Table of Contents

[Business Question: 2](#_Toc159248901)

[Data Sources: 2](#_Toc159248902)

[Variables in the dataset: 2](#_Toc159248903)

[Data Quality Concerns: 2](#_Toc159248904)

[External Supporting Materials: 2](#_Toc159248905)

[Methods: 3](#_Toc159248906)

[Preliminary Results: 3](#_Toc159248907)

[Challenges: 3](#_Toc159248908)

[Team Member Contributions: 3](#_Toc159248909)

**Project Proposal**

**Team 1: Chronic Diseases Risk Prediction Initiative**

Team Members: Rishabh Joshi, Akanksha Nangia, Bhakti Chotalia, Akhil Vidiyala, Stuti Saxena

Course: MISM 3515

# Business Question:

How can we leverage advanced analytics and machine learning techniques to develop an accurate and equitable risk prediction model for chronic diseases to guide targeted preventive interventions and improve population health outcomes?

Why is it Important  
Improving chronic disease prediction accuracy through modern data science approaches can guide preventive care prioritization at both individual and population levels. This data-driven approach is key for reducing disease burden, improving health equity, and ensuring healthcare system sustainability.

Who will be interested in the question?  
This refined approach will interest healthcare professionals, policymakers, insurance companies, and individuals at risk of chronic disease. The results could inform policy development, individual health plans, and broader public health strategies.

Business-related Questions:

Our primary quest revolves around understanding diabetes and its risk factors. Specifically, we aim to address:

* What are the primary risk factors for chronic diseases based on both secondary datasets and primary research?
* How well do different modeling algorithms perform in predicting individual chronic disease risk? Comparing multiple models can reveal the best approach for this problem.
* What lifestyle factors and health behaviors emerge as the most significant predictors of chronic disease risk when analyzed through our integrated dataset from the BRFSS and primary research?

Importance & Outcome:

Chronic diseases like heart disease and diabetes impose heavy burdens worldwide, necessitating more effective prevention and management. This modeling aims to empower individuals and policymakers by quantifying personalized chronic disease risk to guide tailored prevention plans. Integrating clinical, socioeconomic, behavioral and environmental data will provide comprehensive risk insights beyond any single factor. Outcomes include accessible risk scores enabling early interventions for high-risk groups, motivating behavior changes through enhanced risk understanding, and strategically informing resource allocation and policy decisions for population health programs. Ultimately, leveraging advanced analytics to uncover actionable chronic disease risk knowledge can drive targeted prevention efforts for improved outcomes at both individual and community levels.

# Data Sources:

We will be publicly accessible data collected by the government of United States. We’re banking on the latest Behavioral Risk Factor Surveillance System (BRFSS) dataset and WHO Global Health Observatory data.

The BRFSS data offers vital dataset variables relevant to public health surveys, particularly focused on individual health statuses and behaviors, and designed for a comprehensive health risk assessment. Key variables include demographics, health behaviors, socio-economic metrics, and detailed health condition information.

* Demographic variables capture individual details like state residence, gender, and household composition.
* Health behavior variables inquire individuals about exercise, sleep, dental visits, and alcohol and tobacco use.
* Health condition variables are detailed and cover a wide range of common and serious diseases, including diabetes, cancer, heart conditions, and mental health.
* Data on health interventions like vaccinations and screenings are also included, alongside socioeconomic factors like income, employment, and housing.

Links to the dataset being used:

* BRFSS 2022 survey data : <https://www.cdc.gov/brfss/annual_data/annual_2022.html>

# Data Quality Concerns:

Data quality concerns for the BRFSS dataset revolve around ensuring data completeness and accuracy.

* Primary concern would be the issue of missingness, where some respondents may not have provided answers to all survey questions, leading to incomplete data.
* Misreporting is another concern since it is self-reported data through phone surveys, where respondents may not accurately recall or may intentionally misrepresent their health behaviors or statuses.
* Outliers or errors resulting from data entry mistakes or misinterpretation of questions by respondents could affect data reliability.
* Sample bias is a concern if the dataset does not adequately represent the population, especially concerning underrepresented groups.

To mitigate these issues, applying techniques like imputation for missing data, outlier detection and handling, and demographic weighting can be crucial.

# External Supporting Materials:

1. Predict Chronic Diseases from Electronic Health Records:  
    [AI in Healthcare to Predict the Chronic Diseases](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181/)
2. A systematic review and meta-analysis:  
    [ML models to predict Chronic Diseases in the USA](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5876976/)
3. Statistics of chronic diseases in the USA:   
   [Statistics and facts of Chronic Diseases in the USA](https://www.statista.com/topics/8951/chronic-disease-prevention-in-the-us/#topicOverview)
4. The Growing Crisis of Chronic Disease in the United States:  
   [Overview of Chronic Diseases](https://www.fightchronicdisease.org/sites/default/files/docs/GrowingCrisisofChronicDiseaseintheUSfactsheet_81009.pdf)

# Methods:

Our approach begins with a rigorous data cleaning and integration process. Acknowledging the paramount importance of high-quality data in predictive analytics, we will deploy advanced techniques for addressing missing values, outliers, and data inconsistencies. This phase ensures the robustness and reliability of our dataset, setting a solid foundation for in-depth analysis.

1. Exploratory Data Analysis (EDA): We will conduct a comprehensive EDA to understand the underlying structure and relationships within the data. This will include statistical summaries, correlation analysis, and initial visualization to identify trends and patterns.
2. Feature Engineering: This stage involves creating new predictive features from the existing data and selecting the most relevant features for the diabetes prediction model. Techniques such as principal component analysis (PCA) and factor analysis might be employed to reduce dimensionality while preserving critical information.
3. Model Selection and Development: We will evaluate and deploy a variety of machine learning algorithms, including but not limited to:

* Logistic Regression: As a baseline model due to its interpretability and effectiveness in binary classification tasks.
* Decision Trees and Random Forests: These models are beneficial for handling non-linear relationships and interactions between variables.
* Gradient Boosting Machines (GBM): XGBoost or LightGBM, known for their high performance in classification tasks by building an ensemble of weak prediction models.

1. Model Optimization: Each model will be fine-tuned using techniques like grid search or random search for hyperparameter optimization. We will also employ methods like cross-validation to ensure that our model generalizes well to new, unseen data.
2. Model Evaluation: We will use a range of metrics to evaluate model performance, such as accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC-ROC). The choice of metrics will be aligned with the specific requirements and goals of our diabetes prediction task.

# Preliminary Results:

Initial data analysis has identified key variables significantly associated with diabetes risk. Ongoing work aims to integrate these insights into the predictive model.

Initial exploration of the BRFSS dataset has revealed promising variables that could significantly enhance the predictive model. Ongoing work aims to integrate relevant variables for further analysis through data wrangling processes, making the dataset clean to build the prediction models.

# Challenges:

**Data Size and Quality Assurance:**

* Managing extensive datasets, notably the Behavioral Risk Factor Surveillance System (BRFSS) alongside primary survey data, requires meticulous data cleaning and integration due to the volume and complexity involved.
* Ensuring data quality through rigorous checks for accuracy, consistency, and completeness is paramount to maintain high standards of data integrity.

**2. Managing Delays and Optimization:**

* The complexity of large datasets introduces delays, particularly during the data wrangling phase.
* Strategies such as process optimization, frequent progress reviews, and a focus on data quality are essential to mitigate these challenges and maintain progress towards project outcomes.

**3. Precision of Risk Scoring Methodology:**

* Ensuring the accuracy of the risk scoring system relies on several key areas, including data quality and diversity.
* Employing robust statistical analysis techniques, including validation methods, is essential to ensure the precision of the risk scoring system.

**4. Model Generalization and Interpretability:**

* It's crucial that sophisticated machine learning models generalize well to new data and remain interpretable, especially in healthcare.
* Overfitting and lack of interpretability, particularly in complex models like gradient boosting machines, pose significant challenges that need to be addressed.

**5. Long-term Impact Assessment:**

* Assessing the long-term impact of interventions based on predictive models for chronic disease risk is challenging.
* It requires ongoing data collection and analysis beyond the project's scope, highlighting the need for a comprehensive approach to evaluation.

# Team Member Contributions:

**Contributions:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Task** | **Akanksha Nangia** | **Bhakti Chotalia** | **Rishabh Joshi** | **Stuti Saxena** | **Akhil Vidiyala** |
| Spearheading the initial stages of data cleaning and merging datasets |  |  | **✓** | **✓** | **✓** |
| Overseeing the modeling phase |  | **✓** | **✓** |  |  |
| Navigating the risk score development | **✓** |  | **✓** |  |  |
| Clustering operations to understand segmentations | **✓** | **✓** |  | **✓** |  |
| Diving into external research |  | **✓** |  |  | **✓** |
| Visualization and Presentation | **✓** |  |  | **✓** | **✓** |