**Executive Summary**

**Project Title:** *Health Risk Prediction Based on Lifestyle and Demographic Attributes*  
**Course:** DSA2040A – Data Mining  
**Group:** 9  
**Team Members:** Alfred, Satcha, Kevin, Bricole, Geoffrey

**1. Introduction**

This project addresses one of the most pressing concerns in public health: predicting chronic illness based on lifestyle and demographic data. With the increasing prevalence of non-communicable diseases (NCDs) like hypertension, diabetes, and cardiovascular conditions, proactive risk assessment has become essential.

Our goal was to simulate a real-world scenario where a healthcare organization could use data mining to predict high-risk individuals and intervene early. Using synthetic data, we developed a full data pipeline — from extraction and transformation (ETL) to mining, visualization, and reporting.

**2. Dataset Description**

We generated a synthetic dataset of 500 individuals, designed to mimic realistic health records. Features included:

* **Demographics**: Age, Gender
* **Lifestyle**: Smoking Status, Alcohol Intake, Exercise Frequency
* **Health Metrics**: Weight, Height, BMI (calculated)
* **Family History**: Indicator of inherited risk
* **Target Variable**: Chronic Illness (binary: 0 = no illness, 1 = at risk)

The data was stored in a structured CSV format and processed in Jupyter Notebooks using Python.

**3. Objectives**

* Identify significant contributors to chronic illness risk
* Build classification models for risk prediction
* Segment individuals into distinct health risk categories
* Visualize data patterns to support decision-making

**4. Methodology**

**4.1 ETL (Extract, Transform, Load)**

* Synthetic data was generated using NumPy and Pandas
* BMI was computed from height and weight
* Data was cleaned and formatted for modeling

**4.2 Exploratory Data Analysis**

* Correlation matrices identified strong relationships between variables
* Distributions and boxplots revealed high-risk groups
* Key features: BMI, Age, Smoking Status, Exercise Frequency

**4.3 Data Mining Techniques**

* **Classification**:
  + **Logistic Regression**: Modeled probability of illness
  + **Decision Trees**: Provided explainable rule-based predictions
* **Clustering**:
  + **K-Means** grouped individuals into 3 risk clusters:
    1. Healthy
    2. Moderate Risk
    3. High Risk (mostly smokers with high BMI and low activity)

**4.4 Visualization**

* Plotly and Seaborn were used to create intuitive visualizations
* Dashboards summarized key trends in health risk

**5. Key Insights**

* Individuals with BMI > 30 had over **60% probability** of being in the at-risk group
* **Current smokers** and **non-exercisers** were significantly more likely to be labeled ill
* The **age group 55+** with a family history of illness had the highest cluster membership in the "High Risk" category
* Clustering showed distinct behavioral and demographic patterns among the population

**6. Limitations & Ethical Considerations**

* **Synthetic data** may not capture the full variability of real-world health datasets
* Health predictions should not be used in isolation from medical expertise
* Ethical handling of personal health data is critical in real applications

**7. Conclusion**

This project illustrates the practical application of data mining in healthcare. By simulating a real-world risk prediction pipeline, we demonstrated how demographic and lifestyle data can guide proactive interventions. With further refinement and integration with clinical systems, such approaches can drive real impact in population health management.

Our work lays the groundwork for future exploration using real clinical datasets and more advanced modeling techniques (e.g., ensemble learning, deep learning).

**8. Future Recommendations**

* Test the model on real-world healthcare data (e.g., NHANES, WHO datasets)
* Incorporate additional variables like diet, sleep, or medical history
* Use time series forecasting to track individual health trajectories
* Deploy the model in a live dashboard for health administrators or clinics