positions in terms of fertility rates are located in Africa.

The dataset lacks detailed contextual information about the socio-economic, cultural, and healthcare factors influencing fertility rates in the mentioned countries. Without this context, it becomes challenging to draw accurate conclusions or propose effective interventions. Again, fertility rate estimations are subject to statistical methods and assumptions. Variability in data collection methods, reporting practices, and potential inaccuracies in estimating birth rates could affect the reliability of the dataset.

Dataset 15 (Countries with the highest birth rate in 2023) :

The dataset displays the highest birth rate of the countries in the world in 2023.

Among the countries Niger held the top position for the highest global birth rate, recording 46.86 births per 1,000 residents. The trend continued across the African continent, as Angola, Benin, Mali, and Uganda closely followed suit, each experiencing more than 40 births per 1,000 people during the same period.What is particularly noteworthy is that, apart from Afghanistan, all the countries securing positions within the top 20 for the highest birth rates on a global scale were concentrated in the Sub-Saharan African region. From the previous dataset we have seen that Niger also holds the highest fertility rate in the world in 2023 which is obvious but the mentionable fact is that the other countries in the previous dataset do not hold the same serial in birth rate dataset. Some countries' fertility rates and birth rates have a noticeable difference.

The dataset suggests a correlation between low-income countries and higher birth rates without thoroughly exploring the causal factors. This oversimplification may lead to a stereotypical view of the relationship between income and fertility. To enhance the dataset's robustness and applicability, a more comprehensive and nuanced approach would involve incorporating a broader range of factors, ensuring data accuracy, and acknowledging the diversity within regions or countries.

Dataset 16 (Adolescent fertility rate in Bangladesh from 2011 to 2021) :

This dataset represents the trends in a population's reproductive health depending on knowing the number of births per 1,000 women between the ages of 15 and 19. This is known as the adolescent fertility rate. The Bangladesh dataset, which covers the years 2011 through 2021, offers important new information on the trends in teenage fertility in the country during the previous ten years.

In 2011, the teenage fertility rate reached 100.96 births per 1,000 women, attributed to limited access to reproductive health services. In 2012, the rate remained high.In 2013, fertility rates fell to 99.31, indicating potential changes in reproductive health practices. In 2014, they dropped to 88.56, indicating potential improvements in family planning knowledge.In 2015 and 2016, the rates show no significant changes, suggesting that the declining trend may have reached a standstill.The rates showed a slight decline in 2017-2018, with 78.29 and 78.36 respectively, but more significant declines may be challenging to achieve.In 2019, fertility rates decreased slightly to 77.68, but by 2020-2021, they remained high at 76.41 and 75.50, indicating ongoing reproductive health issues.

The dataset has some limitations like it lacks details on socio-economic, educational, and cultural factors influencing adolescent fertility. It mainly focuses on fertility rates, missing data on education, and socio-cultural aspects.

The Bangladeshi dataset on adolescent fertility rates provides valuable insights into reproductive health, but ongoing issues require comprehensive strategies.

Dataset 17 (Total fertility rate in Europe in 2023, by country) :

The total fertility rate (TFR) in Europe for the year 2023 provides a comprehensive overview of the average number of children a woman is expected to have during her lifetime across various European countries.

According to the dataset, the Faroe Islands was predicted to have the highest fertility rate among European nations in 2023. There were 2.71 children born to each woman in the tiny island state in the Atlantic. In 2023, France emerged as the country with the highest fertility rate, with 1.79 children born to every woman, while other small nations like Monaco and Gibraltar also ranked highly. However, Andorra, San Marino, and Malta had the lowest fertility rates in all of Europe; the largest countries with low fertility rates in that year were Italy, Spain, and Ukraine, with an average of 1.3 children per woman.

The dataset contains certain limitations, such as the Total Fertility Rate (TFR), which ignores variables like age, education, and societal changes and simplifies fertility patterns.The dataset may not capture dynamic changes over time because it is dated 2023.

Dataset Summary:

No	**Year: Dataset Short Name**	**Dataset Full Name**	**Dataset Summary**
6	FD	Fertility Dataset	The dataset explores sociodemographic, environmental, health, and lifestyle factors in Bangladesh's fertility rates from 1986 to 2020, but has limitations such as focusing on trends up to 2011, reliance on system dependability, and unclear causes.
7	ASW	Fertility Dataset of Age-Specific Women in Bangladesh for selected years	The dataset, retrieved from the Vital Registration System, includes age-specific and overall fertility rates for Bangladesh from 1986 to 2021. Limitations include the possibility of errors resulting from underreporting, the absence of contextual information, and the emphasis on trends rather than particular causative elements, even though the insights are valuable.

8	CFW	Completed fertility of women between ages 40-50	Databases focusing on women aged 40-50 offer crucial data on fertility trends, including unmarried and birth rates. Understanding married, childbearing, educational, and workforce dynamics is enhanced by examining factors like race, education, and region. Despite limitations, these datasets provide valuable insights for further research.
9	ASWR	Fertility Dataset of Age-specific women in Bangladesh by Residence	The national fertility record from 1998 to 2018 shows a decrease from 2.98 to 2.11, however, it is devoid of socioeconomic factors. Although the urban dataset (2.24-1.71) ignores rural areas and lacks specific urban impact statistics, the rural dataset (2.28-3.0) lacks metropolitan representation and healthcare context. Due to geographic exclusions and data restrictions, interpretation must be done with caution.
10	PFL	Population, fertility rate, Life expectancy	The dataset includes columns for country names, codes, and population counts from 1960 to 2013. It allows analysis of population trends, contrasting population sizes, identifying countries with high and low populations, and calculating percentage changes over defined periods. It also provides insights into global trends.
11	CW	Fertility rate: children per woman	The dataset shows population changes in countries and regions

			between 1950 and 2021, categorized by income levels. It includes absolute and relative changes for each country or region, and categorizes countries into high-income, low-income, upper-middle-income, and lower-middle-income groups.
12	CWHBR	Fertility rate in each continent and worldwide, from 1950 to 2024	From the worldwide Total Fertility Rate (TFR) data, the African continent has the highest fertility rate than any other continent over the time.
13	CWLFR	Countries with the lowest fertility rates in 2023	Among 20 countries from the dataset fertility rate, Taiwan is estimated to be at 1.09 children per woman, making it the lowest fertility rate worldwide in 2023.
14	CWHFR	Countries with the highest fertility rates in 2023	The country with the highest fertility rate in 2023 from the datasets is Niger. The country is from the African continent which we already know has a record of highest fertility rate over the years.
15	FRWC	Countries with the highest birth rate in 2023	As we have seen from the previous dataset that Niger has the highest fertility rate in the world so it is obvious that this country will also have the highest birth rate. But it is not the same for every country. If we compare this dataset from the previous one we can see the fertility rate of countries from highest to lowest in the previous dataset does not match with the series of countries from highest to lowest birth rate.

16	AFB	Adolescent fertility rate in Bangladesh from 2011 to 2021	Bangladesh's adolescent fertility rate has consistently declined from 2011 to 2021, with a significant drop in 2014 and a gradual decline to 75.50 in 2021. This suggests positive trends in reproductive health and family planning interventions...
17	FE	Total fertility rate in Europe in 2023, by country	In 2023, the Faroe Islands had the highest fertility rate among European countries, followed by Monaco and Gibraltar. France had the highest rate at 1.79 children per woman. Conversely, Andorra, San Marino, and Malta had the lowest rates.

Review of techniques applied on these datasets

—---

Instructions section: Do not delete this section. It will be updated and color coded time to time

In this section, write in your own words:

1. An introductory paragraph (500 words or more)
2. Find different techniques and categorically explain them. For example, deep learning based techniques, traditional ML based algorithms, feature fusion etc.
3. Make a table to summarize them
4. You can write anything else related to this.

—---

Write below:

Introductory para (>500 words):

The study focuses on accurate fertility prediction in reproductive health research, integrating statistical techniques, knowledge-based insights, and machine learning. It seeks to understand biological, demographic, and environmental factors impacting fertility dynamics, promoting tailored healthcare plans and interventions through methods like statistical studies and machine learning algorithms.The pursuit of precise and sophisticated fertility prediction has emerged as a central theme in the quickly changing field of reproductive health research, propelling the fusion of innovative approaches and a variety of methodologies. A comprehensive strategy integrating statistical rigor, knowledge-based insights, and machine learning's computational capability is required due to the complex interactions between biological, demographic, and environmental elements. This overview examines the complex field of fertility prediction methods, which includes a range of approaches that together provide a thorough knowledge of reproductive outcomes.

Statistical methods are the foundation of evidence-based decision-making and are at the heart of fertility prediction. Statistical approaches explore patterns and interactions within fertility-related data, ranging from conventional descriptive statistics to sophisticated techniques including regression analysis, time series analysis, and survival analysis. Over time, a sophisticated examination of reproductive processes is achieved through the use of structural equation modeling, confidence interval computation, and dataset disaggregation. Our comprehension of the temporal subtleties inherent in fertility dynamics is improved by techniques such as the Autoregressive Distributed Lag (ARDL) model and hurdle models, which provide insights into the dynamic interactions between variables over various time intervals.

Knowledge-based methods enhance statistical methods by adding a qualitative dimension to fertility prediction. The gap between theoretical comprehension and practical application is bridged by the calibrated spline (CS) estimator, conditional imputation technique, and expert consultations. For example, the CS estimator fits a smooth curve based on known fertility age patterns to refine data. Expert consultations, in the meantime, guarantee that research findings are well-informed, relevant to the situation, and significantly advance the field by adding domain-specific knowledge to the predictive models.

Machine learning algorithms have become strong tools for fertility prediction in the big data era. These strategies use computing power to find complex patterns in large datasets. They range from ensemble learning approaches like random forests and gradient boosting to sophisticated deep learning models. The interpretability problem presented by complicated machine learning models is addressed by explainable AI (XAI) techniques, whilst extreme gradient boosting (XGB) proves to be scalable and effective. Machine learning is versatile and can be used for a variety of tasks, such as class imbalance correction, trend analysis, and fertility rate forecasting. Examples of these tasks include cluster analysis, SMOTE, and predictive modeling.

The field of fertility prediction is always growing as new methods emerge that provide fresh insights. By examining long-term trends, seasonality, and temporal patterns in fertility rates, time series analysis sheds light on how reproductive behavior changes over varying time spans. While geospatial analysis examines geographical variations and hotspots, survival analysis measures the amount of time until particular reproductive events occur. The capacity of techniques to be applied to various elements of reproductive research is demonstrated by the ways in which ensemble learning methods aggregate heterogeneous models, deep learning automatically extracts hierarchical characteristics, and Bayesian methods provide a probabilistic framework.

Technique 1 (Statistical-based techniques):

To make evidence-based decisions regarding reproductive health, fertility prediction researchers and healthcare professionals benefit from statistically based techniques that analyze patterns and relationships within fertility-related data. Examples of these techniques include descriptive statistics, correlation analysis, regression analysis, time series analysis, survival analysis, ANOVA, Chi-Square test, Principal Component Analysis, logistic regression, and quantile regression.

Period Mean:

Period Mean Ages at Birth (1986-2020): Statistical techniques are probably used to compute the mean ages at birth, which are derived by averaging the ages of women during delivery for each particular year (1986-2020). The comprehension of changes in reproductive practices over a 35-year period is aided by these analyses.[2015]

This thorough research provides a sophisticated view of reproductive processes across two unique time periods, from 1986 to 2020 and 2003 to 2018. It includes fertility dynamics, cumulative rates, total fertility trends, and average ages at birth.

Disaggregation:

Disaggregation is the process of dissecting or dividing a more comprehensive aggregated dataset or information into its constituent parts or subcategories.
Statistical techniques used in the analysis of data from the 1993–2014 Ghana Demographic and Health Surveys include the disaggregation of the Total Fertility Rate (TFR) by place of residence, education level, wealth index, and region. Additionally, statistical measures such as Difference, Population Attributable Risk, Ratio, and Population Attributable Fraction are used to estimate inequality. [2022]

Confidence Interval:

The mention of a 95% confidence interval indicates that the repetition of experiment or survey over and over again, will match the results 95 percent of the time. [2022]

Structural Equation Modeling (SEM):

This method is used for modeling and examining the partnerships between observed and unobserved variables. It combines multiple regression and component analysis to enable researchers to investigate intricate direct and indirect interactions between variables within a broad framework.

- **Path analysis:** It is used in the field of structural equation modeling (SEM) to evaluate and quantify the connections between elements in a hypothesized causal model. It is based on data from Uganda Demographic and Health Survey (UDHS) in 2011**. [2018]**

The Autoregressive Distributed Lag (ARDL):

For analyzing the relationship between variables over time, econometricians frequently use the Autoregressive Distributed Lag (ARDL) model, a prominent time series statistical technique. The Autoregressive (AR) and moving average (MA) components are part of the ARDL model, which can handle both lag-free and distributed lags. In its most basic form, the ARDL model is as follows:

$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_{t-1} + \ldots + \beta_{\square} Y_{t-\square} + \beta_{\square+1} X_{t-\square} + \gamma_1 \Delta Y_{t-1} + \gamma_2 \Delta X_{t-1} + \ldots + \gamma q \Delta Y_{t-q} + \gamma q_{+1} \Delta X_{t-q} + \varepsilon_{\square}$ Where:

- Y_t is the dependent variable at time t,

- X_t is the independent variable at time t,

- β_0 is the intercept,

- $\beta_1, \beta_2, \ldots, \beta_{p+1}$ are coefficients for the lagged values of Y and X,

- $\gamma_1, \gamma_2, \ldots, \gamma_{q+1}$ are coefficients for the differenced values of Y and X,

- ϵ_t is the error term,

- p is the number of lagged terms for Y and X,

- q is the number of different terms for Y and X. [January–February 2023]

Hurdle Model:

In the study of the number of children under the age of five living with their mothers the variable has some specific characteristics where it can only be zero, one, two, three, or four, and there are more chances where the number of children can be zero. This type of data often has more zeros than a typical model might expect, this is called the problem of excess zeros. In this case, Hurdle model uses a two-step process:

- **Probit Model (Probability Model):** First step is to use a Probit model to understand the probability of a woman being a mother or not.
- **Count Model (Zero-Truncated Poisson Model):** Second step is to focus only on those women who are mothers (have at least one child). For this group, it explores different count models like negative binomial, ordinary Poisson, and zero-truncated Poisson to anticipate the actual number of children. After trying those models, it is found that the zero-truncated Poisson model is the most suitable. [2016]

Analytical strategies

It is based on descriptive statistics, correlations, and distributed lag non-linear models (DLNMs) for longitudinal data analysis. It calculated TFR and relevant variables for each region using mean and standard deviations. Log-transformed GDP per capita was used due to skewed distribution. Pearson correlation was used to evaluate correlations, and gaussian distribution family DLNMs were used to evaluate variables' effects on TFR.(H Cheng, W Luo, S Si,2022)

Sensitivity analysis

It is a statistical technique that assesses the impact of changes in key parameters on research findings. It helps researchers identify uncertainties and bias sources, enhancing the validity and trustworthiness of statistical models and study findings under various conditions.(H Cheng, W Luo, S Si, 2022)

Multivariate analysis

This type of analysis relies on the correlations between two or more measures to comprehend the relationships between variables. If experimental data indicate that combining correlated measures yields better classification than analyzing each one separately, then this model will be useful. The primary difficulty with multivariate statistical IDs is estimating distributions for high-dimensional data.**(**J Payne ,2004**)**

Weighting

Weighting in statistical analysis assigns different influence to individual observations, ensuring accurate population representation. It accounts for biases and demographic features, and is crucial for producing more accurate estimates and generalizable results. Researchers use weighting techniques in social sciences and public health.**(**J Payne ,2004**)**

Poisson Regression:

In both datasets, Poisson Regression is a crucial statistical technique that aims to model the number of live births while taking into account a wide range of covariates. In these analyses, the main goal of using Poisson Regression is to evaluate changes in fertility rates over time, especially when working with several survey years. This approach provides a statistical framework to comprehend the effects of many variables on the number of live births, which helps shed light on the dynamics of fertility patterns. Notably, Poisson Regression is used to analyze fertility-related data on both the Age-Specific and Total Fertility Rate datasets, demonstrating its adaptability.[2018]

Technique 2 (Knowledge-based Techniques):

CS(Calibrated Spine) Estimator:

Dividing Aggregated Age Groups into Age Groups of One Year:

Using a knowledge-based method called the calibrated spline (CS) estimator to divide data that was first categorized into closed age intervals. This approach interpolates fertility rates by fitting a smooth curve based on established fertility age patterns while maintaining a balance between fit and shape criteria. The CS estimator is a useful knowledge-based method for data refinement since it considers a priori information about the current shapes of age-specific fertility rates.[2015]

Conditional Imputation Approach:

The Conditional Imputation Approach works well with the dataset, which includes a variety of demographic variables like race, Hispanic origin, nativity, educational attainment, labor force status, and region of residence. It fills in the gaps in the paternal age figures, especially as the dataset focuses on women between the ages of 40 and 50. This technique improves the dataset's completeness by including the link between paternal and maternal ages, allowing for a more robust analysis of male fertility patterns across various demographic groups.

This approach addresses missing values in a way that is consistent with the demographic variables that are displayed, and it is consistent with the structure of the dataset. A more thorough comprehension of fertility trends within the designated age range and demographic categories is made possible by the imputed paternal ages.[2021]

The Interpolation Method:

There are cases in the dataset where the parent's age is given as of the end of the year the child was born. The Interpolation Method is used to precisely ascertain the ages of the parents at childbirth. This is especially important for nations like Sweden, where parental ages are given at the end of the birth year in the statistics. The approach makes use of interpolation techniques created specifically for this purpose, by the national data reporting policies of the relevant nations. This technique helps to provide a more accurate and detailed analysis of age-specific fertility rates within the dataset by correcting temporal fluctuations in the reporting of parental ages.[2021]

Expert Consultation

Expert consultation is a cooperative process that improves the breadth and caliber of studies on reproductive health and demography. It acts as a link between theoretical understanding and real-world application, guaranteeing that study findings are informed, pertinent to the context, and make a significant contribution to the field.(M Roser ,2004).

Technique 3 (Machine learning-based techniques):

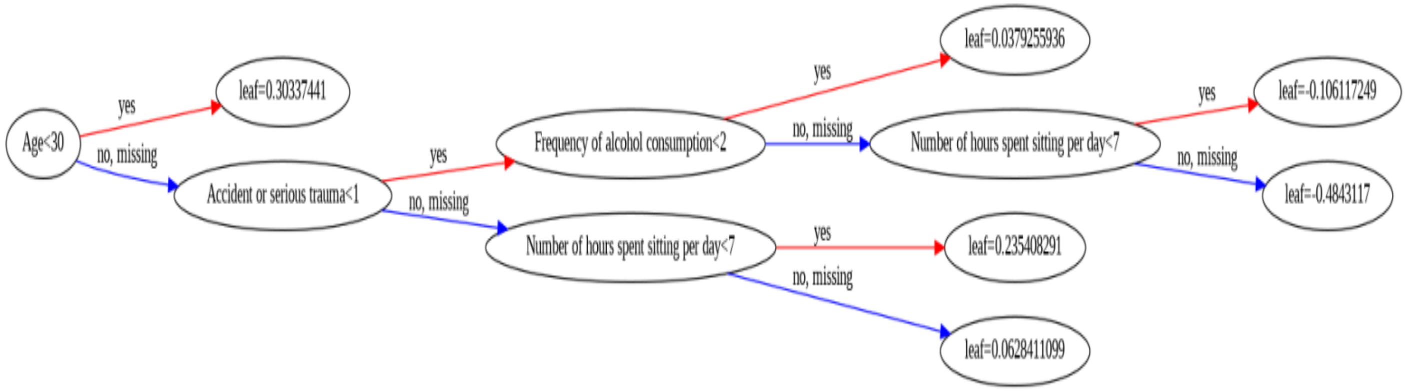
Machine learning techniques are essential for analyzing fertility rates in diverse datasets. The Fertility Dataset for Bangladesh, which examines lifestyle, environmental, and sociodemographic factors, uses methods like support vector machines, decision trees, and random forests. K-nearest neighbors and naive Bayes algorithms are used to analyze age-specific fertility rates. ML approaches are also used to study global population trends, analyze trends over time, and evaluate fertility rates in specific countries. These algorithms contribute to healthcare and demographic research.

- **Extreme Gradient Boosting:**

The scalability and efficiency of the XGB method are demonstrated in the analysis of a dataset containing one hundred semen samples. This technique, which was made popular by Chen and Guestrin in 2016, has a gradient-boosting decision tree basis and is specifically designed to tackle classification, regression, and ranking tasks. Gradient boosting is used to build each tree, which consistently minimizes differences between real and anticipated values to meet the objectives of both regression and classification. By 2015, XGB had proven its mettle by finishing 17 of 29 ML challenges on Kaggle with success.[MDPI]

The accuracy after running Xboosting is 85%

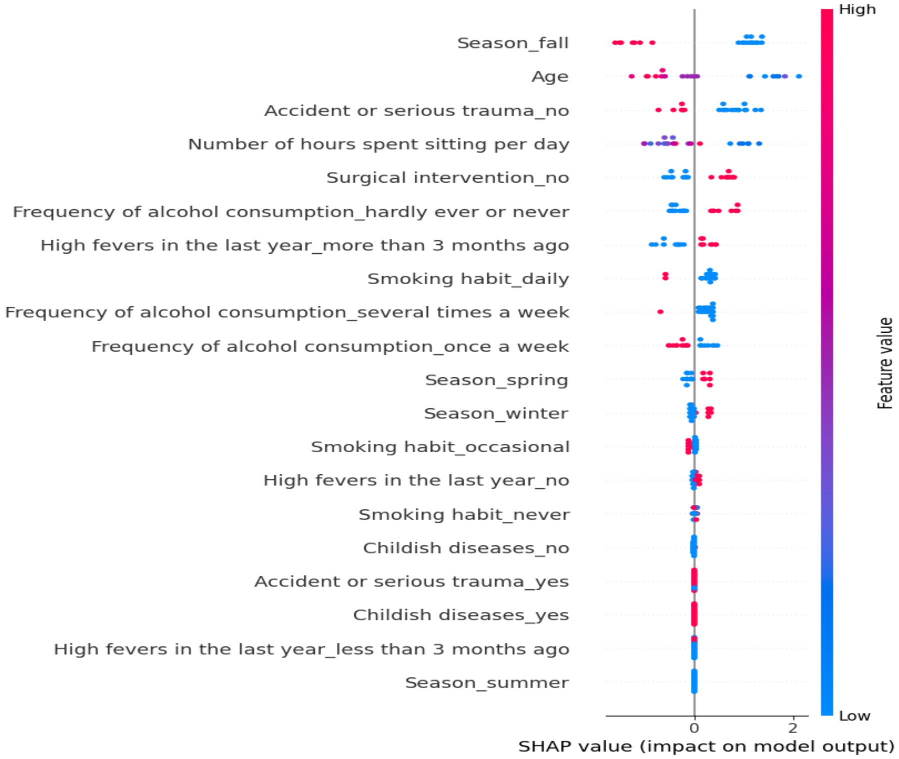
The first tree of the Xboosting method is following:



❖ Explainable AI(XAI):

Black box AI and ML models' incapacity to explain themselves in the setting of 100 semen samples becomes critical. Using this dataset, the Explainable AI (XAI) method turns opaque systems into transparent ones to help interpret prediction reports. Explainability is crucial for understanding AI judgments, learning about the system, and exploring the reasoning behind decision-making. Three main advantages come from using XAI in this semen dataset: (a) clear interpretation that builds model trust; (b) troubleshooting made possible; and (c) identification of the system basis's source. Global and local XAI techniques play important roles in the creation of AI systems for a dataset of 100 semen samples, where explainability and efficacy must be balanced. [MDPI]

The graph of XAI:



Predictive modeling

Through the use of machine learning algorithms, predictive modeling creates models that use a variety of predictors and historical data to predict future fertility rates. The procedure entails gathering pertinent information about fertility rates and the variables that affect them. The model can forecast future fertility rates because it has learned patterns and relationships from training on historical data. Including a variety of predictors, such as socioeconomic factors, educational attainment, and access to healthcare, improves prediction accuracy. In the end, these prediction models provide useful resources to help researchers and policymakers foresee and prepare for future fertility-related demographic trends.(M Roser , 2014).

Cluster Analysis:

Cluster analysis is a machine learning based technique used in grouping a set of objects (can be data points or observations) in such a way that objects in the same group or cluster are more similar to one another than to those in other groups. The objective is to divide a dataset into groups according to the underlying patterns or commonalities found in the data. This method is frequently applied in several domains, such as pattern recognition, machine learning, and data analysis.There are different types of clustering analysis among them the one here used is the hierarchical cluster analysis to compare the 4 countries (Czech Republic, Slovakia, Poland and Hungary) with Scandinavian countries. It was performed to partition the European countries into 4 clusters based on economy and education indicators.

- **Standardization:** Standardize the variables if they have different scales or units to ensure that no single variable dominates the clustering process.
- **Ward's criterion:** It is used for quality assessment.

The values of the characteristics that were taken into consideration had to be standardized initially because of their disparate dimensions. Thus, the distance measure between the items was calculated using a typical Euclidean metric. [2023]

This method helps identify developmental commonalities, regional trends, and family planning efficacy, guiding targeted interventions and policy choices in reproductive health and demographic trends.(M Roser ,2014).

Technique 4 (Synthetic Minority Oversampling Technique (SMOTE)):

The Synthetic Minority Oversampling Technique (SMOTE), an oversampling approach designed to increase the number of samples in the minority class, is used in this study to address the issue of class imbalance. Comparable approaches, such as ESMOTE and SMOTE, have been used in earlier studies to improve the minority class instances. When using this oversampling strategy, the number of occurrences in the minority class is represented by kc, and the number of instances in the majority class is shown by Pc. The symbol for the imbalance ratio between Pc and kc is zc. The number of occurrences in the dataset is calculated using the following equations after the SMOTE oversampling technique has been applied:

Total instances $=Pc+kc+k'c$

In this case, $k'c = (1-zc) * kc)$, denoting the quantity of artificially created instances.

The study uses the SMOTE oversampling technique for a dataset of 100 semen samples from the UCI Machine Learning Repository [2022], aiming to balance normal and altered sperm samples, ensuring the model's performance and avoiding bias in healthcare studies.

Dataset Summary:

	Techniques Name	**Technique Summary**
Statistical-based techniques	**Period Mean**	The study examines reproductive practices over a 35-year period (1986-2020) using statistical techniques, focusing on mean ages at birth, cumulative rates, total fertility trends, and fertility dynamics across two distinct time periods.
	Disaggregation	Disaggregation is the process of dividing a dataset into subcategories, as seen in the 1993-2014 Ghana Demographic and Health Surveys, which used measures like Total Fertility Rate, Difference, and Ratio to estimate inequality.
	Confidence Interval:	The 95% confidence interval suggests that repeated experiment or survey results will be consistent with 95% accuracy.

	Structural Equation Modeling (SEM):	Path analysis is a method used in structural equation modeling (SEM) to analyze the relationships between observed and unobserved variables, based on data from the Uganda Demographic and Health Survey in 2011.
	The Autoregressive Distributed Lag (ARDL):	The Autoregressive Distributed Lag (ARDL) model is a time series statistical technique used by econometricians to analyze the relationship between variables over time.
	Hurdle Model	A Hurdle Model is a statistical based model used when analyzing data with excessive zeros, which are values that are more frequent than expected. It uses two steps which are probability model and count model.
	Analytical strategies	The study uses descriptive statistics, correlations, and distributed lag non-linear models for longitudinal data analysis, calculating TFR and relevant variables for each region, using log-transformed GDP per capita
	Sensitivity analysis	The technique evaluates the influence of changes in key parameters on research findings, identifying

		uncertainties and biases, thus enhancing the validity and trustworthiness of statistical models.
	Multivariate analysis	Multivariate statistical IDs use correlations between measures to understand variables' relationships, but estimating distributions for high-dimensional data is a challenge.
	Weighting	The technique evaluates the influence of changes in key parameters on research findings, identifying uncertainties and biases, thus enhancing the validity and trustworthiness of statistical models.
	Poisson Regression:	Poisson Regression is a statistical technique used to model live births and fertility rates over time, particularly in Age-Specific and Total Fertility Rate datasets. It helps understand the effects of variables on live births and fertility patterns, demonstrating its adaptability

Knowledge-based Techniques	**CS(Calibrated Spine) Estimator**	The calibrated spline (CS) estimator is a knowledge-based method used to divide aggregated age groups into closed age intervals, interpolating fertility rates based on established age patterns, and is useful for data refinement.
	Conditional Imputation Approach	The Conditional Imputation Approach is a method used to fill gaps in data on male fertility patterns, particularly in women aged 40-50. It improves the dataset's completeness by integrating paternal and maternal ages, allowing for a more comprehensive understanding of fertility trends within specific age ranges and demographic categories.
	The Interpolation Method	The Interpolation Method is a technique used to accurately determine parents' ages at childbirth, particularly in countries like Sweden where birth year information is given. This method corrects temporal fluctuations in reporting of parental ages, ensuring a more detailed analysis of age-specific fertility rates
	Expert Consultation	Expert consultation enhances reproductive health and

		demography studies by linking theoretical understanding with practical application, ensuring informed, relevant findings that significantly contribute to the field.
Machine learning-based techniques	**Extreme Gradient Boosting**	The XGB method, popularized by Chen and Guestrin in 2016, uses a gradient-boosting decision tree basis to tackle classification, regression, and ranking tasks. It achieved success in 2015 by completing 17 out of 29 machine learning challenges on Kaggle.
	Explainable AI(XAI)	The Explainable AI (XAI) method is used to interpret prediction reports in a 100 semen dataset, enhancing understanding of AI judgments, learning about the system, and exploring decision-making reasoning. It provides clear interpretation, enables troubleshooting, and identifies the system basis's source, balancing explainability and efficacy.
	Predictive modeling	Predictive modeling uses machine learning algorithms to predict future fertility rates by incorporating various predictors and historical data. This process improves

		accuracy by learning patterns from historical data and incorporating socioeconomic factors, educational attainment, and healthcare access, aiding researchers and policymakers.
	Cluster Analysis	Cluster analysis is a machine learning technique used to group data based on commonalities, aiding in pattern recognition, pattern recognition, and data analysis, particularly in comparing European countries based on economic and education indicators.
Synthetic Minority Oversampling Technique (SMOTE)		This study employs the Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance, a strategy similar to ESMOTE. The minority class's occurrences are represented by *kc*, while the majority class's is represented by *Pc*, with an imbalance ratio symbol.