STAT 515: Final Project - Progress Report

Due before November 22nd 2024, in-person during office hours. **Hard Copy** must be submitted.

Section Number:	STAT-515-001
Group Number:	10
Names of the Members: (As stated on Blackboard)	1. Deekshitha Reddy Kalluru 2. Rushika Arvapalli 3. Sainath Kammiti

Title of the Project:	A Data-Driven Investigation of Heart Disease Risk Factors and Predictive Models
Describe the data set(s) that were considered: (Provide Background information, including the source and the variables)	**Dataset Considered:** These questions are based on the dataset of heart disease, which includes medical and demographic variables for patients. Variables on cardiovascular health include physical characteristics and diagnostic test results. **Source of Dataset:** This data set was obtained from a publicly available data set in OPEN ML and is presented below with the following variables: 1. **Demographics:** age, sex. 2. **Medical Conditions:** exercise_angina, cholesterol, resting_bp (resting blood pressure) max_heartrate, ST_slope, oldpeak (ST depression). 3. **Diagnostic Results**: ST_slope (response of the ST segment during exercise stress test). 4. **Target Variable:** heart_disease (binary: presence or absence of heart disease).
Explain the reasons why the above data set(s) were selected:	1. Relevant to cardiovascular health, which is of real importance. 2. Mix of continuous and categorical variables available, thus suitable for the application of different machine learning models. 3. Variance that correlates well with the research questions to be asked, such as exercise_angina, cholesterol, age, etc. 4. Complexity could be high in the data hence allowing supervised and unsupervised learning methods.
Research Questions:	1. Are there threshold levels of cholesterol, resting blood pressure, or heart rate above which the risk of heart disease increases dramatically? 2. How does the risk of heart disease change across demographic groups (e.g., age-sex combinations) and health conditions? 3. How do various patient groups of comparable health profiles relate to the risk of heart disease? 4. Are historical peaks of ST_slope and ST depression important indications of heart disease?

	5. Is exercise-induced angina contributing factor to heart disease or merely a symptom?
Explain the rational behind each question and how questions relate to the data set:	1. The question looks for nonlinear interactions between physiological markers and heart disease for identifying thresholds for action. 2. Variations in subgroup analysis might provide insight into disparities in heart disease outcomes across populations. 3. Identifying clusters can expose hidden patterns in the dataset that will provide insight into those groups at greater or lesser risk. 4. Whether certain ECG patterns may serve as early warning signals of cardiac disease. 5. This question will try to show the relationship between exercise and angina-whether it is a cause or an association through the assessment of its predictive power and the relative importance in various variable models.
Provide an overview of the statistical methods that will be used to answer the research questions:	1. Decision Trees and Ridge Regression: Decision Trees:automatically determine threshold splits in continuous variables, such as cholesterol and resting-bp. Ridge Regression: To stabilize coefficient estimates and analyze the impact of continuous predictors. 2. Regression logistic and K-means clustering: Logistic Regression: Interaction terms between age, sex, and the other predictors can be considered. K-Means Clustering: To segment the dataset into distinct demographic and health condition groups. 3. K-Means Clustering and Logistic Regression: K-Means Clustering: To identify natural groupings in the data. Logistic Regression: To assess heart disease risk within each cluster. 4. PCA and Logistic Regression/Random Forest: Principal Component Analysis (PCA): To identify latent patterns between oldpeak and ST_slope. Logistic Regression/Random Forest: From the basis of PCA components for Predictive Modeling. 5. Logistic Regression and Random Forest Logistic Regression: To predict the likeliness of heart disease while controlling other variables. (e.g., max_heartrate, ST_slope). Random Forest: Exploring the importance of exercise_angina in predicting heart disease.

Explain the rationale behind the methods used:	1. Random Forest: Purpose: Non-linear model to deal with intricate interaction and feature importance ranking. Rationale: Handles non-linear relationships and provides insight into important predictors. 2. Ridge Regression: Purpose: Regularized regression to address multicollinearity and stabilize variable estimates. Rationale: Robust estimation of the coefficients for continuous variables. 3. Logistic Regression Purpose: Model binary presence/absence heart disease outcomes Rationale: As one of the simplest and most powerful models, the information of the relations among variables and hypothesis test. 4. Decision Trees Purpose: To illustrate thresholds in continuous predictors. Justification: A simple way for intuition about nonlinear decision boundary. 5. K-Means Clustering. Objective: To separate patients into clusters, based on their demographics and health status. Rationale: Unsupervised learning in order to extract the hidden structure of data. 6. The goal of PCA is to reduce dimensionality in searched-for patterns among linked datasets. Rationale: This allows modeling of complicated relationships for prediction.
Any Challenges or issues faced/facing:	1. Data imbalance: A specific strategy might be the need to oversample or undersample the target variable of interest, heart disease. 2. Multicollinearity: Cholesterol, max heart rate, and resting blood pressure may confuse any regression-type analysis. 3. Interpretability: While powerful, clustering and Random Forests are not as interpretable as linear models. 4. Cluster Validation: Determining the optimal number of clusters for K-Means might be challenging.
Comments (If any):	We can expand the predictors by including additional health factors like BMI, blood sugar levels, or smoking habits to enhance the model's accuracy and scope. Advanced machine learning models like Gradient Boosting or Neural Networks can be explored to improve prediction performance. These generalizations and dependable models may be tested on other datasets externally. Clustering techniques, such as hierarchical clustering, could be developed or extended to make the understandings of patient groupings clear. Furthermore, analytics results have to be interpreted in clinically applicable guidelines in collaboration with healthcare practitioners.