**Data 300 Project Proposal**

**Project Title: Disease Risk Prediction Using Machine Learning Algorithms**

**Team Members:  
Rahma Semma, Nishtha Sharma**

**Description of the project:**

The goal of this project is to forecast the likelihood of illnesses like heart disease and diabetes by analyzing patient information such as age, lifestyle, and health measurements. Algorithms such as Logistic Regression, Random Forest, and Support Vector Machines (SVM) will be utilized to create models that evaluate and forecast disease probabilities using different variables. Our objective is to execute, contrast, and assess these models in order to comprehend how various algorithms manage medical data and forecast disease results. Moreover, college students are experiencing a rise in heart conditions and other chronic illnesses because of stress, unhealthy eating habits, and lack of physical activity. This project aims to address both the technical side of disease prediction and increasing awareness about the common occurrence of these conditions in young adults. Through the utilization of this predictive model, our goal is to inform both college students and the wider community regarding their health vulnerabilities and the significance of detecting and preventing them early on. We will utilize datasets that are accessible to the public, such as the Framingham Heart Study dataset and diabetes datasets sourced from platforms like Kaggle. We will utilize these comprehensive patient datasets to train, test, and validate our machine learning models. Our main objective is to demonstrate how machine learning can enhance healthcare results, particularly for younger demographics.

**Algorithms to be Utilized:**

* **Logistic Regression** for baseline binary classification of disease risk.
* **Random Forest** for identifying complex relationships between biochemical markers and disease risk.
* **Support Vector Machines (SVM)** for more nuanced classification and decision boundary formation.

We will evaluate the algorithms based on performance metrics such as **accuracy**, **precision**, **recall**, and **F1-score**, focusing on their ability to predict disease risk based on biochemical inputs.

**Roles and Contributions of Team Members:**

* **Rahma Semma**: Will focus on the biochemical markers' data preparation and feature engineering, implementing Random Forest algorithms, and visualizing the relationship between markers and disease risk.
* **Nishtha Sharma**: Will handle logistic regression and SVM implementation, parameter tuning, and co-lead on evaluating model performance using economic and demographic variables.
* Both members will collaborate on data cleaning, report writing, and finalizing the project documentation.

**Topics Related to Course:**

* **Machine Learning Algorithms**: Logistic Regression, Random Forest, and SVM.
* **Data Preprocessing**: Cleaning and preparing data, feature selection focused on biochemical markers.
* **Model Evaluation**: Using precision, recall, F1-score, and confusion matrix.
* **Visualization**: Creating clear and intuitive visual representations of the models and their predictions.

**Dataset Description:**

We will use publicly available health datasets, such as the **Framingham Heart Study dataset** and other relevant datasets from **Kaggle** and **UCI Machine Learning Repository**. These datasets contain detailed patient information including biochemical markers (cholesterol, blood pressure, glucose levels) as well as lifestyle factors, which are ideal for building a predictive model of disease risk.

**Schedule of Project:**

* **Week 1-2:** Gather and clean datasets, prepare features related to biochemical markers.
* **Week 3-4:** Implement Logistic Regression and Random Forest models, train models, and begin evaluation.
* **Week 5-6:** Implement SVM, fine-tune parameters, and complete model evaluation.
* **Week 7:** Finalize model evaluation, complete visualizations, and analyze results.
* **Week 8:** Prepare the final report, update the GitHub repository, and develop the project presentation.

**Final Products and Presentation:**

* A fully implemented machine learning model predicting disease risk based on biochemical markers, with detailed analysis of **Logistic Regression**, **Random Forest**, and **SVM**.
* Visualizations showing the importance of biochemical markers in predicting disease risk.
* A poster summarizing the project's goals, methodology, and outcomes.
* A well-documented GitHub repository with code, datasets, and regular updates.

**Plan for Evaluation:**

Our models will be evaluated using the following metrics:

* **Accuracy**: To assess how often the model correctly predicts disease risk.
* **Precision and Recall**: To evaluate how well the model identifies true disease risks while minimizing false positives.
* **F1-Score**: To provide a balanced evaluation of precision and recall.
* **Confusion Matrix**: To understand the distribution of true positives, true negatives, false positives, and false negatives. We will also use visualizations to demonstrate how each model considers biochemical markers in predicting disease risk.

**Objective and Aim of the Project:**

This project aims to not only build a predictive model for disease risk but also to emphasize the role of biochemical markers in disease prediction. By focusing on these markers, we hope to provide valuable insights into the prevention and management of chronic diseases. Additionally, we aim to raise awareness about the impact of health risks in young adults, especially in light of rising cases of heart disease and other chronic conditions. By sharing our findings, we seek to encourage proactive health management in both the college community and the broader public.