Exploring the factors affecting Casaerean Section delivery across India

Neha Gunta
Department of Computer Science and Information Systems
Birla Institute of Technology and Sciences, Hyderabad Campus
Hyderabad, India
f20200073@hyderabad.bits-pilani.ac.in

Shreya Koditala
Department of Computer Science and Information Systems
Birla Institute of Technology and Sciences, Hyderabad Campus
Hyderabad, India
f20200176@hyderabad.bits-pilani.ac.in

Sanjuktha Anem
Department of Computer Science and Information Systems
Birla Institute of Technology and Sciences, Hyderabad Campus
Hyderabad, India
f20200001@hyderabad.bits-pilani.ac.in

Anurag Choudhary
Department of Electrical and Electronics Engineering
Birla Institute of Technology and Sciences, Hyderabad Campus
Hyderabad, India
f20201797@hyderabad.bits-pilani.ac.in

***Abstract*—This study aims to examine the prevalence and trends of C-section deliveries in India using the National Family Health Survey (NFHS) datasets. The data was cleaned and preprocessed using techniques such as sampling, normalization. The highest prevalence of C-section deliveries was found in urban areas (32.9%), followed by rural areas (14.9%). Data Analysis techniques like Logistic Regression and Naïve Bayes were used and the features highly affecting C-section were identified. Further research is needed to understand the factors contributing to the increasing trend of C-section deliveries in India and to identify appropriate interventions to reduce unnecessary C-sections**

Keywords—caesarean, C-section, pregnancy, India, economic status, data

I. Problem Motivation

A Caesarean section (C-section) is a surgical procedure used to deliver a baby through an incision in the mother's abdominal wall and uterus. C-sections may be planned in advance if a vaginal delivery is deemed too risky, or they may be done as an emergency procedure if complications arise during labour. Caesarean is generally considered to be a safe procedure, but it comes with a lot of risks such as, infection, bleeding, blood clots, and injury to internal organs.

Apart from the factors concerning health, a lot of social-economic practices also come into play when a mother considers C-section over normal delivery. Awareness about the procedure, availability of facilities and guidance, financial and family support are few of the key factors that influence this decision. These factors vary to a great extent between the rural and urban population of India. Hence, we aimed to find out the major factors that influence a mother's consideration for C-section, mainly focussing on the parameters related to health(BMI, anaemia etc), financial status, awareness and availability of facilities..

II. Background

A. The dataset- NHFS-4

The dataset selected is the National Health and Family Survey (NFHS) conducted in India in 2015-16.The NFHS-4 [8] conducted in 2015-16 is the fourth round of the survey and collected data from over 600,000 households and 700,000 women and men across India. The questionnaire used for this included a wide range of topics including fertility, nutrition, maternal and child health etc.

III. Literature Survey

Health care institutions around the world are concerned about an increase in Caesarean section (CS) rates. An increase in the CS rate in a developing nation has a significant impact on the scarce health care resources. A research by the ICMR found that CS rates rose from 21.8% in 1993-1994 to 25.4% in 1998-19994 in 30 teaching hospitals in India[1].

According to the World Health Organisation's (who) guidelines, modified in 1994, the Caesarean birth rate in any population group should range between 5% and 15% (who 1994).An analysis of the National Family Health Survey data shows that the rate of this form of delivery in states like Kerala, Goa, Andhra Pradesh, West Bengal and Tamil Nadu is alarmingly high[2].

In the past, the percentage of deliveries that were performed via Caesarean section served as a gauge for determining the severity of complications and population groups' access to high-quality obstetric care. The same indicator is currently acting as an alarm for potential medical intervention overuse. The immediate consequence for the household is the economic burden since the economic cost of a c-section far exceeds the cost of a normal delivery.

Research shows that the risk of maternal death following a caesarean section is five to seven times higher than vaginal birth (report of the NIHS 2006)[3].Hence, the trend of c-section deliveries analysed from 1992-93 to 2005-06 shows an upward trend in c-section rates.

In 1998-99,it was around 4.8% in rural areas and 14.9% in urban areas. It increased to 6.2% in rural and 17.8% in urban during 2005-06.The Janani Suraksha Yojana (JSY) programme may have a significant impact on how willingly disadvantaged women accept institutional births[4].In 1998-99,it was around 4.8% in rural areas and 14.9% in urban areas. It increased to 6.2% in rural and 17.8% in urban during 2005-06.The Janani Suraksha Yojana (JSY) programme may have a significant impact on how willingly disadvantaged women accept institutional births[4].

The increase in CS in India may be due to growing institutional delivery. Perhaps, the most important factor impacting CS now is the place of delivery (private or public medical facility). The health system is burdened by an increase in caesarean section delivery rates.[5]

By NH4 and NH5,it is said that compared to other regions of India, the southern states have a greater rate of C-section births. When it comes to C-section deliveries, literacy is crucial. Births by C-section are more likely to occur in those who are 30 to 40 and older. Women in the median wealth index category had the highest likelihood of having a C-section (OR-CI, 1.62 [1.55-1.66]), followed by women from wealthy households (OR–CI, 1.46 [1.41–1.52]).[6]

IV. Methodology

For the analysis of the data, we have manually chosen all the related features from the dataset and have first pre-processed it, followed by analysis, based on which we have derived our conclusions. We have elaborated on our process and our findings in this and the subsequent sections.

A. Data Pre-processing

After manually selecting the most suitable features, we have first applied data cleaning, followed by sampling of the data. We have introduced a new feature called BMI (Body Mass Index), and have performed normalization on BMI and haemoglobin readings.

1) Data Cleaning:

After going through the existing literature around the prevalence of Cesarean Births under different conditions, we selected around 200 columns from the original data set that we thought were relevant to our problem. Then we had to clean the data in the selected columns, the first step of which was to replace all the empty spaces in the data set with np.nan values, making the analysis more convenient using NumPy [7]. Then, we had to convert the data type from 'object' to Float, to make calculations on the data possible. In some of the columns selected, the number of missing values was of a large proportion to the total number of values (>20%), therefore we had to drop those columns as handling the missing values by either ignoring those rows or by interpolating the values would have been detrimental to the veracity of the analysis. were left with 85 independent variables for our target variable, that is, the frequency of Cesarean births.

2) Sampling:

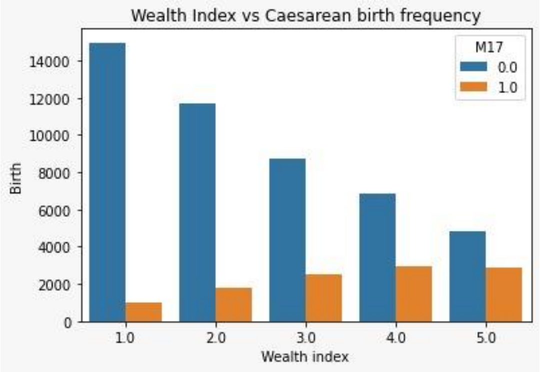
The data set that we were working with had a huge number of data points for each state, and to reduce the numerosity of the data, we applied simple random sampling to the data state-wise. We sampled 25% of the data from each state to give our final reduced dataset.

3) Normalization:

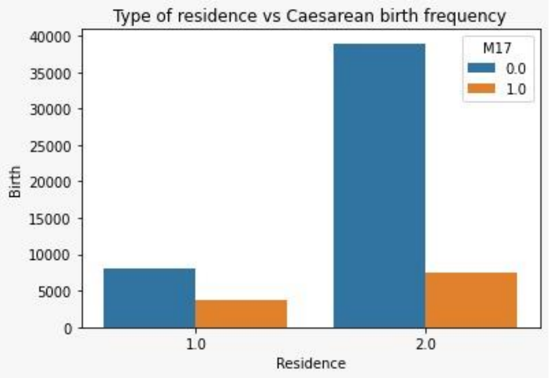
Due to the nature of the survey conducted for collecting the data, a lot of the variables were categorical, and to bring the continuous variables to a standard scale, to improve the results of further analyses, we normalized them using the min-max normalization technique. Some of the variables that we've normalized are age, height, weight, and so on.

B. Data Exploration

Count plots for the two variables most likely to affect the target variable were plotted based on our understanding of the literature- the residence type(rural/urban) and the wealth index of the subjects. The plots were made with the help of the Seaborn [8] (a Python data visualization library based on matplotlib.) The plots show that Urban and well-off women are significantly more likely to have Caesarean births compared to rural and economically backward women. This is consistent with the literature around the problem and provides a base to further analyse the finer details of the relationships of the target variable with the other factors.



A count plot depicting the relationship between the Wealth Index and the Caesarean birth frequency.



A count plot depicting the relationship between the type of residence and the Caesarean birth frequency.

C. Data Analysis

After selecting various socio-economic and health-related features based on existing literature to explain the birth type, the more relevant features were selected using the Chi-Square test and Pearson Correlation coefficients. After applying the Chi-Square test on the dataset, we took the first 20 features with the most significant correlation with the target variable. Now it is sometimes possible that this test misses out on some of the explanatory variables that significantly influence the target variable. To prevent this possibility, we took 20 variables ranked by their Pearson correlation coefficient and took the union of the said 40 features to arrive at our final set of features for further analysis. Some features in this set were redundant, while some indicated the same physical variable. To further reduce the number of features, we clubbed the markers that indicate the same variable to make a single feature; for example, exposure to TV, internet, and radio was clubbed into a single feature, exposure to media. After this, we had a set of 28 features for data analysis.

1) Logistic Regression

The first Classification technique we used was Multivariate Logistic Regression. This popular supervised machine learning technique effectively predicts binary target variables given a set of labelled datasets to train on. Here the binary target variable for the model was the birth type, and the independent variables were the columns we had selected. A Logistic Regression model gives weights corresponding to each independent variable, which act as coefficients to get a linear combination of tuples of the independent variables, which is a scalar value. If this value is lesser than a pre-defined threshold value, we label that particular tuple as a negative example and positive if the value exceeds the threshold. The optimal set of weights of the model is arrived upon by using the Gradient Descent Algorithm on the cost function, which is a function of the weight vector.

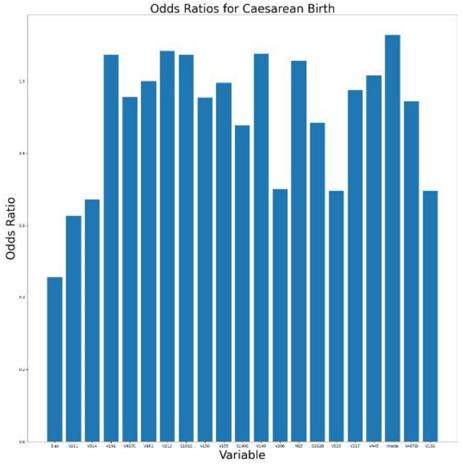
On applying the algorithm to our model, by splitting it into training and testing sets, we got a model for the data, which gave weights for all our features. We observed that the weightage of the previous birth type(Caesarean/Normal) feature was much more significant than the other features, having an odds ratio of 161.76. From this observation, we can infer that the odds of the following birth type being Caesarean are 161.76 times more likely if the previous birth type was Caesarean.

It was observed that the weightage in the model of the other variables, compared to the previous birth type, was significantly less. So, to find further finer details from the regression model, we trained another model where we omitted the last type of birth as a feature. In this model, we could see that no feature had as significant an influence as the previous birth type in the last model.

In the second model, the odds ratio and the weightage of the different variables indicate that the variables that had the strongest influence on the predictions were BMI, media exposure, no. of hospital visits, education level, wealth index, place of delivery, type of residence(rural or urban) and the variables that indicate lifestyle facilities that the subject woman possesses and some health parameters. We also plotted the partial dependence plots of the variables, indicating a positive or negative relationship between the target variable and the independent variables.

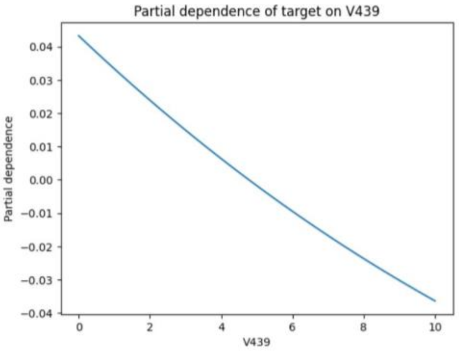
We can conclude from the regression analysis that the factors that most influence the birth type were the ones that indicate the physical health state of the woman and their socio-economic status. From the partial dependence plots, we conclude that the odds of the birth being Caesarean increase with an increase in BMI, wealth index of the woman and exposure to media. We also observe that the odds are higher for women living in urban areas. Another factor that had a large influence on the birth type was place of delivery, with the frequency of Caesarean births being much larger in private hospitals than in Government hospitals.

A graph depicting the odd ratios for Cesarean births:

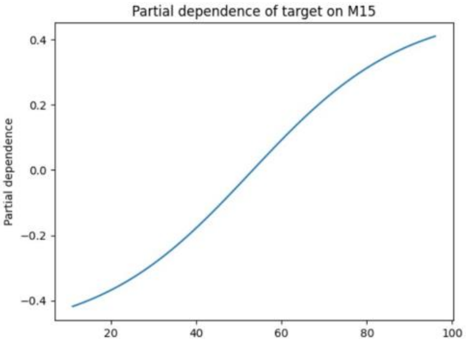


Visualization of Partial Dependence of a few features:

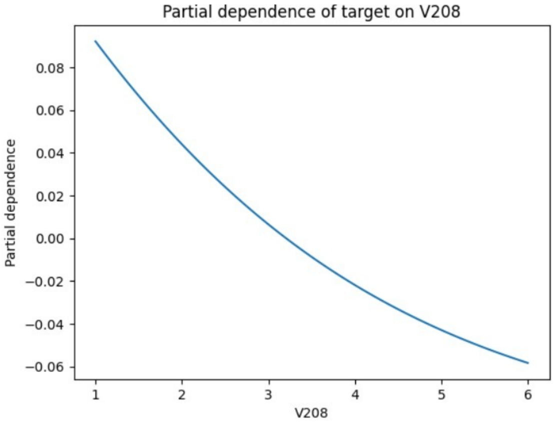
a) Height/Age Percentile



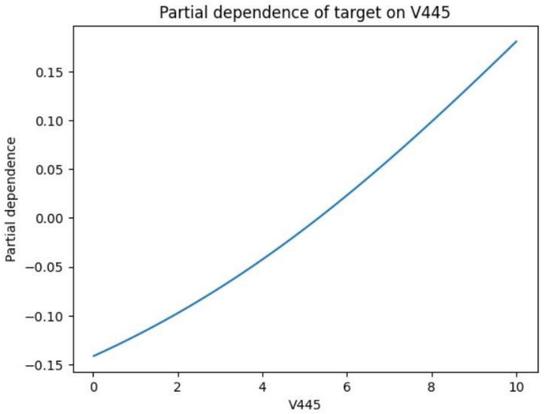
b) Place of delivery



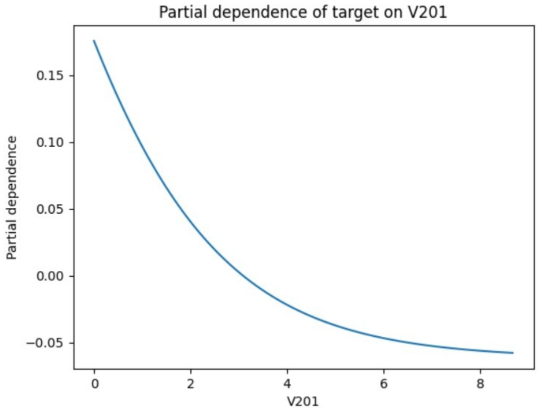
c) Births in last five years



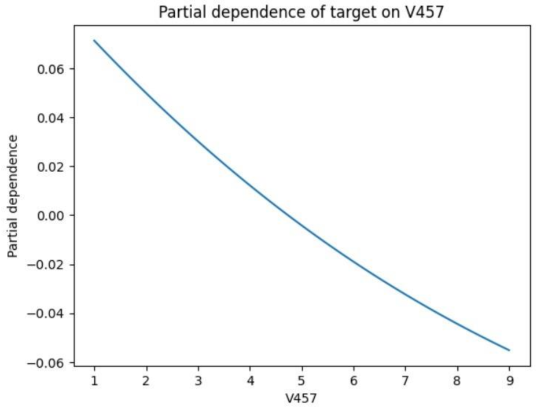
d) Body Mass Index (BMI)



e) Total children ever born



f) Anemia level



2) Naïve Bayes

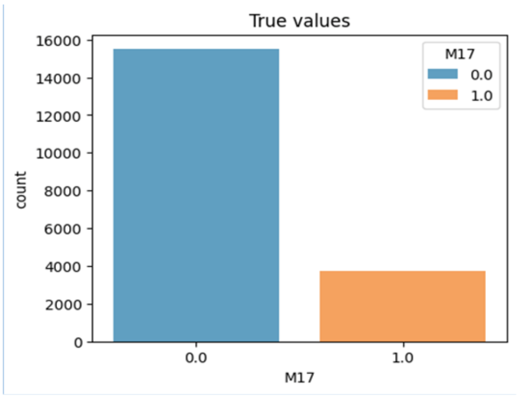
We then used a Naive Bayes classification algorithm to build a model that could predict whether a birth would result in a caesarean delivery or not based on various factors, such as maternal age, previous births, and wealth index

The dataset we used for our analysis was the National Family Health Survey (NFHS), which is a large and nationally representative survey conducted in India. We performed attribute subset selection to select the most relevant features for the classification task. This involved selecting the subset of features that had the most significant impact on the prediction accuracy, while also considering factors such as interpretability and computational efficiency.

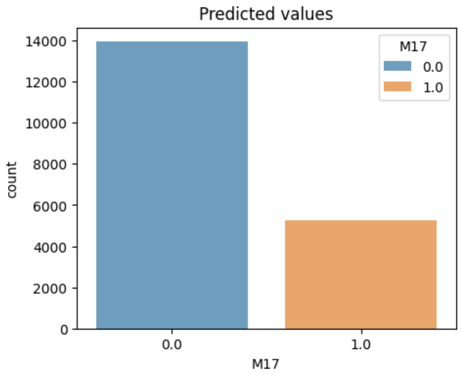
One advantage of using Naive Bayes in this context is that it is a simple and fast algorithm that can handle large datasets with high-dimensional feature spaces. Additionally, Naive Bayes is robust to noise and irrelevant features, making it a good choice for classification tasks with a lot of noisy or irrelevant data.

Our Naive Bayes classification model was able to predict caesarean births with a high degree of accuracy. Specifically, the model achieved an accuracy of 86%, indicating that the model is reliable and can be used to predict the likelihood of caesarean births in India.

In conclusion, Naive Bayes is a powerful and versatile classification algorithm that can be used to predict the outcomes of various real-world applications, including medical diagnosis. Our study demonstrates the effectiveness of Naive Bayes in predicting caesarean births and highlights its potential for use in other medical applications.



A count plot showing true values for Caesarean column, with 1 representing Caesarean and 0 representing Natural Birth.



A count plot showing predicted values for Caesarean column, with 1 representing Caesarean and 0 representing Natural Birth.

3) Other Methods

Apart from these methods, we have also tried to apply k-means clustering on specific columns of the dataset in order to derive inferences of the regular patterns in various clusters, but the results of clustering were not satisfactory and had low accuracy.

V. Conclusion

We have investigated a few methods to identify the elements that significantly influence caesarean births in India. Our study, which used data from the National Family Health Survey 2019–21, found significant correlations between the likelihood of caesarean births and maternal age, mother's education, region, place of a child's birth, mothers' wealth profiles, and the area of residence. The findings showed that urban residence, high educational attainment, and a better socioeconomic background were independently related with caesarean delivery in women. A proper understanding of how to prevent prenatal problems can also aid in lowering the likelihood of caesarean section malpractice. Qualitative study is also required to comprehend the cultural values, psychological aspects, and potential perspectives of Indian women.

As for further scope, there are several areas that could be explored in future research. Firstly, incorporating more advanced techniques such as ensemble methods (e.g., Random Forests, Gradient Boosting) or deep learning models (e.g., Neural Networks) could potentially improve the predictive performance of our models. Secondly, conducting more in-depth feature engineering, including feature selection, feature scaling, and feature transformation, could help optimize the models' performance. Additionally, exploring different evaluation metrics and performing cross-validation to assess the models' generalization ability could enhance the reliability of our results.

References

[1] A. Amjad, U. Amjad, R. Zakar, A. Usman, M. Z. Zakar, and F. Fischer, “Factors associated with caesarean deliveries among child-bearing women in Pakistan: Secondary Analysis of data from the demographic and Health Survey, 2012–13,” *BMC Pregnancy and Childbirth*, vol. 18, no. 1, 2018. [Accessed Mar. 20,2023]

[2] Ghosh, S and K S James (2010): "Levels and Trends in Caesarean Births: Cause for Concern?", Economic & Political Weekly. [Accessed Mar. 20,2023]

[3] B. Unnikrishnan, “Trends and indications for caesarean section in a tertiary care obstetric hospital in Coastal South India.,” *Australasian medical journal*, pp. 821–825, 2010. [Accessed Mar. 20,2023]

[4] S.Shabnam,"Caesarean section delivery in India: causes and concern"

[5] L. Bogg, V. Diwan, K. S. Vora, and A. DeCosta, “Impact of alternative maternal demand-side financial support programs in India on the caesarean section rates: Indications of supplier-induced demand,” *Maternal and Child Health Journal*, vol. 20, no. 1, pp. 11–15, 2015. [Accessed Mar. 20, 2023]

[6] V. Mishra, N. Roy, P. Mishra, V. Chattu, S. Varandani, and S. Batham, “Changing scenario of C-section delivery in India: Understanding the maternal health concern and its associated predictors,” *Journal of Family Medicine and Primary Care*, vol. 10, no. 11, p. 4182, 2021, doi: 10.4103/jfmpc.jfmpc_585_21. [Online]. Available: http://dx.doi.org/10.4103/jfmpc.jfmpc_585_21 [Accessed Mar. 20, 2023]

[7] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[8] “Statistical Data Visualization#,” *seaborn*. [Online]. Available: https://seaborn.pydata.org/#:~:text=Seaborn%20is%20a%20Python%20data,introductory%20notes%20or%20the%20paper . [Accessed: 21-Mar-2023].

[9] International Institute for Population Sciences (IIPS) and ICF. 2017. National Family Health Survey (NFHS-4), 2015-16: India. Mumbai: IIPS.

[10] K. Kalaivani and P. Ramachandran, “Interstate Differences in Institutional Delivery and Caesarean Section Rates in India,” *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4036695. [Online]. Available: http://dx.doi.org/10.2139/ssrn.4036695 [Accessed Mar. 20, 2023]

[11] [1]M. Bhatia, K. Banerjee, P. Dixit, and L. K. Dwivedi, “Assessment of Variation in Cesarean Delivery Rates Between Public and Private Health Facilities in India From 2005 to 2016,” *JAMA Network Open*, vol. 3, no. 8, p. e2015022, Aug. 2020, doi: 10.1001/jamanetworkopen.2020.15022. [Online]. Available: http://dx.doi.org/10.1001/jamanetworkopen.2020.15022 [Accessed Mar. 20, 2023]

[12] [1]A. Sengupta, M. Sabastin Sagayam, and T. Reja, “Increasing trend of C-section deliveries in India: A comparative analysis between southern states and rest of India,” *Sexual & Reproductive Healthcare*, vol. 28, p. 100608, Jun. 2021, doi: 10.1016/j.srhc.2021.100608. [Online]. Available: http://dx.doi.org/10.1016/j.srhc.2021.100608 [Accessed Mar. 20, 2023]

[13] [1]S. K. Mohanty, B. K. Panda, P. K. Khan, and P. Behera, “Out-of-pocket expenditure and correlates of caesarean births in public and private health centres in India,” *Social Science & Medicine*, vol. 224, pp. 45–57, Mar. 2019, doi: 10.1016/j.socscimed.2019.01.048. [Online].Available: http://dx.doi.org/10.1016/j.socscimed.2019.01.048 [Accessed Mar. 20, 2023]

[14] Roy, A., Paul, P., Chouhan, P. *et al.* Geographical variability and factors associated with caesarean section delivery in India: a comparative assessment of Bihar and Tamil Nadu. *BMC Public Health* **21**, 1715 (2021). https://doi.org/10.1186/s12889-021-11750-4

[15] A. Choudhary, S. Koditala, S. Anem, and N. Gunta, "Casaerean Births Analysis," https://colab.research.google.com/drive/1XqBzTEMRUHuoeYw_ZS4dpTj12-SZGinW?usp=sharing