DATA 22100: FINAL PROJECT

GABRIEL REYES ESCLASANS, FEDERICO GUGLIELMOTTI, GISELLE MIRALLES

Abstract. In this paper, we explore the "Adding it Up" All Women Dataset from the Guttmacher Institute. This dataset provides statistics on a variety of factors relating to women's reproductive health, including abortion rates, healthcare costs, and more. It includes statistics for over 100 countries. This report explores a subset of the dataset's features. In addition, 2 machine learning models, Logistic Regression and Neural Networks, were created to predictively classify a country's continent based on its reproductive health statistics. We see that both models have difficulty predicting countries that belong to Oceania and hypothesize that this is due to a limit of data within this region. We also observe that the models are weaker when predicting the continent of Asian countries and speculate that this is due to substantial variation in features among countries in Asia. We conclude that both multi-class classification models perform similarly and both have an average accuracy above 85%.

Contents

1. The Dataset

The dataset used in this write-up comes from the "Adding It Up" [1] dataset. It is part of the 2019 "Adding It Up" project [2] by the Guttmacher Institute, a non-governmental organization advocating for sexual health and the expansion of reproductive rights worldwide. The project aims to illustrate the healthcare needs of all women of reproductive age in low- and middle-income countries and illustrate the scale of investment required to satisfy them. The "Adding It Up" dataset aggregates and draws from datasets released by international organizations or research institutions such as UNFPA, IIP at Johns Hopkins, and WHO, with data from 2015 to 2019. The dataset consists of a single Excel sheet whose rows are low- and middle-income countries and whose columns are many features relating to maternal healthcare needs and outcomes for each country (e.g. the cost of an abortion in USD, or the percentage of expectant mothers attending more than 4 prenatal care visits before childbirth, etc.)

2. Review of Existing Literature on this Dataset

Dawson et al. [3] use the Adding It Up dataset to investigate how to integrate the essential services defined in the report into health coverage in the 11 Pacific nations. They find that a large investment is needed as capacity building is essential for data collection to plan sexual health services effectively. Routine, systematic, standardized data collection on contraception, abortion, and reproductive coercion should be implemented and reported annually. This data can be used to enhance services up to international standards and assess SRH approaches' success.

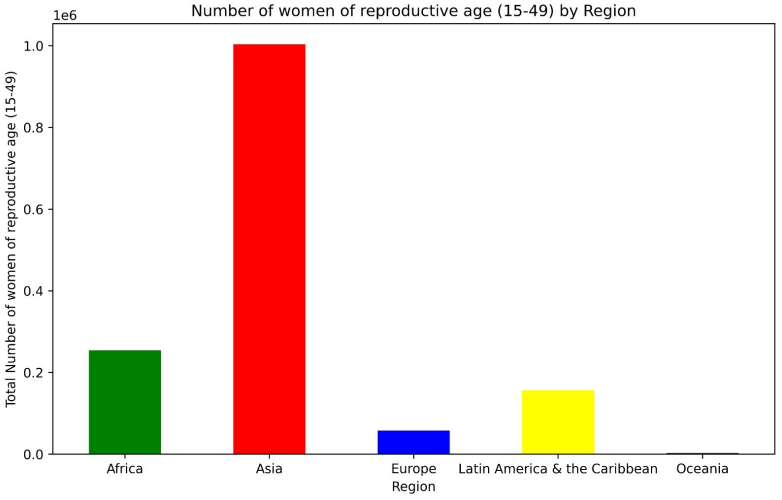
Sathar et al. [4] sought to identify levels of family planning financing in Pakistan and assess whether current funding is sufficient to meet national and international goals. They estimated family planning costs nationally using the Adding-It-Up dataset and methodology and by assessing expenditure trends between 2017 and 2021. They concluded that family planning services in Pakistan cost $81 million, with a $93 million gap to provide services to all women with unmet contraception needs. Despite an upward trend in government expenditure, it was slow and varied across regions.

3. Data Exploration

After examining the data, we decided to build models that would allow us to predict the region to which countries belong based on values for a given set of variables. For the purposes of this project, it is important to keep in mind the difference in size between regions in terms of women of reproductive age (15-49). We will also transform all the data to relative terms/per capita to standardize it.

Variable	Definition
Country	Location of the abortion data entries by countries
Region	Location of the abortion data entries by region
Anc_4plus	Percent of women 15-49 with four or more ANC visits
Upreg	Unintended pregnancies, total number
Upreg_no_cp	Number of unintended pregnancies at no contraceptive care
Upreg_abortion_no_cp	Number of abortions at no contraceptive care
Upreg_std	Upreg no_cp divided by Upreg
Upreg abortion_std	Upreg abortion_no_cp divided by Upreg
Curr_costs_cp_percap	Total contraceptive care costs per capita
Curr_costs_prnc_percap	Total pregnancy-related and newborn care costs per capita
Curr_abortion_pac_costs_percap	Total abortion-related care costs per capita
Rate_matdeath	Total maternal deaths per 100,000 live births
Sti_nocarc	Number of women with one of the 4 curable STIs needing but not receiving care
Inneed_married	Number wanting to avoid pregnancy, all married women
Inneed_formerlymar	Number wanting to avoid pregnancy, formerly married women
Inneed_nevermar	Number wanting to avoid pregnancy, never married women
Inneed	The addition of Inneed_married, Inneed_formerlymar, Inneed_nevermar divided by wra
Wra	Number of women of reproductive age, 15-49

Chosen Variables and Definitions



Women by Continent (in Bln)

We chose 18 variables out of more than 300 in the dataset based on data availability and quality, relevance to the goal of our models, and potential issues like significant collinearity. We also use some of these variables to normalize other values, like "wra" and "upreg," and we even created some variables out of other ones, like Inneed.

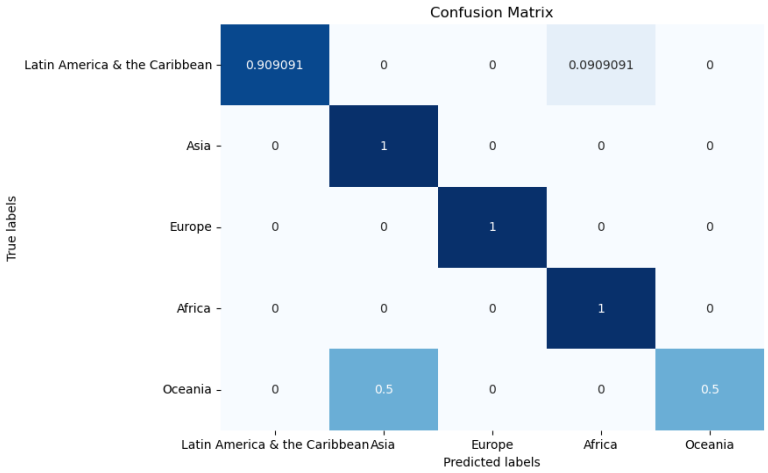
4. Logistic Regression

The first model we chose was a non-binary logistic regression. We tested other modes like the Lasso and Ridge classifiers, but the logistic regression outperformed all of them as our classification algorithm. This regression aims to model the probability that a given observation, think in terms of the abortion characteristics of a country, belongs to a particular category, in this case, our five regions, based on one or more explanatory variables. Thus, we are developing and training a model that will tell us the probabilities, given certain abortion data, of a country belonging to either Africa, Asia, Europe, LATAM, or Oceania. The latest results of our multinomial logistic regressions were promising. We obtained an accuracy of 91%, and with the exception of LATAM, our precision for each region was greater than 85%.

Classification Report

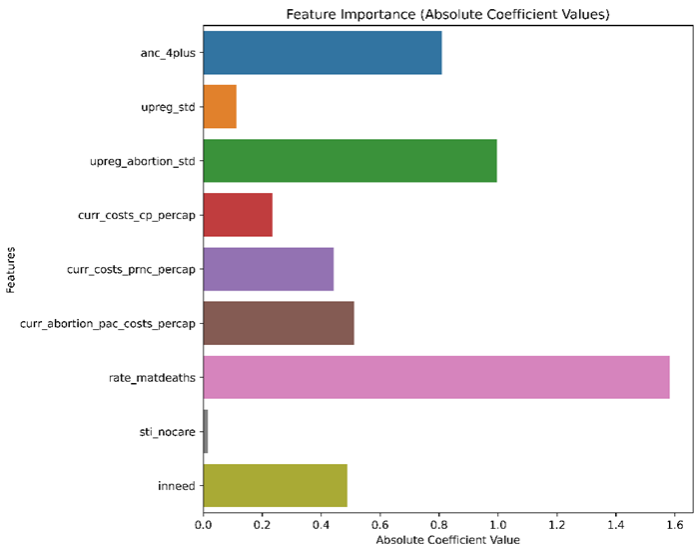
Region	Precision	Recall
Africa	1.0	0.9090909
Asia	0.8	1.00
Europe	1.0	1.00
Latin America	0.5	1.00
Oceania	1.0	0.50
	Accuracy:	0.91

Classification Report



Confusion Matrix

When examining the most significant coefficients (keep in mind all the data is standardized), we obtained that Rate_matdeath, Upreg_abortion_std, and Anc_4plus. Although Rate_matdeath having the largest coefficient might be surprising at first, it makes logical sense given that it is a strong differentiator between regions, and it would be surprising to see, for example, a European country having a Rate_matdeath extremely different from another European nation but not so surprising if comparing it to another country in Latam.



Absolute coefficient value in multinomial logistic regression

It might be argued, perhaps, that it is the most characteristic feature of these regions among the chosen variables. Lastly, given that we are using a test set and random states, readers could be concerned about changes in accuracy as the random state changes. We shared the same concern, and that is why we decided to create a loop to obtain the average accuracy for our model after changing the random state several times. We obtained an overall accuracy of 91% and an average accuracy of 86% for the five models and a standard deviation of just 9%. These are strong results when considering we are working primarily with limited data and regarding human behavior.

5. Neural Network

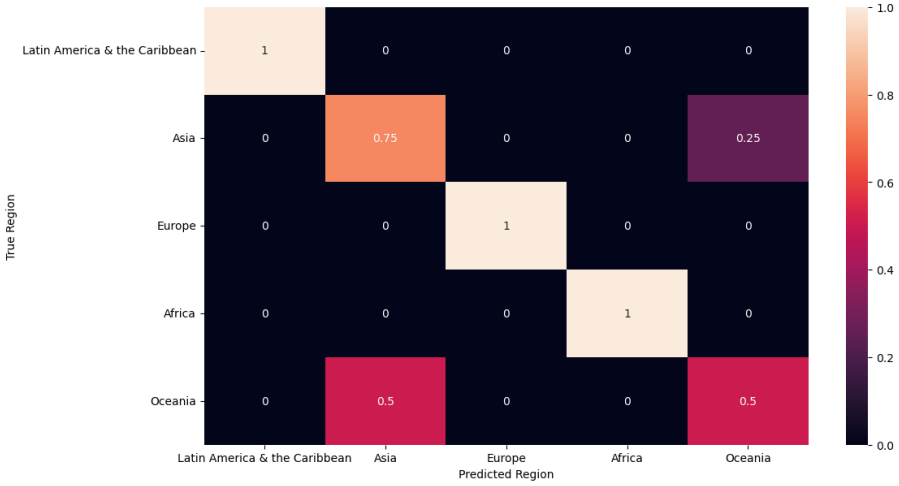
This model uses a linear neural network to predict a country's region based on statistics relating to women and sexual/reproductive healthcare. All 14 originally selected features were standardized and inputted to the network. No additional feature selection was done. When designing the model, we believed the data was most suited for a simple linear network rather than a convolutional network because of the number of features we have (14) and the number of region categories (5). Additionally, since we were working with a set of features rather than image data, a linear Neural Network was a better choice than a CNN. Some of the hyper parameters used are shown in the table below.

The model consists of one hidden layer and uses an Exponential Linear Unit (ELU) activation function. After experimenting with different hyper parameters and network structures, including batch size, learning rate, and different activation functions, we found this model to perform the best. After splitting the training and testing data with an 80:20 ratio, training the network, and validating using the test data, we find a 90.9% accuracy. The confusion matrix describing the predicted labels is shown below. As we see through the confusion matrix, the model has the most difficulty predicting countries in Oceania. We foresaw this difficulty due to the small number of countries in this region. We can also see through the confusion matrix that Asian countries are also most commonly confused with Oceania. Some of these limitations we believe to be primarily due to the scarcity of complete data in Oceania.

Hyperparameter	Value
Input dimension (in dim)	14
Output dimension (out dim)	5
Hidden dimension	64
Second hidden dimension (hidden dim2)	32
Number of epochs (n epochs)	40
Batch size	8
Learning rate	1e-3

Classification Report

Region	Precision	Recall
Africa	1.00	1.00
Asia	0.75	0.75
Europe	1.00	1.00
Latin America	1.00	1.00
Oceania	0.50	0.50
	Accuracy:	0.877



Confusion Matrix

The network was run multiple times, varying splits of training and testing data. The network was run 20 times, and the average accuracy over these runs is 87.7%. The standard deviation of these accuracies is 6.2%. This signifies that the percent accuracy is variable based on the training and test split, however the standard deviation is relatively agreeable within the social sciences, meaning there is not a detrimental variation within the accuracies.

Conclusions. We can observe that both the neural network and the logistic regression perform satisfactorily well in predicting the continent of a country based on a set of characteristics relating to its sexual/reproductive health. While this does not clearly point us to the solution to health disparity, it does suggest that there are regional patterns in women's healthcare. This tells us that solutions to disparities might not be found in standardized approaches for all countries, but that regional considerations (e.g. societal attitudes and customs, environmental conditions, religious beliefs, etc.) have to be given some weight when working towards reducing healthcare inequality. The fact that Asia is the continent for which both models are weaker predictors is consistent with this view, since Asia is simultaneously the largest and most diverse continent, including with respect to national income and healthcare. This could also explain the fact that the most important coefficients in the logistical model are maternal deaths, number of prenatal visits, and standardized number of abortions with no contraceptive care (all potential proxies for a country's healthcare access and quality). The similarity in overall performance between the two models suggest that one is not necessarily suited more properly to the data and with fine tuning, both logistic regression and neural networks can act as successful classifiers for this dataset. We can also say that perhaps another model or method we did not implement has the possibility to increase the accuracy of predictions above 91%.

References

[1] E. SUlly et al., (2019). Adding It Up: Investing in Sexual and Reproductive Health 2019 - Methodology Report Supplementary Materials. https://osf.io/m85s9/. Accessed 4 March 2024.
[2] Guttmacher Institute, (2019). Adding It Up: Investing in Sexual and Reproductive Health. https://www.guttmacher.org/adding-it-up. Accessed 4 March 2024.
[3] Dawson, A., Ekeroma, A., Wilson, D. et al. How do Pacific Island countries add up on contraception, abortion and reproductive coercion? Guidance from the Guttmacher report on investing in sexual and reproductive health. https://doi.org/10.1186/s12978-021-01122-x
[4] Zeba Sathar, Susheela Singh, Sabahat Hussain, Maqsood Sadiq, (2023) Financing gaps for Pakistan's contraceptive prevalence goals: Analyses using the Guttmacher adding-it-up methodology https://www.sciencedirect.com/science/article/pii/S0010782422004334

Notes on Individual Contributions

A rough approximation of how we distributed the work amongst each other follows:

- Gabriel did the initial data cleaning, wrote up the exploratory data analysis, made 4 visualizations, and coded, analyzed, and wrote up the logistic regression section.
- Giselle coded, analyzed, and wrote up the neural network section, created its visualization, and wrote the abstract.
- Federico coded the LaTeX file and transcribed everyone's parts in it, wrote the dataset explanation, lit review, and the conclusion of the models.

Everyone looked at the overall report and edited it as needed at various stages.