# AI CoLab Internship Report

**Yara Yaghi**

**Computational Data Science Sophomore**

**George Mason University**

**June 2024 - August 2024**

# Table of Contents

# 1. Introduction

During my internship at MedStar Health's AI Colab, I had the unique opportunity to explore the intersection of healthcare and artificial intelligence. This experience was invaluable in providing hands-on exposure to cutting-edge AI applications in the healthcare sector. I was involved in projects that required the integration of large-scale medical data with advanced AI techniques, highlighting the importance of data-driven approaches in modern healthcare. The AI Colab served as a dynamic environment where innovation met clinical relevance, allowing me to contribute to meaningful projects aimed at improving patient outcomes. This internship was pivotal not only in advancing my technical skills but also in deepening my understanding of how AI can be harnessed to address some of the most pressing challenges in healthcare today.

# 2. Project Details

## 2.1 SLE Data and Project

### 2.1.1 Objective

The primary objective of this project was to investigate the relationship between the SLE Index and various clinical and laboratory markers of disease activity in patients with Systemic Lupus Erythematosus.

### 2.1.2 Tasks

The tasks involved data cleaning, Exploratory Data Analysis (EDA), correlation analysis, longitudinal analysis, predictive modeling, and summarizing findings with clear visualizations.

### 2.1.3 Skills Used

The skills used included data analysis, programming in Python (pandas, matplotlib, seaborn, scikit-learn), data visualization, project management, and interdisciplinary collaboration.

## 2.2 Social determinants of health: predicting maternal risk using ML

#### 2.2.1 Objective

The primary objective of this project was to leverage machine learning (ML) algorithms to predict maternal risk, focusing on complications and mortality, by analyzing the impact of social determinants of health. This project aimed to create a synthetic population that would enable accurate risk prediction while preserving patient privacy and ensuring that the findings could be generalized across diverse sociodemographic groups.

#### 2.2.2 Tasks

The tasks undertaken in this project were about my understanding of maternal health and the role of machine learning in the field of medicine. To achieve this, I conducted extensive research using library databases, academic journals as well as other scholarly materials. Therefore, I studied literature on maternal mortality rates, healthcare disparities and ML applied to predict health outcomes. Thereafter, I collected and combined all such information resulting into a complete essay that highlighted the present problems in maternity care, how ML can tackle them and the ethicality associated with employing AI in medical care.

### 2.2.3 Skills Used

The primary skills I applied in this project included research and literature review, which involved critically analyzing scholarly articles and extracting relevant information. I also honed my writing skills by structuring the information into an essay format, ensuring clarity and coherence. Additionally, I developed a deeper understanding of how machine learning is applied in healthcare, particularly in addressing disparities in maternal health. This project enhanced my ability to synthesize complex information from various sources and communicate it effectively, which is essential for both academic and professional settings.

## 2.3 Determining an Association between Family History of Cardiometabolic Disease and the Incidence of Preeclampsia in Pregnant Women of Different Racial and Age Groups using Artificial Intelligence and Machine Learning

#### 2.3.1 Objective

The objective of this project was to determine whether there is an association between family

history of cardiometabolic diseases (such as hypertension, diabetes, and cardiovascular diseases) and the incidence of preeclampsia among pregnant women across different racial and age groups. The project aimed to utilize artificial intelligence (AI) and machine learning (ML) to identify at-risk populations and enhance predictive analytics in maternal health.

#### 2.2.2 Tasks

The tasks for this project involved conducting extensive research to understand the relationship between cardiometabolic disease history and preeclampsia. This included a literature review to explore the existing evidence on family history as a risk factor for preeclampsia, as well as investigating racial and age disparities in maternal health outcomes. Additionally, the project required the preparation and analysis of a synthetic dataset, applying machine learning models such as logistic regression to assess the association between the predictor variables (e.g., family history, race, age) and the incidence of preeclampsia. The final task involved synthesizing the findings and preparing them for presentation, including a detailed analysis of how different racial and age groups are affected by these health conditions.

#### 2.2.3 Skills Used

The project required the application of various skills including research and literature review, data analysis, and machine learning. I enhanced my understanding of maternal health and the impact of cardiometabolic diseases by critically analyzing scholarly articles. I developed technical skills in data preparation, including dataset merging and the application of binary features, as well as model training and evaluation using Python. Additionally, I applied logistic regression techniques to analyze the data and extract meaningful conclusions. This project also improved my ability to work with a multidisciplinary team and communicate complex findings effectively, both in written and presentation formats.

# 3. Learning Experience

## 3.1 New Skills Learned

In the AI CoLab internship, I was exposed to a wide array of skills within data science and artificial intelligence. Developing further complex statistical methods like Logistic Regression and Linear Regression, working with machine learning algorithms in Random

Forest, SVM, and GBM, some data preprocessing techniques used include encoding categorical variables and handling missing data. I used pandas for data manipulation and processing, scikit-learn for the implementation of the model, and matplotlib and seaborn in case some visualization would be required. I learned class balancing and hyperparameter tuning for better results of the model. I got strong experience in handling and analyzing large datasets and interpreting the results from the model's results based on which meaningful conclusions can be drawn.

I further gained many academic and research skills, especially in data science and health. I have learnt how to write a research article w.r.t. structure, methodology and clear discussion of the results. I also developed the ability for searching, critical analysis of scholarly articles, and building on prior knowledge that helps base our research on a credible academic foundation. From all these experiences, a sharp mind research skills learn what to do with them to navigate the academic publishing process in order to bring meaningful projects lying at the junction of AI and healthcare into limelight.

Apart from the technical and research skills, I had a great experience in collaborative teamwork and project management. I did a number of interdisciplinary projects with students, data analysts, and health professionals where I learnt to communicate technical concepts to non-technical stakeholders in an effective manner. This improved my organizational skills and time management skill since we were dealing with a number of deadlines. Such collaboration has helped me enhance my technical abilities and the capacity for working in a productive and innovative team environment.

## 3.2 Challenges Faced

During the internship, there were so many challenges that basically shaped my growth. One of them was improvement in terms of accuracy in the logistic regression model. This was quite challenging because it required a lot of patience while trying out several techniques for feature engineering, hyperparameter tuning, and class balancing. With every try, more knowledge was gained, as well as complexities with regard to refining predictive models. The scheduling of mentors and group members was another challenge due to differences in availability and responsibility. In doing so, effective communication and flexibility were quintessential to be

made the most out of collaborative effort. Besides that, I had to face some of the coding problems that I hadn't faced during my studies. That required quick learning of techniques and debugging issues that I was not familiar with. Overcoming these challenges has strengthened my problem-solving skills, enhanced my technical knowledge, and made me realize that adaptability is a crucial attribute in any professional workplace.

# 4. Mentorship

## 4.1 Dr. Mindi Messmer, Dr. Omar Aljawfi , Dr. Charles Amankwa

## 4.2 Support Received

Throughout my internship, I was fortunate to receive invaluable support from Dr. Mindi Messemer, Dr. Omar Aljawfi, and Dr. Charles. Dr. Mindi, my mentor, played a crucial role in guiding me and my group through the research process. She helped us identify a suitable project, pointed out the limitations in our study, and consistently provided direction on our next steps. Dr. Omar, the AI CoLab coordinator, was instrumental in our learning experience. He reviewed our findings, highlighted errors, and provided us with new learning materials. His guidance was essential in setting up our project, and he even supported the creation of a future poster presentation. Dr. Charles, our Data Analyst, offered expert advice on handling our data and refining our research questions. His insights were critical in selecting the most appropriate models for our project. The combined support from these mentors greatly enhanced the quality of our work and my overall learning experience.

# 5. Final Deliverables

## 5.1 Project Summary

Our project focused on determining the association between family history of cardiometabolic disease and the incidence of preeclampsia among pregnant women across different racial and age groups using artificial intelligence and machine learning. The logistic regression model developed for this analysis achieved an accuracy of 71%, showing reasonable discriminatory power and good calibration, especially at higher probabilities. The key findings revealed that African American, Asian, and 'Other' racial groups had a 4-5% higher risk of preeclampsia

compared to White patients. Age trends indicated that younger women (18-24) faced the highest preeclampsia risk (~59%), which decreased with age, with the lowest rate (~55.5%) observed in the 35-44 age group. Additionally, our analysis highlighted that routine obstetric care dominated across races, but higher cesarean delivery rates among White women may explain their increased preeclampsia risk with age. These results emphasize the critical need for targeted monitoring and intervention strategies considering both racial and age-related factors.

### 5.2 Presentations

The final presentation was on the relationship between family history of cardio-metabolic disease and risk of preeclampsia in pregnant women of various racial and age groups using artificial intelligence and machine learning. We went further to give the methodology we had used in doing this project, and that included data preparation, model training, and finally, the evaluation. We shared key findings, such as the accuracy of our logistic regression model at 71%, racial disparities indicating higher risk for preeclampsia with African American, Asian, and 'Other' racial groups, and that younger women aged 18-24 had the highest risk for preeclampsia, which decreased with age. These findings were illustrated with visualizations through decile analysis and cross-tab rates between race and age groups. We concluded with the implications of these results and some further research, such as investigating the causes of the disparities found or the generalization of this study to a larger and more representative sample.

## 6. Reflection

### 6.1 Personal Growth

The internship at AI CoLab has been transformative both personally and professionally. It provided me with hands-on experience in combining medical research with AI, which significantly broadened my understanding of data science applications in healthcare. Collaborating with health professionals and fellow data scientists enhanced my teamwork and communication skills, crucial for interdisciplinary projects.

This internship has also played a crucial role in enhancing my coding skills by exposing me to

real-world healthcare datasets and challenging analytical tasks. This experience improved my proficiency in data manipulation, statistical analysis, and implementing machine learning models, all essential in data science. In terms of research skills, I gained practical experience in conducting through literature reviews, which involved synthesizing existing research findings and identifying gaps for future investigation. Furthermore, I developed strong presentation skills through regular updates and meetings with colleagues and stakeholders. Communicating complex technical concepts effectively become second nature, enabling me to articulate findings and insights from data analyses with clarity and confidence.

This experience sharpened my research abilities and expanded my view on how data science can address critical global health challenges. Overall, the internship not only solidified my career aspirations in healthcare data science but also instilled a deeper sense of purpose in using technology for social impact.

## 6.2 Future Plans

As a result of this experience, I have made a firm commitment to pursue a career in healthcare data science in the future. I became aware of the immense potential of data driven solutions for improving healthcare outcomes through my work at the intersection of medical research and artificial intelligence. In addition to increasing my technical skills, I gained a deeper understanding of the role data science plays in tackling global health challenges through hands-on experience with healthcare datasets, coding challenges, and machine learning applications.

I was also able to expand my perspective on interdisciplinary teamwork as well as the importance of transforming technical insights into actionable strategies when I collaborated with health professionals and researchers. I grew passionate about using data science to help underserved regions and conflict zones as a result of this experience.

In the end, I have been motivated by this internship to continue developing by kills and expertise in healthcare data science, with the goal of contributing meaningfully to advancements in medical research, healthcare delivery, and health equity globally.

# 7. Graphs and Figures

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.47 | 0.07 | 0.12 | 1769 |
| 1 | 0.72 | 0.97 | 0.82 | 4323 |
| accuracy | | | 0.71 | 6092 |
| macro avg | 0.60 | 0.52 | 0.47 | 6092 |
| weighted avg | 0.65 | 0.71 | 0.62 | 6092 |

Figure 1: Learning to build a Logistic Regression Model on python

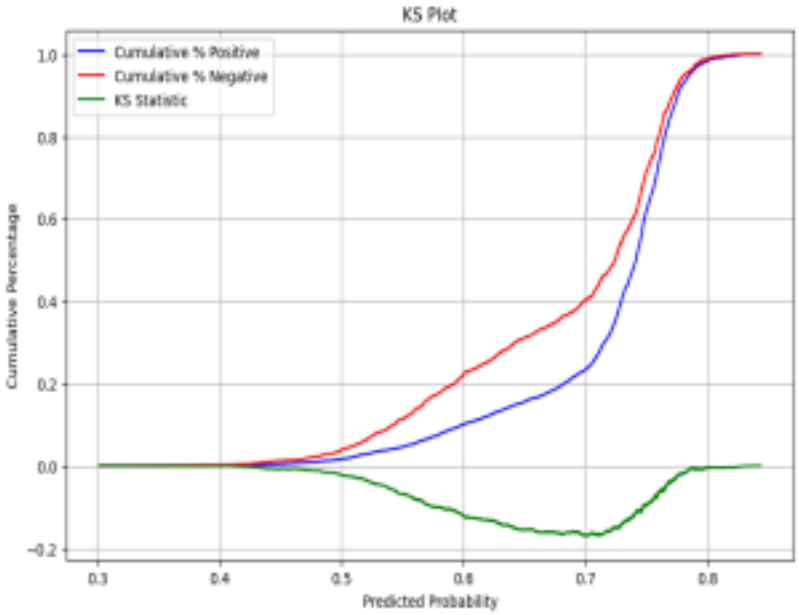


Figure 2: Learning to build a KS Plot on Python

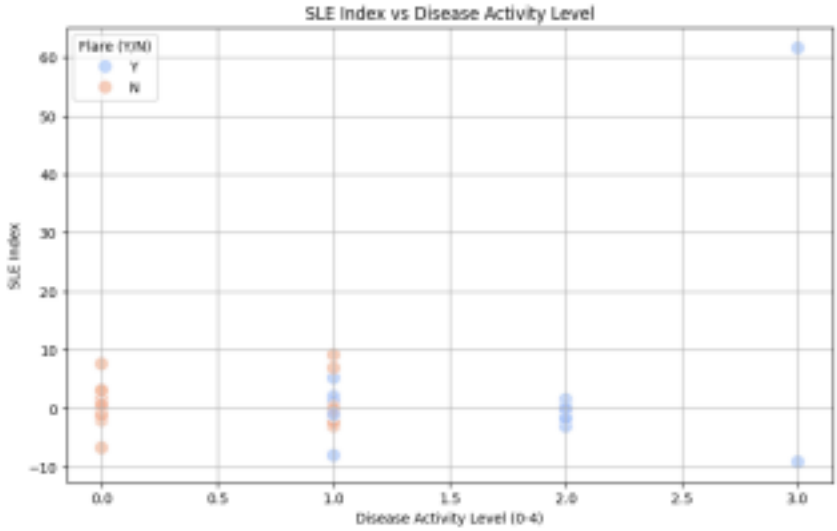


Figure 3: Learning how to build and analyze complicated graphs