STATS 503: Project Report
Predicting Diabetes and Stroke Risk based on Lifestyle, Dietary and Demographic attributes using NHANES Data

Rishikesh Ksheersagar, Sudhanshu Agarwal, Sandeep Jala
{rishiksh, sudhagar, jsandeep}@umich.edu

April 21, 2024

1 Introduction

Public health is a dynamic and complex field that relies heavily on data-driven insights to tackle chronic diseases and promote healthier lifestyles across populations. Our project uses NHANES data to explore two key public health questions:

1. **Can lifestyle attributes such as diet, physical activity, smoking habits, and alcohol consumption be used to accurately predict the risk of developing diabetes?** This question is grounded in the increasing prevalence of diabetes worldwide, making it a major public health concern. We analyze NHANES data to develop a model that predicts diabetes risk based on diet, physical activity, and substance use. This model aims to aid healthcare providers and policymakers in preventive strategies.
2. **What are the key demographics, dietary habits, and biomarkers associated with stroke incidence?** This study aims to predict stroke incidence by analyzing key demographic factors, dietary habits, and biomarkers from NHANES data. Given stroke's significant impact on mortality and disability, understanding its predictors is vital for effective prevention and treatment. This research offers insights into stroke risk factors, potentially informing personalized interventions and enhancing prevention efforts and patient outcomes.

The NHANES datasets offer a valuable resource for addressing these questions, thanks to their comprehensive health, nutritional, and demographic data drawn from a diverse U.S. population. Our study not only aims to answer key health-related queries but also enhances NHANES's role as a powerful tool for impactful health research.

These questions transcend academic curiosity, aligning closely with real-world health challenges. Carefully crafted to be answerable with available data, their insights hold practical significance for healthcare professionals, researchers, and policymakers alike. Our research is poised to make a substantial contribution to the literature on lifestyle and health, offering evidence-based recommendations for fostering healthier communities globally.

2 Data Description

This section describes the datasets utilized in our analysis to answer the project's questions, Table 1 and 2 focus on the datasets used for both Diabetes Prediction and Stroke Risk Prediction respectively.

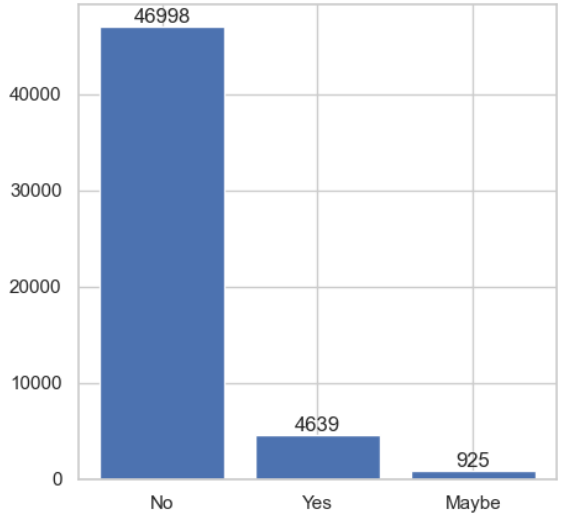
Each dataset has been carefully selected for its relevance to the key lifestyle attributes and health indicators pertinent to the study's objective. The utilization of these diverse datasets allows for a comprehensive analysis of factors influencing the risk of diabetes. For the first part where we predict if the patient might have diabetes or not, we use NHANES data from the years 2011 to 2020. For the second part of the project where we aim to find the features affecting the chances of stroke for predicting Stroke Risk, we use the NHANES data from years 2015 to 2020.

2.1 Question 1 - Diabetes Risk

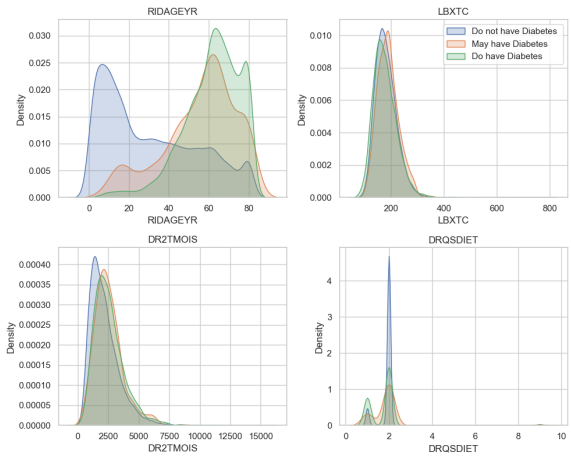
Dataset Name	Description
ALQ	Alcohol usage data
DEMO	Demographic information
MCQ	Medical conditions questionnaire
SMQRTU	Smoking - recent tobacco use
BMX	Body measures data
DIQ	Diabetes questionnaire data
OCQ	Occupational questionnaire
SMQSHS	Secondhand smoke data
BPQ	Blood pressure questionnaire
HSQ	Hearing status questionnaire
PAQ	Physical activity questionnaire
SMQ	Smoking status questionnaire
BPXO	Orthostatic blood pressure
INQ	Income questionnaire
SLQ	Sleep disorders data
TCHOL	Total cholesterol data
CDQ	Cardiovascular health data
KIQ	Kidney conditions - Urology data
SMQFAM	Family smoking data
WHQ	Weight history data

Table 1: Datasets used for analyzing diabetes risk based on lifestyle attributes.

Response Variable for Diabetes Risk Prediction The feature *DIQ010* extracted from DIQ dataset indicates if a doctor has told a patient that they have Diabetes. From this feature, we use {Class 1: Yes, Class 2: No, Class 3: Borderline / Moderate Risk} for further analysis.
We present a summary of all datasets collected to predict Diabetes Risk in Table 1. We also present some exploratory plots showcasing the data distributions in Figure 1.



(a) Number of patients with Diabetes



(b) Density Plots for some Key Features against Diabetes Risk

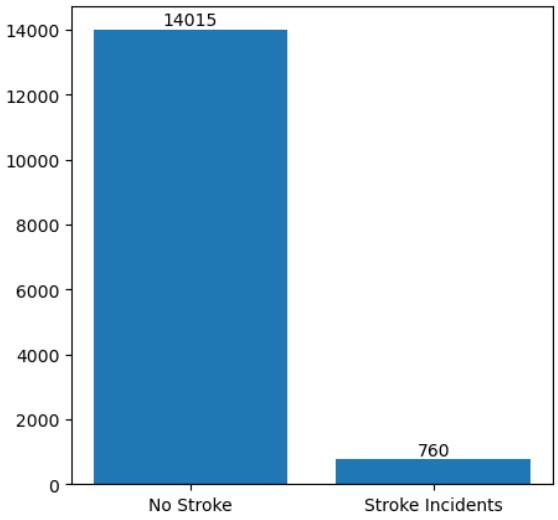
Figure 1: EDA on data gathered for diabetes Prediction

2.2 Question 2 - Stroke Risk

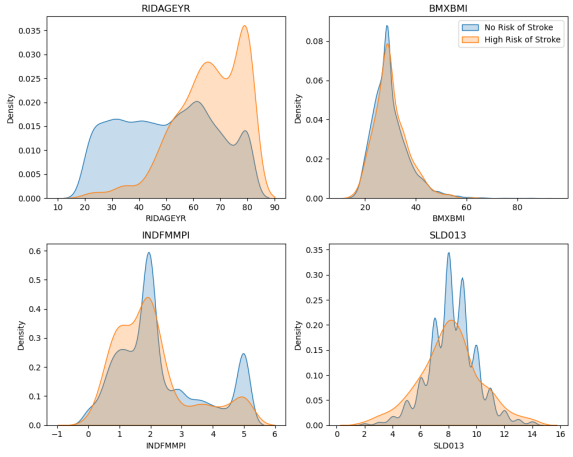
Dataset Name	Description
DEMO	Demographic information
HDL	High-density lipoprotein (HDL) cholesterol data
PAQ	Physical activity questionnaire
TCHOL	Total cholesterol data
ALQ	Alcohol usage data
DIQ	Diabetes questionnaire data
HSQ	Hearing status questionnaire
SLQ	Sleep disorders data
TRIGLY	Triglycerides level data
BPQ	Blood pressure questionnaire
DR1TOT	Dietary recall - day 1 total nutrient intakes
INS	Insulin level data
SMQRTU	Smoking - recent tobacco use
WHQ	Weight history data
CDQ	Cardiovascular health data
DR2TOT	Dietary recall - day 2 total nutrient intakes
MCQ	Medical conditions questionnaire
SMQ	Smoking status questionnaire

Table 2: Datasets used for analyzing stroke-associated factors.

Response Variable for Stroke Risk Prediction : The feature *MCQ160F* derived from MCQ dataset, indicates whether a doctor has informed the patient if he/she has had a stroke. We use the classes {1: Yes and 2: No} for further analysis.
Table 2 shows details of all NHANES datasets used for Stroke Risk Prediction. We also show some EDA plots highlighting Number of Stroke Incidents in the data and their correlation with some key features.



(a) Number of Stroke Incidents



(b) Density Plots for some Key features against Stroke Risk

Figure 2: EDA on data gathered for Stroke Risk Prediction

3 Methods

3.1 Data Pre-processing

The quality and reliability of predictive models hinge on meticulous data preprocessing to handle missing values and identify pertinent features. In this section, we elucidate our comprehensive approach to data preprocessing, ensuring the integrity and efficacy of our predictive framework.

3.1.1 Question 1 - Diabetes Risk

We consider all the features extracted for predicting risk of developing diabetes, as discussed in Section 2.1.

Missing Value Treatment Handling missing data is crucial for maintaining model accuracy and robustness. In our preprocessing, we excluded columns with over 30% missing data. For the remaining data, continuous variables were imputed with the median, and categorical variables with the mode. This approach ensured the integrity of our dataset, resulting in 52,590 records and 192 features, thoroughly prepped for further analysis and modeling.

Feature Selection We employed the RandomForestClassifier Feature Importance technique to identify the most influential features for predicting diabetes incidence.

$$\text{Feature Importance} \quad I(f) = \sum_{t \in T} p(t) \Delta i(s_t, t)$$

Where $I(f)$ is the importance of feature f, $p(t)$ is the proportion of samples reaching node t, and $\Delta i(s_t, t)$ is the impurity decrease from split s_t at node t. The feature importance plot was analyzed to pinpoint the most predictive features for diabetes.

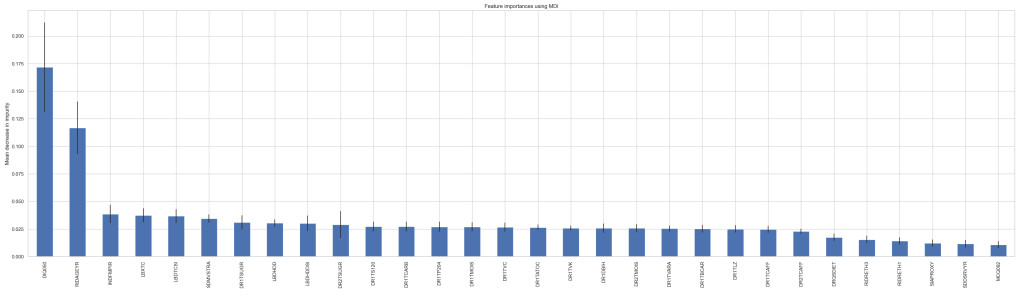


Figure 3: Random Forest Classifier Feature Importance

By leveraging this technique, we extracted the top 30 most important features, thereby reducing dimensionality while retaining critical predictive information. Notable features identified through this process include the patient's dietary indicators like 'sugar intake', 'caffeine intake', 'carbs intake' and other chemical intakes along with 'Age' and 'Family Income'.

Final Dataset Preparation : Prior to model training, we performed one-hot encoding on all categorical columns, a standard preprocessing step essential for rendering categorical variables amenable to machine learning algorithms.

3.1.2 Question 2 - Stroke Risk

We consider all the features extracted for predicting a stroke incidence as discussed in Section 2.

Missing Value Treatment : We perform similar missing value treatment as in Question 1. Following the process for Question 2, the final dataset comprised 14,775 records and 132 features, ready for subsequent analysis and modeling endeavors.

Feature Selection:

3.1.2.1. Logistic Regression with LASSO penalty and Cross Validation :

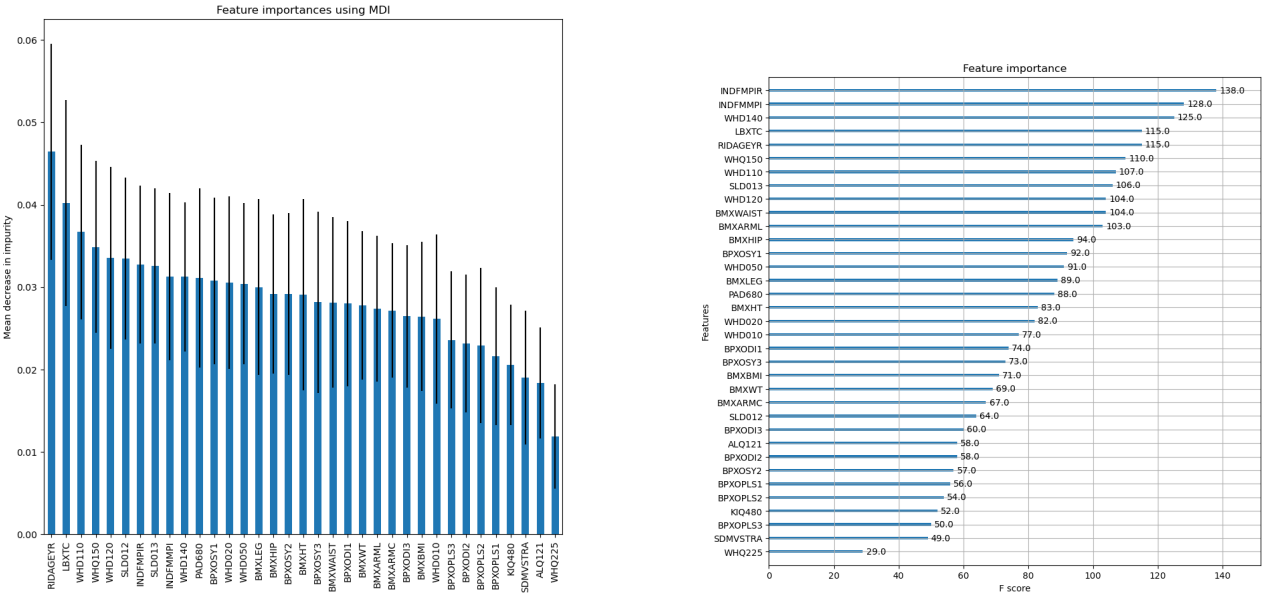
We began by applying a Logistic Regression Model with an L1 penalty, known as LASSO regularization, to the entire dataset. We utilized 5-Fold Cross Validation to identify the Regularization Parameter $C = 0.01$. The LASSO penalty helps in feature selection by shrinking some coefficients towards zero, thus removing less influential predictors. The model is expressed as:

$$\text{minimize} \quad \mathcal{L}(\beta) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] + \lambda||\beta||_1$$

Here, the L1 penalty ($\lambda||\beta||_1$) promotes sparsity in coefficients, automatically selecting significant features. After fitting this model, 30 out of 130 features remained with non-zero coefficients, indicating their importance, while 100 features were excluded as their coefficients shrank to zero.

3.1.2.2. Tree-based Feature Selection :

We utilized RandomForestClassifier and XGBoostClassifier for their robust feature selection capabilities. RandomForestClassifier assesses features using Gini importance, while XGBoostClassifier uses gain-based metrics.



(a) Random Forest Classifier Feature Importance

(b) XGBoost Classifier Feature Importance

Figure 4: Comparison of Feature Importance

We compare the top 50 features selected by each method in Figure 3 and 4, discerning those features that were consistently deemed important across both methodologies. This stringent criterion ensured the retention of features exhibiting robust predictive power, validated using 2 different methods. As a result, we arrive at a curated subset of 42 features. The identified features encapsulate crucial dimensions of the dataset, spanning demographic, physiological, and lifestyle factors pertinent to stroke prediction. Noteworthy among these are 'Age', 'Weight', 'Family Income', 'Physical Activity Data' along with several other features like 'Weight History' of the patients.

Final Dataset Preparation : Prior to model training, we performed one-hot encoding on all categorical columns as we did for Question 1.

3.2 Modeling

3.2.1 Question 1 - Diabetes Risk

Following the identification of 30 key features in Section 3.1.1, we trained a predictive model using the XGBClassifier algorithm. XGBClassifier, or Extreme Gradient Boosting Classifier, utilizes gradient boosting to train decision trees sequentially. Each tree improves upon the errors of its predecessor, enhancing the model's capability to capture complex data relationships and provide accurate predictions. The 30 selected features were used as inputs for the XGBClassifier. We use 20% of the data as Test Set to validate our trained model. This approach not only streamlined the modeling process but also enhanced interpretability by focusing on the most informative predictors. The performance results of this model on the Test Set are discussed in the next section.

3.2.2 Question 2 - Stroke Risk

3.2.2.1. Logistic Regression with LASSO penalty : As discussed in Section 3.1.2, utilizing the trained Logistic Regression Model with LASSO penalty, we proceeded to predict the likelihood of stroke occurrence based on the 30 selected features for identifying Stroke Risk on 25% of the data kept aside as Test Set. This initial modeling approach served as a foundational step in evaluating the predictive capability of the dataset and provided insights into the relevance of individual features in stroke prediction. We discuss the results of this model in the Results section.

3.2.2.2. XGBoost Classifier with Cross Validation : We follow a similar process for modeling XGBoost Classifier as in Question 1 (Section 3.2.1) on features identified as per Section 3.1.2. The 42 selected features for identifying Stroke Risk were used as inputs for the XGBClassifier. We used 5-Fold Cross Validation to evaluate the performance of this model. The results of this model being tested against 25% of the data (Test Set) held-out while training, are discussed in the next section.

4 Results

4.1 Question 1 - Diabetes Risk

The XGBoost classification report for diabetes risk highlights its varied performance across risk categories. While accurately identifying individuals at "No Risk," with high precision and recall (F1-score of 0.96), its performance diminishes for "High Risk" and "Moderate Risk" categories, indicating challenges in correctly identifying individuals at these levels. Despite this, the model achieves an overall accuracy of 0.93. Further refinement may be needed to enhance predictive capability, particularly for higher-risk individuals.

Table 3: Diabetes Risk - Classification Report for XGBoost

Class	**Precision**	**Recall**	**F1-Score**	**Support**
High Risk	0.87	0.48	0.62	983
No Risk	0.93	0.99	0.96	9360
Moderate Risk	1	0.15	0.26	170
Accuracy	0.93			
Macro Avg	0.94	0.54	0.61	10513
Weighted Avg	0.93	0.93	0.92	10513

4.2 Question 2 - Stroke Risk

As discussed in Section 3.2.2, we employed two models on NHANES data to predict Stroke Risk: Logistic Regression with LASSO penalty and XGBoost. The XGBoost Classifier's superior performance, evidenced by higher accuracy and stronger class-wise metrics and an AUROC of 0.91, underscores its efficacy in handling imbalanced datasets and capturing complex data relationships. Conversely, Logistic Regression with LASSO penalty exhibited limitations, particularly in addressing the minority class, resulting in lower precision and recall and an AUROC of 0.79. These findings emphasize the importance of selecting appropriate modeling techniques tailored to the dataset's characteristics for optimal classification. Detailed results are presented below.

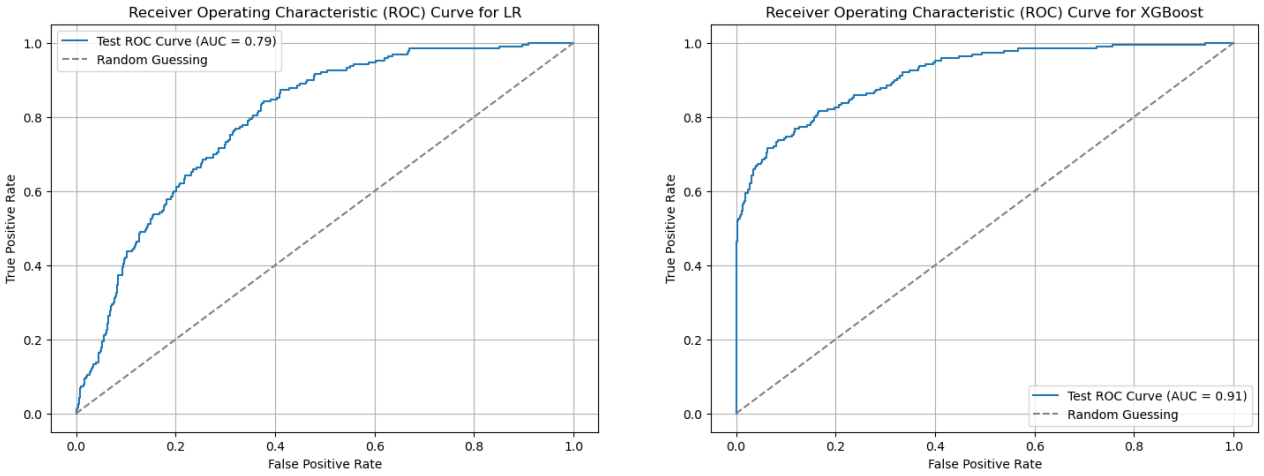
Table 4: Stroke Risk - Comparison of Results from both Models

Class	**Precision**	**Recall**	**F1-Score**	**Support**
No Risk	0.95	1.00	0.97	3504
High Risk	0.29	0.04	0.07	190
Accuracy	0.95			
Macro Avg	0.62	0.52	0.52	3694
Weighted Avg	0.92	0.95	0.93	3694

Table 5: Logistic Regression with LASSO

Class	**Precision**	**Recall**	**F1-Score**	**Support**
No Risk	0.97	1.00	0.99	3504
High Risk	0.90	0.49	0.64	190
Accuracy	0.97			
Macro Avg	0.94	0.75	0.81	3694
Weighted Avg	0.97	0.97	0.97	3694

Table 6: Boosted-Trees (XGBoost)



(a) Logistic Regression AUROC

(b) Random Forest Classifier AUROC

Figure 5: Comparison of AUROC

5 Conclusion

In this project, we successfully identified vital lifestyle, dietary, demographic, and biomarker features for accurate prediction of diabetes and stroke risk. With meticulous data preprocessing and the use of Tree-based models, we achieved impressive weighted F-1 scores of 92% and 97% for Diabetes and Stroke Prediction, respectively. Leveraging techniques such as LASSO penalties, Logistic Regression, Tree-Based Methods, and k-fold Cross Validation contributed significantly to our project.

Improvements:

1. Addressing imbalanced class distributions is crucial for reliable predictions. Imbalanced datasets

can inflate accuracy metrics but lower precision and recall for identifying true stroke instances. While oversampling the underrepresented class can help, caution is needed in healthcare due to data sensitivity. Oversampling may distort the label distribution, requiring careful monitoring.

2. Analyzing healthcare data requires domain expertise. Enhancements to results can be achieved by selecting patient cohorts meticulously and adjusting predictors based on subject-matter expertise. This iterative process refines predictive models for more accurate insights.

6 Contributions

This project was a collaborative effort where each of the three team members played key roles aligned with their strengths, ensuring comprehensive project execution. Below are the summarized contributions:

- **Sudhanshu Agarwal:** Led the data preparation and management, handled preprocessing, and was pivotal in statistical analysis, focusing on decision tree models for diabetes risk prediction.
- **Rishikesh Ksheersagar:** Spearheaded the analysis for stroke incidence using Lasso regression and XGBoost, focusing on identifying significant predictors and interpreting their impacts on stroke risks.
- **Sandeep Jala:** Oversaw the study design and alignment with research objectives, conducted the literature review, and was key in data gathering and pre-processing. Played a key role in analyzing features related to Diabetes Risk Prediction.

Together, the team engaged in regular discussions, reviewed progress comprehensively, and collaborated effectively and equally in all facets of this project to present a scientifically robust final report.

7 Reproducibility

To facilitate the reproducibility of our analysis, here is a succinct guide that outlines the necessary steps:

1. **Download the Datasets:** Access the NHANES website and download the required datasets for our analysis as listed in Section 2.
2. **Data Preparation:** Run the *data_prep.ipynb* Jupyter notebook to clean and prepare the data. This step ensures data is formatted correctly for analysis.
3. **Diabetes Risk Prediction:** Execute the *Q1.ipynb* notebook, which contains the statistical models for predicting diabetes risk based on lifestyle attributes.
4. **Stroke Risk Prediction:** Run the *Q2.ipynb* notebook to analyze key factors associated with stroke incidence and predict Stroke Risk.

We assume successful installation of packages: pandas, numpy, sklearn, xgboost, matplotlib, and seaborn prior to running the analysis Jupyter notebooks.

These steps are designed to be straightforward, ensuring that anyone can replicate our findings efficiently. For further details or troubleshooting, this report includes comprehensive documentation and contact information for assistance.