AI CoLab Internship Report

Yara Yaghi

BS Computational & Data Science 2025

George Mason University

May 2025 - August 2025

Table of Contents

1. Introduction

The AI CoLab Internship Program is an interdisciplinary research program that offers students an opportunity to use machine learning and artificial intelligence to tackle tough biomedical science and healthcare problems. The program is designed as an inclusive community that aligns interns with subject-matter experts and senior mentors, where technical skill-building and subject-matter knowledge are combined to develop sound, data-driven solutions.

I was an intern in data science research at AI CoLab on a research project that evaluated the relationship between kidney function, β-cell failure, and type 2 diabetes risk. My group and I used NHANES data to investigate the relationship between physiological markers such as the estimated glomerular filtration rate (eGFR) and blood urea nitrogen (BUN) and lifestyle markers such as smoking and alcohol consumption. Our responsibilities were data cleaning and preprocessing, feature engineering, model estimation and selection, and the interpretation of statistical results in a biomedical setting.

Program design offered a balance of liberty and periodic supervision, creating a reproducibility-friendly culture, interpretability, and ethical responsibility—values highly relevant to healthcare solutions. Use of real-world biomedical data sets also encompassed tackling real-world issues like missing data, measurement noise, and class imbalance that needed to be addressed judiciously so that the validity of findings is not affected.

This project enhanced my technical skill, increased my exposure to the analysis of clinical data, and reinforced the importance of working in teams to provide quality research. Most importantly, however, it illustrated the ability of AI, if utilized properly, to inform patient care choices, determine target interventions for screening, and create public health interventions.

2. Project Details

2.1 Project Title

Predicting β-cell Dysfunction from Kidney Function and Lifestyle Factors in Type 2 Diabetes Using Machine Learning

This research study explored the relationship among kidney function, lifestyle variables, and β-cell dysfunction in adult patients with Type 2 Diabetes (T2D). β-cell dysfunction, a prime cause of impaired insulin secretion, is a significant factor in diabetes progression and management. Using a nationally representative large dataset, we developed and validated machine learning models to predict β-cell function with the aim of identifying such key predictors that can be utilised in early intervention and clinical decision-making. The study also used state-of-the-art data preprocessing techniques so that we can supply a stable, high-quality input to the modeling procedure.

2.2 Objective

The principal objective of this project was to apply machine learning techniques to examine how markers of kidney function, i.e., eGFR and BUN, interacting with lifestyle factors such as smoking and alcohol consumption impact β-cell function quantified in terms of HOMA-B. We aimed to develop predictive models to identify individuals at high risk of β-cell failure within the T2D population and contribute to earlier diagnosis and therapy. The project also hoped to assess the value of clinical and behavioral parameters as predictors, compare various algorithmic approaches, and report results in a clinically applicable and translatable manner to real-world healthcare environments.

2.3 Tasks

To achieve these objectives, the project embraced a systematic, multi-stage workflow:

- Data Acquisition and Integration: We integrated a number of years of NHANES data sets (1999–2020) in order to produce a unified sample of U.S. adults with documented Type 2 Diabetes.
- Inclusion/Exclusion Criteria: We applied filters to retain only likely T2D cases, removing Type 1 Diabetes and persons with end-stage kidney disease (eGFR < 30 mL/min/1.73m²).

- Data Cleaning: This included removal of duplicates, missing value treatment using median/mode imputation and reasoning by logic, and capping of outliers wherever necessary.
- Feature Engineering: We calculated derived variables such as HOMA-B (β-cell function) and did transformation of skewed variables (e.g., log transformation of ACR). We engineered categorical and binary features as well, e.g., "Smoker" vs. "Non-Smoker."
- Model Development: We compared and tested different machine learning models such as Random Forest, XGBoost, Bagging, LightGBM, etc., employing R^2, MSE, and feature importance to evaluate their performance.
- Model Interpretation and Visualization: Feature importance plots and model performance plots were generated to identify the most influential predictors and assess model accuracy.

2.4 Skills Used

The work required a multi-disciplinary skill set, i.e., domain knowledge in health research along with excellent technical skills in data science.

- Programming & Tools: Python and R, using libraries such as pandas and NumPy for data processing, scikit-learn for machine learning workflows, XGBoost and LightGBM for gradient boosting models, and matplotlib/seaborn for plots.
- Machine Learning Techniques: Utilized ensemble learning techniques (Random Forest, Bagging, Gradient Boosting) and hyperparameter search via GridSearchCV to improve model performance.
- Statistical Analysis: Used performance measures based on regression metrics (R^2, MSE) to numerically quantify model performance so it could be interpreted clinically.
- Data Preprocessing: Utilized scaling, encoding, feature skew transformation, and logical imputation of missing values.
- Domain-specific expertise: Use of diabetes pathophysiology, measurement of kidney function, and β-cell biology in guiding feature selection, model interpretation, and exclusion criterion selection.

3. Learning Experience

3.1 New Skills Learned

We learned through this project that feature selection, scaling, and transformation are all vital to model interpretation and performance. Data preprocessing was key to our project since we had to lean heavily on it to influence our models' accuracy and validity directly. Exploratory data analysis (EDA) played a critical role in discovering meaningful patterns, informing our feature engineering, and making our models clinically relevant. We also understood the need to evaluate models based on more than a single criterion—like AUC, Mean Squared Error (MSE), and feature importance scores—so we could get a highly balanced sense of their virtues and vices. Aside from technology, we understood the need to leverage each member's strength. Through delegation of work according to our areas of specialization, we worked well, ensured high standards, and completed the project within time..

3.2 Challenges Faced

One of the biggest challenges we faced was that the NHANES dataset did not contain specific clinical biomarkers—such as HbA1c, C-peptide, and specific medication details—that limited the clinical validity and richness of our models. Information merging within multiple NHANES cycles took considerable efforts in merging variables, cleaning, standardizing, and scaling as well as survey year consistency maintenance. Missing data was especially difficult to deal with, and wherever possible logical imputation methods had to be employed to maintain clinical relevance in variables. Cross-sectional design of the dataset also restricted us from assessing longitudinal trends in β-cell function or viewing temporal changes. Despite these limitations, we adapted our approach to maximize the value of current data, finding a compromise between statistical precision and clinical relevance.

4. Mentorship

4.1 Mentors’ Names

- Dr. Nawar Shara, PhD – Founding Co-Director, AI CoLab, MedStar Health Research Institute
- Omar Aljawfi – Informatics Analyst & Data Scientist, MedStar Health Research Institute
- Raed Darwish – Mentor, AI CoLab
- Diyang Lyu – Mentor, AI CoLab
- Zannatul Ferdous – Mentor, AI CoLab
- William Mea – Mentor, AI CoLab
- Maryam Solimany – Program Coordinator, AI CoLab, MedStar Health Research Institute

4.2 Support Received

We are deeply grateful for the valuable feedback and encouragement provided by our mentors and program coordinators throughout the course of this project. Dr. Nawar Shara and Omar Aljawfi offered insightful feedback on data science techniques, project design, and clinical relevance of our project. Our technical mentors Raed Darwish, Diyang Lyu, Zannatul Ferdous, and William Mea provided hands-on technical guidance, assisting us in iterating over our preprocessing procedures, optimizing our machine learning algorithms, and interpreting our findings with both statistical soundness and clinical insight. Maryam Solimany ensured we had the organizational and logistical support we needed to proceed without a hitch, from scheduling mentorship meetings to ensuring we had access to required resources. Their combined expertise, feedback, and encouragement greatly contributed to the quality, depth, and relevance of our project.

5. Final Deliverables

5.1 Project Summary

We found that Random Forest performed best with the highest rate of prediction ($R^2 = 0.8531$) and least error (MSE = 0.1469). XGBoost was the next to be classified, and Bagging and LightGBM performed less well with increasing values of both error and MSE. Renal function (eGFR, BUN), albumin-to-creatinine ratio (ACR), age, and socioeconomic status such as income-to-poverty ratio had the greatest impact on predicting β-cell dysfunction according to

feature importance and SHAP values. These findings emphasize the key relevance of clinical biomarkers and demographic factors to forecast diabetes outcomes and demonstrate the strength of ensemble methods for finding intricate interactions.

5.2 Presentations

For the AI CoLab project, we demonstrated our project in the final presentation to faculty, peers, and mentors. We discussed in the presentation our research issue, objectives, data preprocessing work, feature engineering methods, and the use of machine learning models. We also provided visualizations such as model flow charts, feature importance charts, and performance comparison plots in order to make our findings relatable to both technical and non-technical audiences. The presentation finished by summarizing the clinical significance of our results, the difficulties encountered with the NHANES dataset, and suggestions for future directions, including longitudinal datasets and other biomarkers.

Separately, I completed an individual project with my mentor, Diyang Lyu, to help create the materials page of the interactive workshop "2025 GHUCCTS-AI CoLab Summer Intensive in AI, Biostatistics, & Data Science." I built an interactive Jupyter Notebook to help the participants learn and practice data exploration techniques. The notebook guided users through the entire workflow of working with a real-world dataset, starting from data cleaning, missing value imputation, exploratory data analysis (EDA), and basic feature engineering. Although I could not participate in the workshop itself, we wanted to make it more hands-on by including practice questions and exercises with solutions for students to attempt step-by-step and observe how preprocessing choices affect model performance. This gave me the opportunity to give back not just as a learner but also as a creator of learning materials for future students.

6. Reflection

6.1 Personal Growth

This internship enhanced my professional and technical growth enormously. I enhanced my capabilities in machine learning, data analysis, and model explainability and enhanced my communication skills by sharing results to different audiences. These tasks of cleaning and merging

enormous datasets assisted me in learning persistence, accuracy, and the importance of teamwork. In addition to technical capability, I have developed leadership and teamwork capabilities by actively engaging in collaborative work and problem-solving and this has enabled me with the capabilities to excel in future professional and research settings.

6.2 Future Plans

This experience has only served to strengthen my resolve to pursue careers in data science for biomedical research and health care. Having witnessed the potential of AI and machine learning to generate actionable intelligence in the clinical environment has stimulated me to learn even more about predictive modeling for diseases such as diabetes and cardiovascular disease. In the future, I want to continue specializing in advanced modeling techniques and develop projects that integrate technical innovation with real-world healthcare needs. Lastly, this internship experience again reinforced my long-term goal of applying data-driven methods to improve patient outcomes and evidence-based medical decision-making.

7. Graphs and Figures

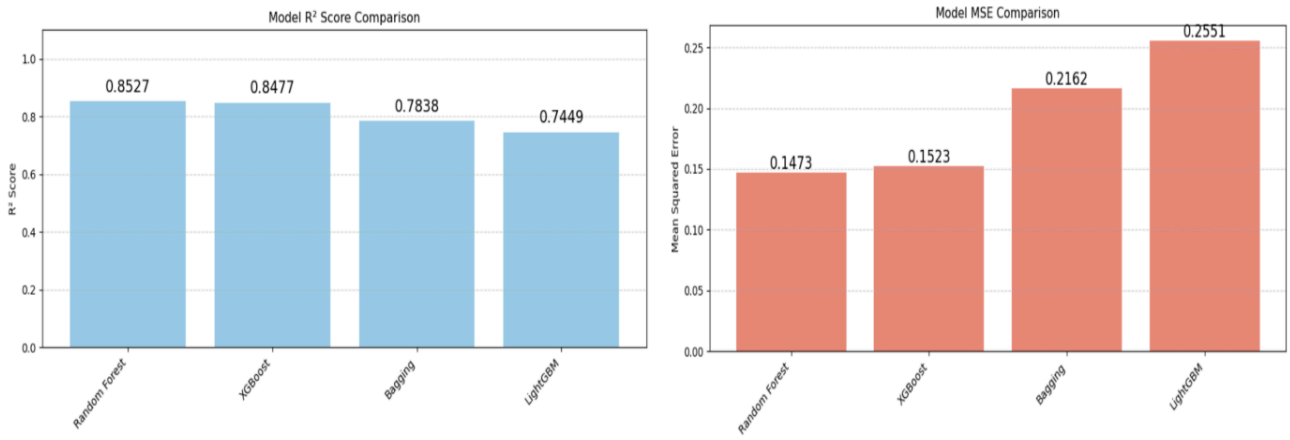

Figure 1: Comparison of Model Performance Using R² Score and Mean Squared Error (MSE)

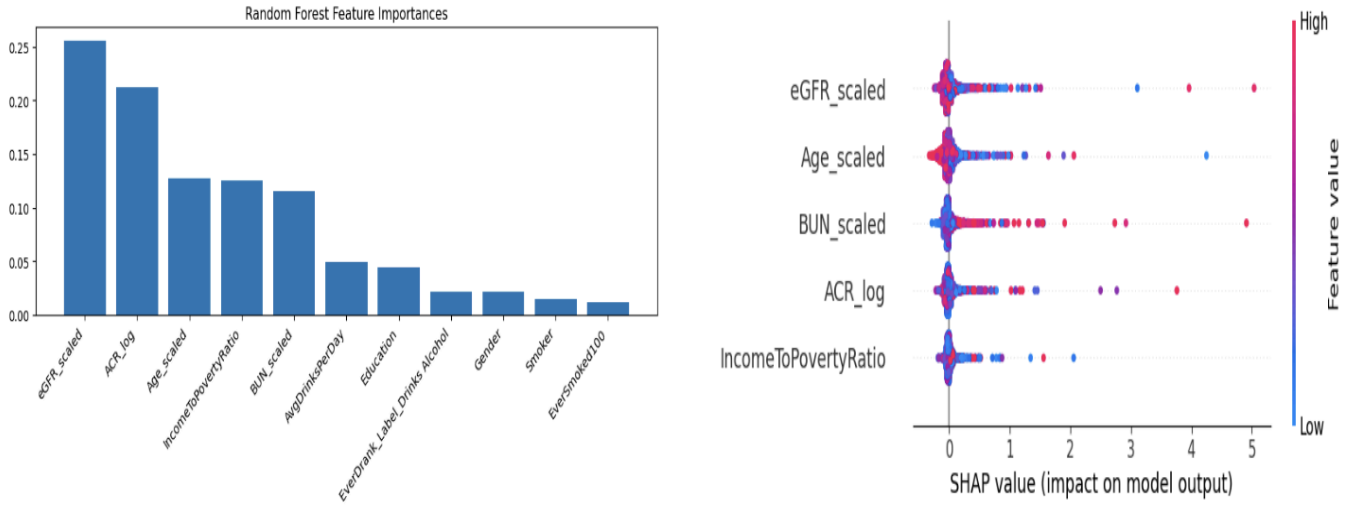

Figure 2: Random Forest Feature Importance Rankings and SHAP Value Analysis Showing Feature Impact on Model Predictions

References

Akhuemonkhan, E. & Lazo, M. (2017). Association between family history of diabetes and cardiovascular disease and lifestyle risk factors in the U.S. population (NHANES 2009–2012). *Preventive Medicine, 96*, 129–134. https://doi.org/10.1016/j.ypmed.2016.12.015

CDC. (2024, May 15). About Type 2 Diabetes. Diabetes. https://www.cdc.gov/diabetes/about/about-type-2-diabetes.html

Dludla, P. V., Mabhida, S. E., Ziqubu, K., Nkambule, B. B., Mazibuko-Mbeje, S. E., Hanser, S., Basson, A. K., Pheiffer, C., & Kengne, A. P. (2023). Pancreatic β-cell dysfunction in type 2 diabetes: Implications of inflammation and oxidative stress. *World Journal of Diabetes, 14*(3), 130–146. https://doi.org/10.4239/wjd.v14.i3.130

Formatting Instructions

- **Font:** Times New Roman, 12 pt
- **Line Spacing:** 1.5 lines for the main content
- **Page Margins:** 1 inch on all sides
- **Page Numbers:** Bottom right corner
- **Graphs and Figures:** Should be numbered and include captions (e.g., "Figure 1: Data Analysis Results")

Example Graph and Figure Captions

- **Figure 1:** Distribution of Task Completion Times
- **Figure 2:** Skill Proficiency Levels Before and After Internship
- **Figure 3:** Project Workflow Diagram