MIE368: Analytics in Action
Final Report:

Maternal Mortality Intervention within Impoverished States

Group 7: [redacted] Nada Al Aker, [redacted]

1.Problem Statement:

Research Question: What is the optimal maternal health care intervention an impoverished country can implement in order to minimize their maternal mortality rate?

Context: As background, we refer to the World Health Organization's definition of maternal death, defined as "The death of a woman while pregnant or within 42 days of termination of pregnancy from any cause related to the pregnancy". [1] Every day, 810 women die from pregnancy or childbirth related causes. [2] Most of these deaths are preventable. With a discrepancy between resources and need, maternal death affects impoverished regions most significantly. With this motivation, the team analyzed multiple features that contribute to maternal health. We determined if a quantitative relationship exists between a country's maternal healthcare and maternal mortality rate. We then identified the single most significant area of intervention, with the most impactful application for countries considered 'Low-Middle Income' (LMIC). The scope of this analysis ranges 194 countries over 34 years.

2. Data Description:

For this analysis, the team referenced a collection of maternal and reproductive health data available from the World Health Organization's (WHO) "Global Health Observatory Data Repository"[3] and the World Bank Open Data Repository [4]. Over 20 datasets were examined; the final dataset is a combination of 8 datasets merged on the basis of "Country" and "Year" [5][6][7][8][9][10][11][12]. The dataset is comprised of 13 features and 1 target variable, as detailed in the data dictionary in Figure 1. Beyond these features, areas of maternal health, such as "Births by Caesarean Section" [13], "Female Genital Mutilation" [14], "Contraceptive Prevelance" [15] and "Women Married or in a Union Before Age 15" [16], were also explored. However, these attributes did not have satisfactory data available for analysis. This is notable as the team believes these, among other features, contribute to maternal health and are crucial to the development of a comprehensive model. As a result, for the purpose of analysis, we made the assumptions that the considered features provide a holistic representation of maternal health and the populations considered for each data feature are equivalently representative.

Feature	Definition
Country	Country name (String)- Not included in analysis, used for categorization
Year	Year of recorded data (Integer)
Prevalence of HIV	Percentage of women (15-24) who have HIV (%, Float)
Prevalence of Undernourishment	Percentage of total population considered undernourished (%, Float)
Current Health Expenditure	Percentage of Gross Domestic Product spent on healthcare (%, Float)
Adolescent Birth Rate	Percentage of adolescent women (aged 15-19) who gave birth (%, Float)
Antenatal Care Coverage	Percentage of pregnant women who received antenatal care 4+ times (%, Float)
Birth Attended by Health Personnel	Percentage of births which were supervised by skilled health personnel (%, Float)
Institutional births	Percentage of births which took place in a healthcare facility (%, Float)

Income Code	Income level of a country based on: {0:Low Income, 1:Low-Middle Income, 2: Middle-High income, 3: High Income} (Categorical Integer)
Sub Saharan African Region	If the data country of origin is from Sub-Saharan Africa (Binary Integer)
Condition: Less than 80% of Births attended by skilled health personnel	If the percentage of births attended by a skilled health personnel was above or below 80%, taking values 0 or 1 respectively (Binary Integer)
Condition: Less than a 5% adolescent birth rate	If the percentage of adolescent women (aged 15-19) who gave birth was above or below 5%, taking values 0 or 1 respectively (Binary Integer)
Condition: Less than 67% of births take place in a institutional facility	If the percentage of births taken palace in an institutional facility was above or below 67%, taking values of 0 or 1 respectively (Binary Integer)
Target: Maternal Mortality Ratio	**Percentage of births that resulted in maternal death (%, Float)**

Figure 1. Data Dictionary

3. Analysis Methodology Implementation

With respect to methodology, there were 4 major areas of this analysis: data cleaning, feature engineering, exploratory data analysis (EDA) and regression modelling. Extensive data cleaning was performed to develop a functional dataset. Then, feature engineering and EDA were conducted over several iterations. Next, 6 regression model design alternatives were implemented. Lastly, results analysis was conducted in 3 areas: a total performance analysis, geographical analysis and income level analysis. In this section, we detail each method, discussing their motivation, implementation and notable insights.

3.1 Data Cleaning

Over the course of two major iterations, the team investigated 20 datasets in total. Before the final dataset was created, extensive data cleaning was performed on each individual dataset. This was required as the individual datasets were presented in a very disorganized and non uniform manner.

First, a uniform indexing structure was implemented for the columns and rows. Then, each numerical feature was converted to a uniform format of *percentage.* The original data included values expressed in terms of "total sum","per 1000" or "per 100 000" of the population. Next, the datasets included unhashable data, like confidence intervals, in many feature columns. For the purpose of analysis, the unhashable data was omitted. However, a model accounting for these confidence intervals would create more accurate results. Lastly, the "Year" data was inconsistent throughout all the datasets, presented in multiple formats including overlapping intervals and singular year values. To combat this inconsistency, the *split_years(df)* function was developed and applied to each dataframe. This function separated every interval into individual year entries. Then for each set of country and year pairs, only one entry was stored, taking the *mean* of all corresponding entries. This data cleaning strategy introduces the assumption the dataset is representative of the empirical annual data.

Once data cleaning was applied, the datasets were merged. Through preliminary investigation, there was significant variability in the dataset sizes; certain datasets contained rows to the order of *1000's* while others were only to the order of *100's.* Through the merging of overlapping rows, the smaller datasets drastically limited the range of trainable data. To maintain sufficient data for analysis, the latter group of datasets were removed from the scope of analysis. After both major iterations, this resulted in the removal of 9 datasets. Based on the correlation coefficient analysis of the removed features, found in Appendix A, the removed features are *not* heavily correlated to the included features. This observation emphasizes the need to include these aspects of maternal healthcare in the model, as they are not sufficiently represented within the current range of features. This also emphasizes our team’s belief that the removed datasets would be crucial to the development of a more accurate and comprehensive model if more data were available.

3.2 Iteration 1: Feature Engineering

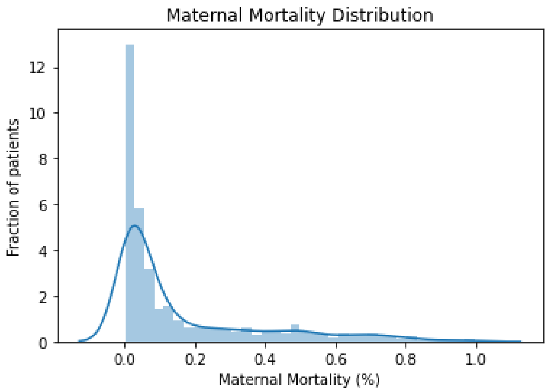
Once the cleaned data was merged, an initial analysis was conducted. The dataset had a total of 5402 rows and 10164 missing values. In other words, 33.96% of the values were missing and, more significantly, 47.55% of the numerical data was missing. With a high volume of missing data, removing missing entries would negatively impact the dataset. This signaled a need for data imputation.

First, we implemented strategic data dropping. We assessed the impact of removing rows based on 3 different thresholds of maximum NaN’s per row (4, 3 and 2 = maximum number of NaN’s). Through this assessment, 3 provided optimal results with 3283 rows and 6195 missing values. This threshold choice reduced the number of missing data points by 39%. Next, to impute the missing data, we grouped the data by country. Then, we attempted to impute the data based on each country’s central tendency measures, namely the mean and median. When not enough data was available, we imputed based on the central tendency measures of the entire column. Through comparing the central tendency measures against the mean, median, range and standard deviation of each column, the mean was consistently more representative. Thus, the team implemented the replacement with the mean. It should be noted that more accurate results can be drawn by implementing an imputation strategy for each column, as certain features are better contextually expressed with different measures.

3.3 Iteration 1: Exploratory Data Analysis

In order to highlight inconsistencies and gather supporting evidence, the team conducted the first round of exploratory data analysis. Through a statistical summary, the feature means and medians were more heavily skewed towards the maximum, as seen in Appendix B. Meanwhile, the target variable was more heavily skewed towards the minimum, highlighted by the histogram in Figure 2. Based on these results, our team concluded a need to implement balanced class weights in our regression modelling.

Furthermore, through plotting histograms for each feature, a relationship between maternal healthcare and maternal mortality is supported. The histograms plot the frequency of maternal mortality for each feature. Each feature is separated in two categories, “above the feature average” and “below the feature average”. The numerical value of the average is not specifically significant in this visualization. Rather, the average is used to demonstrate a trend; distinguishing maternal mortality in the presence of maternal healthcare versus its absence. Each histogram follows the same general trend displayed by Figures 3, with further examples in Appendix C. Beyond contextual consistency, the results demonstrate factors of maternal healthcare have a notable correlation to maternal mortality. The example in Figure 3 shows when the adolescent birth rate is below the average, maternal mortality ratio is also low. However, as the adolescent birth rate exceeds the average, the mortality rate increases in variability and value.



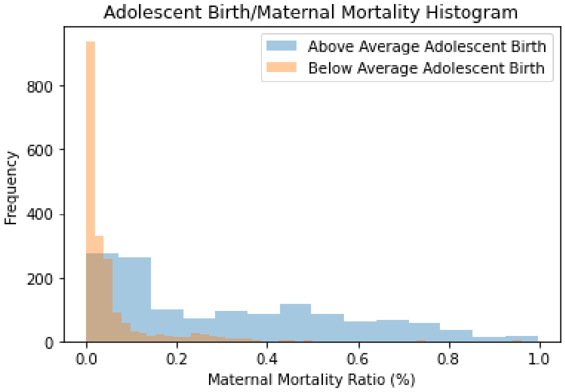


Figure 2. Maternal Mortality Histogram

Figure 3. Maternal Mortality vs. Adolescent Birth

Once EDA was completed, a basic linear regression model was designed as a benchmarking tool. Our training and testing data split was done on a standard 7:3 ratio[17]. The models are scored based on the coefficient of determination, R^2, measuring how close the data points are to the regression line. The training and testing scores were respectively 0.563 and 0.570, signifying needed improvement. As a result, our team conducted a second major iteration of feature engineering and EDA.

3.4 Iteration 2: Feature Engineering

In the second feature engineering iteration, 2 feature generation techniques were employed to create a more well rounded representation of maternal health. First, further data collection was done, implementing new features. Two of the added string features, “Income Level” and “Geographical Region”, were converted to categorical data. The income code was stored as one categorical feature, taking values [0-3] based on the following map, { 0: Low Income, 1: Low-Middle Income, 2: Middle-High income, 3: High Income}. The geographical region data was converted to 7 binary variables. This categorization differentiation is to account for the future contextual interpretability of the model coefficients. The income codes have an increasing correlation while the regional data is distinct.

Next, new features were generated by the manipulation of a CART model to capture non-linear relationships. The CART model held a maximum depth of 3 levels and a minimum number of 2 samples within a split, determined heuristicly through benchmarking with MIE368 lab material [17][18]. Through the model splits in Appendix D and the relative feature importance in Appendix E, the 4 most impactful conditions were determined and implemented as binary features in the final dataset. The other regional features were also dropped from the final dataset as only the African regional feature displayed significant importance. The 4 most important conditions are as follows:

- ➢ Rate of Births Attended by a Skilled Health Professional <= 79.625 %
- ➢ Data from the Region of Sub-Saharan Africa
- ➢ Rate of Institutional Births <= 67.1 %
- ➢ Adolescent Birth Rate (women aged 15-19) <= 4.587%

3.5 Iteration 2: Exploratory Data Analysis

A second iteration of EDA was then conducted to ensure the result quality had not decreased. Then, the team further investigated the role of the new categorical features. Among other observations, the distribution of mean maternal mortality rate categorized by income level was of interest. The distribution is as follows: { Low Income: 0.433%, Low-Middle Income: 0.278%, Middle-High income: 0.0744%, High Income: 0.0201%}. This is also shown graphically in Figure 4, on page 5. This correlation shows a decreasing trend. As a nation’s economic strength increases, their maternal mortality rate decreases; supporting a meaningful relationship between a nation's income level and maternal mortality rate.

Lastly, a correlation matrix is generated and analyzed in order to assess the relationship between the features. The correlation matrix heatmap can be found in Appendix F. From the correlation matrix, it can be observed that the engineered CART model conditions are highly correlated with the real corresponding features. It should be noted, this is expected and will impact the interpretability of the final model coefficients. If there are too many engineered features, the model may have a difficult time assessing which feature is actually contributing to the target. Beyond these features, there are also a selection of other features which are correlated above the suitable threshold of 0.7. It is also important to mention that the objective of this analysis is to assess which maternal healthcare factors impact maternal mortality most significantly, however a high correlation does not necessarily imply causation. Further investigation in this respect is required. As a result of the moderately high level of correlation, a “K-Best Features” model will be developed and assessed in order to reduce the risk of multicollinearity.

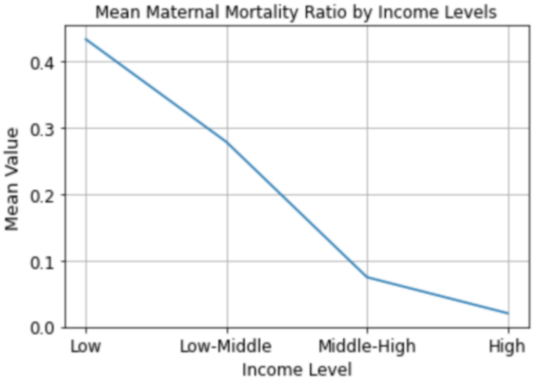
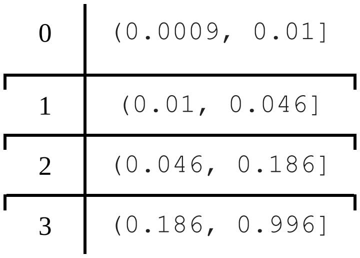


Figure 4. Average Maternal Mortality by Income Level

3.6 Regression Modelling

With the final dataset, the team implemented 2 base regression models and 4 design extensions to assess the final performance. Following the baseline model, our models are scored based on R^2. For this analysis, the threshold of a statistically meaningful score has been defined as an R^2 value of 0.7 or greater, characterizing a strong relationship between the regression model and data. This is based on benchmarking against similar research [19]. The two base models developed are a linear regression model and a multinomial logistic regression model. As the model has a relatively small number of features, regularization was not utilized in the linear regression model. With respect to the multinomial logistic regression, the maternal mortality target variable was split into categories. A categorization based on quartiles was employed in order to account for an even and ample distribution of data points. The quartile ranges and associated categorical codes are shown in Figure 5. In Figure 5, it can be observed that due to the high density in *low* maternal mortality rates, the interval sizes are very uneven in width. Other strategies such as intervals of equal width, increased splits, or deciles can also be investigated to assess the changes in performance.



0	(0.0009, 0.01]
1	(0.01, 0.046]
2	(0.046, 0.186]
3	(0.186, 0.996]

Figure 5. Maternal Mortality Quartile Ranges and Respective Integer Category

The multinomial logistic regression model also implements balanced class weights as previously defined in the EDA. Additionally, a standard approach of 5-fold cross validation is employed in order to select the model hyperparameters. Lastly, through investigation, it has been found that a minimum of 10 000 iterations must be implemented for a convergence of the model score.

After testing the base models, two modifications were implemented to each base model. First, data scaling was incorporated. With the addition of several categorical features, this step is important because there is an increased variability in the order of magnitude of the features. Next, a feature selection model was added, extending the first modification. As recognized by EDA, the feature variables are moderately correlated. In order to limit multicollinearity, the Analysis of Variance (ANOVA) F-value is evaluated for the best 10 features; this number of features has been heuristically set in order to maintain a high number of features. The performance summary of all 6 models is displayed below in Figure 6.

Model		**Training Score (%)**	**Testing Score (%)**
A	Base: Multinomial Logistic Regression	73.8	70.9
A.1	Scaled Log. Regression	74.2	70.1
A.2	**Feature & Scaled Log. Regression**	**74.2**	**71.4**
B	Base: Linear Regression	70.0	65.3
B.1	Scaled Lin. Regression	64.1	57.6
B.2	Feature and Scaled Lin. Regression	61.0	55.0

Figure 6. Regression Model Design Alternatives Scoring Summary

4. Results and Analysis

The final model performance is now measured. First, we assess if there is a statistically significant linear relationship between maternal healthcare and maternal mortality. Then, we assess the most meaningful features contributing to maternal mortality. This is with the objective to identify the most impactful intervention a “Low-Middle Income” country can implement to most significantly reduce maternal mortality rates. As observed in Figure 6, the highest performing model is A.2, the multinomial regression model with data scaling and feature selection. This model has a testing score of 0.714. As the model scores above 0.7, this model shows a strong relationship with potential for further development. A total performance is assessed. Then a geography-based, and income-based sensitivity analysis were each done.

4.1 Total Performance Analysis

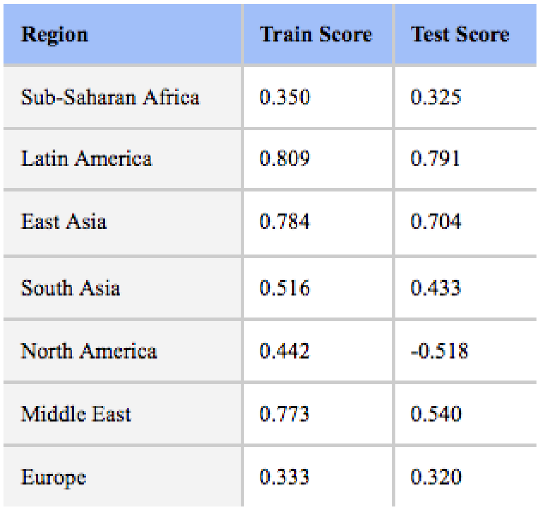
Although model B.2 has the lowest performance, it is referenced for all 3 analyses. This is as a linear regression model allows for better contextual coefficient interpretation. It is hypothesized this model performs poorly as it's more difficult to infer a continuous target rather than a categorical target. Note, as data scaling has been implemented, the comparison of feature coefficients is assumed as appropriate[20]. The final equation representing the *total model* and the 10 most significant features are as follows:

$$Y = 0.025X_1 + 0.011X_2 + 0.00037X_3 - 0.0038X_4 + 0.017X_5 - 0.339X_6 + 0.137X_7 + 0.849X_8 - 0.759X_9 + 0.157X_{10} + 0.931$$

X_1 : Prevalence of HIV (%)
X_2 : Prevalence of Undernourishment (%)
X_3 : Antenatal care coverage - at least four visits (%)
X_4 : Institutional Births (birth taken place in a facility) (%)
X_5 : Births attended by skilled health personnel (%)
X_6 : Income Code
X_7 : Region Sub-Saharan Africa
X_8 : Less than 80% Births Attended by Skilled Health Personnel Condition
X_9 : Less than 5% Adolescent Birth Rate (Women Aged 15-19) Condition
X_{10} : Less than 67% Institutional Births Condition
As a result, the 3 most impactful features contributing to maternal mortality are the Skilled Health Professional Condition (ie. less than 80% of births with a health professional), the Adolescent Birth Rate Condition (ie. less than 5% adolescent birth rate for women aged 15-19), and the Income Code.

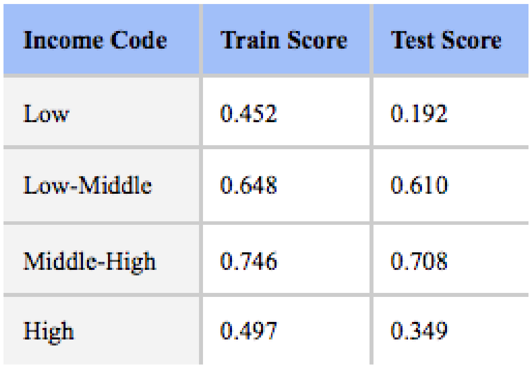
4,2 Geographical Analysis

Next, the geographical analysis was conducted by evaluating the model performance separately for each geographical region and contrasting the results. A summary of the scores are shown in Figure 7. The model performed poorly for some regions and very well for others. This is hypothesized to be attributed to the variability in income levels and the quantity of data. As shown in Appendix G, the regions which combined a well distributed balance of country income levels, as well as a high number of data points, performed most successfully. While investigating the European region, in addition to low income level variability, there is also much smaller variance in the other features as shown in Appendix H. Conversely, regions of Latin America and East Asia perform very well. They notably provide a combination of income level variability and a large quantity of data. For both of these high performing regions, births attended by health professionals is the 1st or 2nd most impactful feature, in line with the general model results. This is shown in Appendix I. It is significant to note the region of Latin America has achieved a score of 0.791, higher than the general model; this is concluded to be just attributed to the specific data distribution. Alternatively, the region of North America has a negative testing score of -0.518. This not only reflects the model is not appropriate for this dataset, but that it is a worse prediction than a horizontal line, or random chance. This juxtaposition reflects the importance of data quantity and variability.



Region	Train Score	Test Score
Sub-Saharan Africa	0.350	0.325
Latin America	0.809	0.791
East Asia	0.784	0.704
South Asia	0.516	0.433
North America	0.442	-0.518
Middle East	0.773	0.540
Europe	0.333	0.320

Figure 7. Geographical Analysis Summary



Income Code	Train Score	Test Score
Low	0.452	0.192
Low-Middle	0.648	0.610
Middle-High	0.746	0.708
High	0.497	0.349

Figure 8. Income Analysis Summary

4.3 Income Level Analysis

The income analysis conclusions are very similar to the geographical analysis. The dataset was separated based on each income level and each segment was tested against the model individually. Similarly to the geographical analysis, the model performed well for certain regions and poorly for others. The highest performing datasets are "Low-Middle" and "Middle-High" Income countries. This observation is justified through a contextual lens. Countries with mid-range income levels are more likely to have larger variance in healthcare services; this is as they are neither completely impoverished, nor completely developed. Countries on the extremities of income level, as in "High" or "Low" Income countries, will demonstrate much lower variance in health care services. In these countries, the majority of the population is more significantly affected by the national income level, either with or without access to appropriate health care respectively [21]. This is loosely quantitatively demonstrated through a summary of maternal mortality variance based on Income level in Appendix J. With respect to feature importance, births attended by health professionals is also the 1st or 2nd most impactful feature, shown again in Appendix I

5. Reflections and Next Steps

Based on the model coefficients in each analysis, the binary feature of "Less Than 80% of Births with a Skilled Health Professional" is the most impactful in predicting maternal mortality. However, maternal mortality is much more complex than the 10 features we have presented, graphically shown in Appendix K. Many factors contributing to maternal mortality can not be easily translated to quantifiable metrics. As a result, other socio-economic and political maternal mortality risk factors must be explored to strengthen this model. This analysis involves a selection of assumptions outlined as the scope has been narrowed by many limitations. The major limitation has been a lack of data, there have been a variety of interesting variables which were omitted due to insufficient data. Furthermore, this has impacted the model accuracy due to the high quantity of data imputation as well as the non-uniform data populations.

Moving forward, there are several interesting areas to investigate further to optimize the model performance. Firstly, based on several aspects of the EDA, the relationship between the feature and target variables appears to follow an exponential curve, rather than a linear curve. From this observation, it will be interesting to investigate if a different distribution will model the data more accurately, if at all possible. This is potentially why a logistic regression model is a more accurate inference than a linear regression model. Secondly, as the region of Sub-Saharan Africa has the highest quantity of Low & Low-Middle income countries, which experience the highest rates of maternal mortality, it will be valuable to develop a personalized model for this region. To develop this model, a further investigation into the source of our missing data will be required. A comparison between which regions had the largest quantity of data imputation contrasted against the model scores will reveal weakness areas of the model. After further data collection, research should be conducted to determine what makes the healthcare system in low income sub-saharan african countries unique. This is with the goal to introduce wider variability in feature values, which is correlated to high model performance. Through these future steps, the intervention recommendation will be more specifically applicable to Low-Middle income countries.

6.Conclusion

In summary, our team has explored the relationship between maternal health care and maternal mortality. We have gathered maternal healthcare data from over 20 datasets and incorporated extensive feature engineering in order to extract meaningful characteristics of maternal healthcare. We have then implemented 6 different regression strategies, resulting in a high performing model with a score of 71.4. Our results reflect a statistically significant relationship, stronger in regions with variable income levels. From our combination of analyses, to most significantly reduce maternal mortality, an LMIC intervention should be targeted in achieving over 80% of births with a skilled health professional in attendance.

References

[1] Wikipedia, “Maternal death,” (2020, October 05). en.wikipedia.org. [Online]. Available: https://en.wikipedia.org/wiki/Maternal_death. [Accessed Oct. 09, 2020]

[2] World Health Organization, “Maternal Mortality,” who.int, Sep. 19, 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/maternal-mortality. [Accessed Oct. 09, 2020]

[3] World Health Organization, “The Global Health Observatory,” who.int, [Online]. Available: https://www.who.int/data/gho. [Accessed Nov. 15, 2020]

[4] The World Bank Group, “World Bank Open Data,” data.worldbank.org. [Online]. Available: https://data.worldbank.org. [Accessed Nov. 15, 2020].

[5] The World Bank, “Prevalence of HIV, female (% ages 15-24), Regions and Countries Income Levels,” SH.HIV.1524.FE.ZS datasheet, 1990 [Revised 2019]

[6] The World Bank, “Prevalence of undernourishment (% of population),” SN.ITK.DEFC.ZS datasheet, 2000 [Revised 2018]

[7] The World Bank, “Current health expenditure (% of GDP),” SH.XPD.CHEX.GD.ZS datasheet, 2000 [Revised 2017]

[8] World Health Organization, “Adolescent birth rate (per 1000 women aged 15-19 years),” node.main.REPADO39?lang=en datasheet, 2000 [Revised 2020]

[9] World Health Organization, “Antenatal care coverage - at least four visits (%),” node.main.ANTENATALCARECOVERAGE4?lang=en datasheet, 1985 [Revised 2020]

[10] World Health Organization, “Births attended by skilled health personnel (%),” node.main.SKILLED BIRTH ATTENDANTS?lang=en datasheet, 2000 [Revised 2020]

[11] World Health Organization, “Institutional births (%),” view.main.SRHIBv datasheet, 1985 [Revised Oct. 2020]

[12] World Health Organization, “Maternal mortality ratio (per 100 000 live births),” node.main.15?lang=en datasheet, 2000 [Revised 2020]

[13] World Health Organization, “Births by Caesarean Section,” node.main.BIRTHSBYCAESAREAN?lang=en datasheet, 1990 [Revised 2020]

[14] World Health Organization, “Female Genital Mutilation,” node.main.FGM?lang=en datasheet, 2000 [Revised 2020]

[15] World Health Organization, “Contraceptive Prevalence,” view.main.HEMRHI_AGEv datasheet, 2000 [Revised 2020]

[16] World Health Organization, “Women Married or in a Union Before Age 15,” node.main.CHILDMARRIAGE?lang=en datasheet, 1995 [Revised 2020]

[17] T. Chan. MIE 368. Class Lab 5, Topic: “Model Engineering.” Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Oct. 28, 2020

[18] T. Chan. MIE 368. Class Lab 2, Topic: “Classification and Regression Trees (CART) and Random Forest.” Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Sep. 30, 2020

[19] M. I. M. Salleh, “What is the acceptable R-squared in the information system research?,” *ResearchGate*, 18-Jan-2016. [Online]. Available: https://www.researchgate.net/post/What-is-the-acceptable-R-squared-in-the-information-system-research-Can-you-provide-some-references. [Accessed: 09-Dec-2020].

[20]siamiisiamii 1, Nick Cox, Monica Reinstate, “Should you ever standardise binary variables?,” *Cross Validated*, 01-Jul-1962. [Online]. Available: https://stats.stackexchange.com/questions/59392/should-you-ever-standardise-binary-variables. [Accessed: 09-Dec-2020].

[21]C. G. Orach, “Health equity: challenges in low income countries,” *African health sciences*, Oct-2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2877288/. [Accessed: 09-Dec-2020].

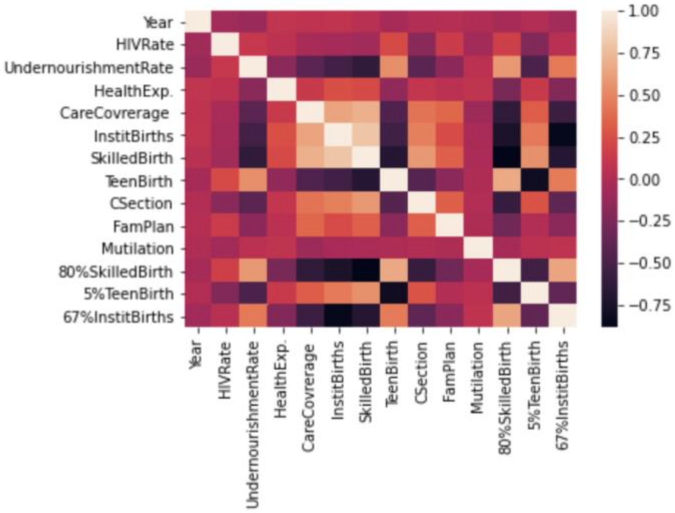
[22]Google Colab Link: https://colab.research.google.com/drive/11WKVq5V_vhMpD68kjWFa_ueQ6rXhMbkE?usp=sharing

[23]“Maternal mortality,” *World Health Organization*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/maternal-mortality. [Accessed: 09-Dec-2020].

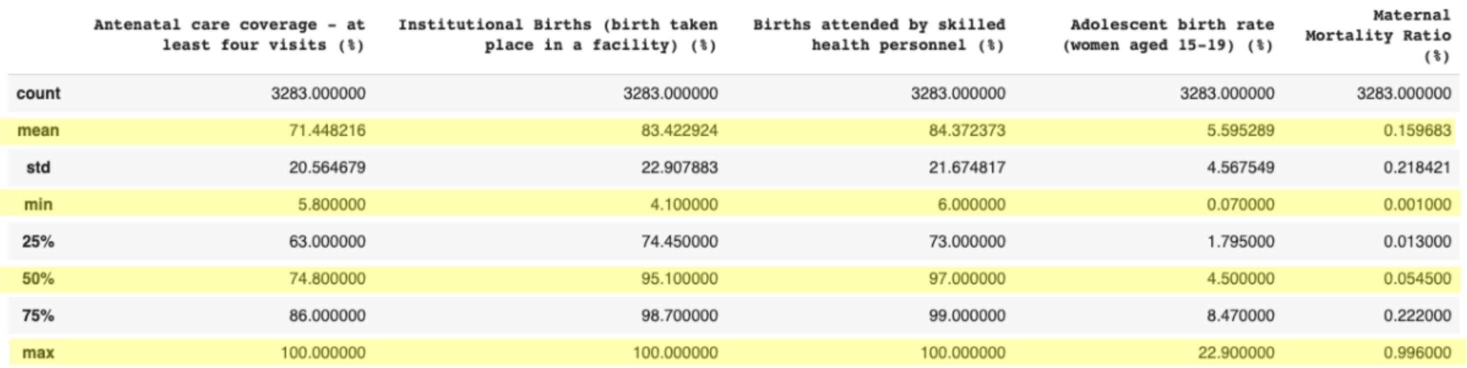
Appendix

Appendix A: Correlation Analysis of Removed Features

The removed features are labelled as “CSection”, “FamPlan” and “Mutilation”:



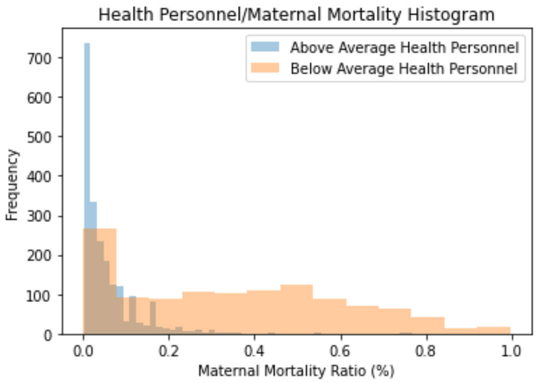
Appendix B: Overview of Feature Measures of Central Tendency

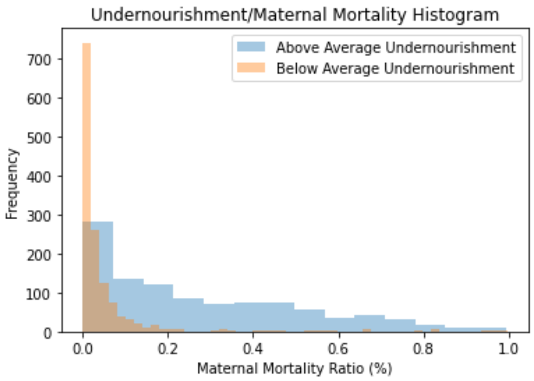


	Antenatal care coverage - at least four visits (%)	Institutional Births (birth taken place in a facility) (%)	Births attended by skilled health personnel (%)	Adolescent birth rate (women aged 15-19) (%)	Maternal Mortality Ratio (%)
count	3283.000000	3283.000000	3283.000000	3283.000000	3283.000000
mean	71.448216	83.422924	84.372373	5.595289	0.159683
std	20.564679	22.907883	21.674817	4.567549	0.218421
min	5.800000	4.100000	6.000000	0.070000	0.001000
25%	63.000000	74.450000	73.000000	1.795000	0.013000
50%	74.800000	95.100000	97.000000	4.500000	0.054500
75%	86.000000	98.700000	99.000000	8.470000	0.222000
max	100.000000	100.000000	100.000000	22.900000	0.996000

Appendix C: Selection of Histograms of Maternal Mortality Frequency by Feature

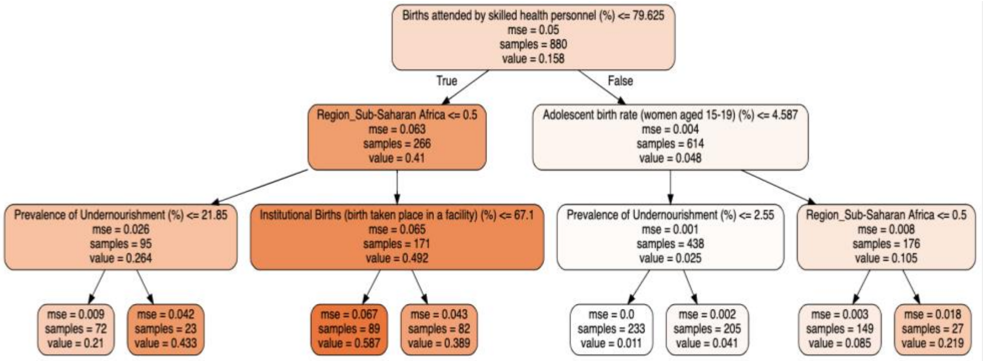
Further examples of the trends showcased in Section 3.3:



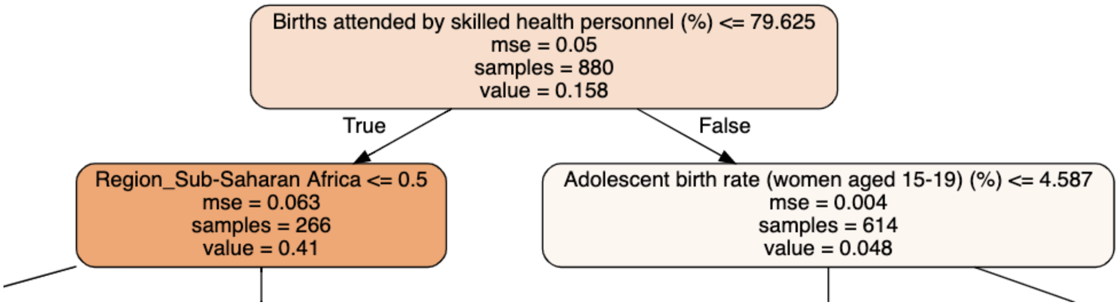


Appendix D: Visualization of CART Model Splits

A visualization of the general CART Model shape (titles negligible):



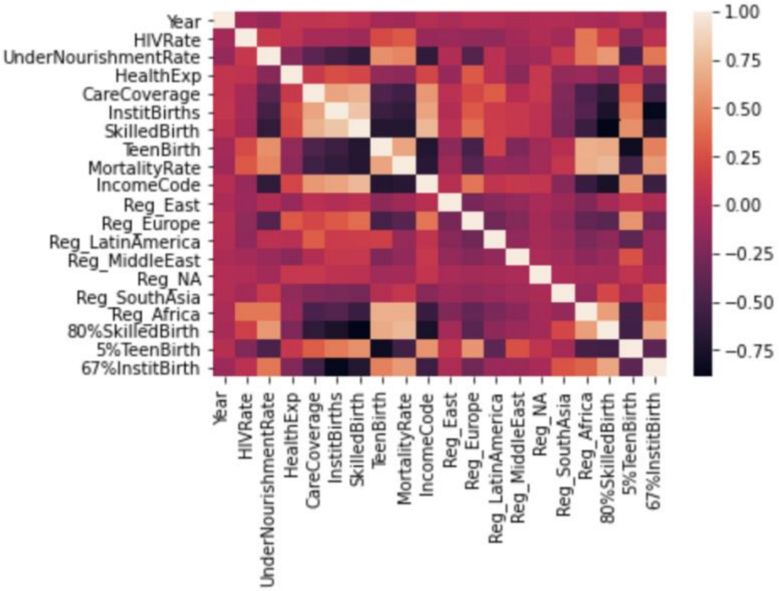
Based on this model, the 3 most important splits are shown below:


Births attended by skilled health personnel (%) <= 79.625
mse = 0.05
samples = 880
value = 0.158
True
False
Region_Sub-Saharan Africa <= 0.5
mse = 0.063
samples = 266
value = 0.41
Adolescent birth rate (women aged 15-19) (%) <= 4.587
mse = 0.004
samples = 614
value = 0.048

Appendix E: Relative Importance of CART Model Splits

Feature Name	Relative Importance
Births attended by skilled health personnel (%)	0.776422
Region_Sub-Saharan Africa	0.114290
Institutional Births	0.052869
Adolescent birth rate (women aged 15-19) (%)	0.025731

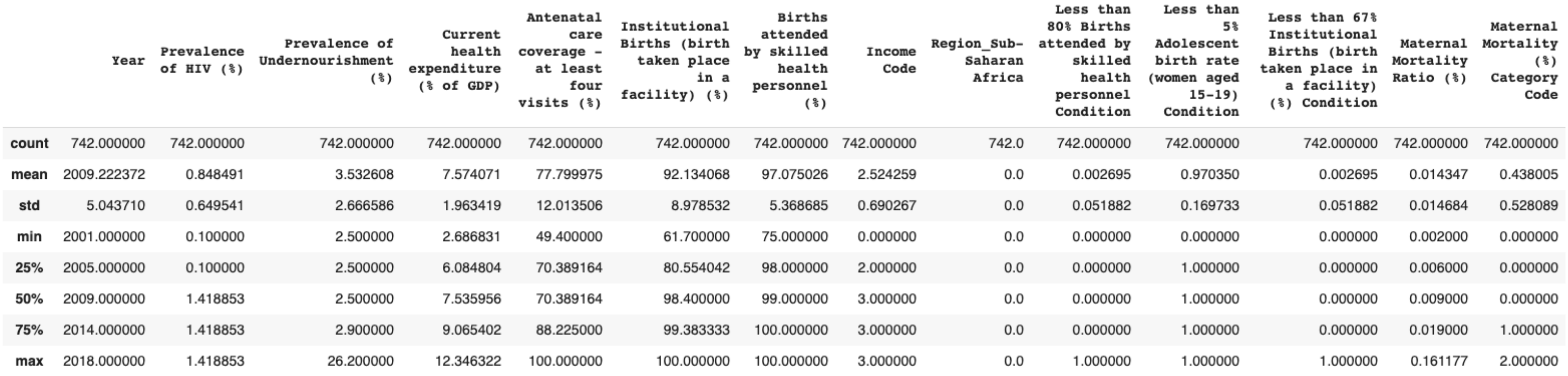
Appendix F: Final Dataset Correlation Matrix Heat Map



Appendix G: Geography Analysis- Income Level Count Distribution

	Sub-Saharan Africa	Latin America	East Asia	South Asia	North America	Middle East	Europe
Low Income Data	263	17	0	14	0	6	17
M-L Income Data	206	51	125	96	0	51	33
M-H Income Data	54	262	124	17	0	35	236
High Income Data	34	98	109	0	18	123	456
Total # of Data	557	428	358	127	18	215	742
Testing Score	0.325	0.791	0.704	0.433	-0.518	0.540	0.320

Appendix H: Europe Region Statistical Summary



	Year	Prevalence of HIV (%)	Prevalence of Undernourishment (%)	Current health expenditure (% of GDP)	Antenatal care coverage - at least four visits (%)	Institutional Births (birth taken place in a facility) (%)	Births attended by skilled health personnel (%)	Income Code	Region_Sub-Saharan Africa	Less than 80% Births attended by skilled health personnel Condition	Less than 5% Adolescent birth rate (women aged 15-19) Condition	Less than 67% Institutional Births (birth taken place in a facility) (%) Condition	Maternal Mortality Ratio (%)	Maternal Mortality (%) Category Code
count	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.000000	742.0	742.000000	742.000000	742.000000	742.000000	742.000000
mean	2009.222372	0.848491	3.532608	7.574071	77.799975	92.134068	97.075026	2.524259	0.0	0.002695	0.970350	0.002695	0.014347	0.438005
std	5.043710	0.649541	2.666586	1.963419	12.013506	8.978532	5.368685	0.690267	0.0	0.051882	0.169733	0.051882	0.014684	0.528089
min	2001.000000	0.100000	2.500000	2.686831	49.400000	61.700000	75.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.002000	0.000000
25%	2005.000000	0.100000	2.500000	6.084804	70.389164	80.554042	98.000000	2.000000	0.0	0.000000	1.000000	0.000000	0.006000	0.000000
50%	2009.000000	1.418853	2.500000	7.535956	70.389164	98.400000	99.000000	3.000000	0.0	0.000000	1.000000	0.000000	0.009000	0.000000
75%	2014.000000	1.418853	2.900000	9.065402	88.225000	99.383333	100.000000	3.000000	0.0	0.000000	1.000000	0.000000	0.019000	1.000000
max	2018.000000	1.418853	26.200000	12.346322	100.000000	100.000000	100.000000	3.000000	0.0	1.000000	1.000000	1.000000	0.161177	2.000000

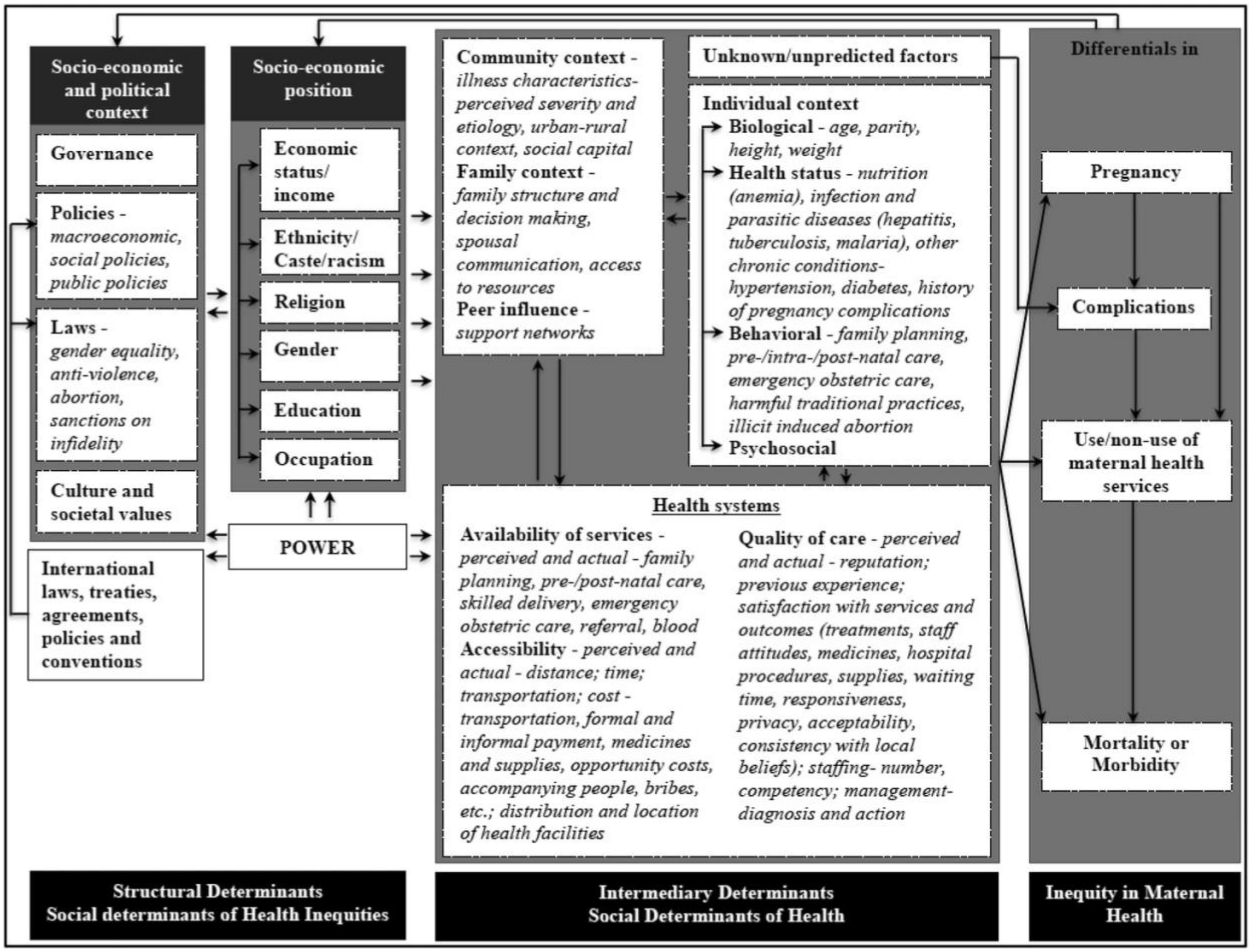
Appendix I: Summary of Best Performing Models and their Coefficients

Best Performing Models Name	HIV	Undernour.	Health Exp.	Antenatal Care	Inst. Births	Skilled Birth	Income Code	<80% Skilled Birth	<5% Adolescent	<67% Inst. Births
East-Asia	0.04	0.007	0.003	-0.001	-0.002	0.0002	0.018	0.064	-0.040	-0.050
Latin-America	-0.0003	0.009	0.002	-0.003	0	0.0008	-0.005	-0.025	-0.138	0.016
L-M Income	-0.003	0.008	0.002	0.004	-0.002	0.00035	-0.0005	0.228	0.088	0.036
M-H Income	-0.0008	-0.004	0.002	-0.000670	0	-0.002	0.182	-0.078	-0.039	0.058

Appendix J: Maternal Mortality Variance by Income Level

Income Level	Maternal Mortality Ratio Standard Deviation
Low Income(0)	1.186
Middle-Low Income(1)	0.704
Middle-High Income(2)	0.765
High(3)	0.693

Appendix K: Socio-Economic Factors Contributing to Maternal Mortality


Socio-economic and political context
Governance
Policies - macroeconomic, social policies, public policies
Laws - gender equality, anti-violence, abortion, sanctions on infidelity
Culture and societal values
International laws, treaties, agreements, policies and conventions
Socio-economic position
Economic status/ income
Ethnicity/ Caste/racism
Religion
Gender
Education
Occupation
POWER
Community context - illness characteristics- perceived severity and etiology, urban-rural context, social capital
Family context - family structure and decision making, spousal communication, access to resources
Peer influence - support networks
Unknown/unpredicted factors
Individual context
Biological - age, parity, height, weight
Health status - nutrition (anemia), infection and parasitic diseases (hepatitis, tuberculosis, malaria), other chronic conditions- hypertension, diabetes, history of pregnancy complications
Behavioral - family planning, pre-/intra-/post-natal care, emergency obstetric care, harmful traditional practices, illicit induced abortion
Psychosocial
Health systems
Availability of services - perceived and actual - family planning, pre-/post-natal care, skilled delivery, emergency obstetric care, referral, blood
Accessibility - perceived and actual - distance; time; transportation; cost - transportation, formal and informal payment, medicines and supplies, opportunity costs, accompanying people, bribes, etc.; distribution and location of health facilities
Quality of care - perceived and actual - reputation; previous experience; satisfaction with services and outcomes (treatments, staff attitudes, medicines, hospital procedures, supplies, waiting time, responsiveness, privacy, acceptability, consistency with local beliefs); staffing- number, competency; management- diagnosis and action
Differentials in
Pregnancy
Complications
Use/non-use of maternal health services
Mortality or Morbidity
Structural Determinants
Social determinants of Health Inequities
Intermediary Determinants
Social Determinants of Health
Inequity in Maternal Health

Source Reference found in Reference list [23]

Appendix L: Google Colab Link of Complete Analysis Script

https://colab.research.google.com/drive/11WKVq5V_vhMpD68kjWFa_ueQ6rXhMbkE?usp=sharing

*Note: There is also a comprehensive Table of Contents in the Colab Python file for easy navigation.