**Educational Research Proposal: Predictive Modeling of Heart Disease Using Clinical Data**

**Introduction**

The growing burden of heart disease worldwide necessitates a deeper understanding of its risk factors and the development of predictive models to aid in early diagnosis and intervention. This project, "Predictive Modeling of Heart Disease," seeks to leverage clinical data to predict the likelihood of heart disease among patients based on various health indicators. Our research explores relationships between key variables, such as cholesterol levels, blood pressure, and age, to build robust models that can forecast heart disease risk.

**The team comprises**

1. Krishan Pandey,
2. Hardeep Gumber,
3. Anvita Iyer,
4. Aleksandra Kutz,
5. Daniel Allen

Each team member will play a critical role in the project's success, with responsibilities ranging from statistical analysis to visualization and presentation delivery.

**Project Objectives**

1. Exploratory Data Analysis (EDA):

- The primary objective of this phase is to investigate the data, identify patterns, detect anomalies, and summarize critical statistics such as mean, median, and standard deviation.

- Assumption: Data is clean and follows a normal distribution for critical variables like cholesterol and blood pressure, allowing for straightforward EDA and hypothesis testing.

2. Statistical Hypothesis Testing:

- Several hypotheses will be tested to determine the statistical significance of observed relationships. For instance, we hypothesize that patients with higher cholesterol levels or resting blood pressure are more likely to have heart disease.

- Assumption: Relationships between key variables like cholesterol, age, and blood pressure will yield statistically significant results, justifying further modeling.

3. Predictive Modeling:

- Develop predictive models such as logistic regression, random forest, and K-nearest neighbors (KNN) to classify the likelihood of heart disease.

- Assumption: Due to balanced class distributions, models trained on these features will achieve high accuracy and generalize well to unseen data.

4. Feature Selection and Dimensionality Reduction:

- Techniques like principal component analysis (PCA) will be applied to reduce noise and identify the most significant predictors of heart disease.

- Assumption: Feature selection will enhance model accuracy without significantly reducing interpretability.

**Methodology**

1. Exploratory Data Analysis (EDA)

- Summary statistics (mean, median, standard deviation) will be calculated for critical variables like age, cholesterol, and resting blood pressure.

- Visualization techniques such as histograms, scatter plots, and correlation matrices will explore the relationships between variables like age and maximum heart rate.

2. Statistical Hypothesis Testing

- Hypotheses developed (from [9] and [10]) will be tested using methods like:

- T-tests to compare means between groups (e.g., males vs. females).

- ANOVA will test differences in cholesterol levels across chest pain types.

- Logistic regression is used to assess the probability of heart disease, given predictors like age and cholesterol levels.

- Example Hypothesis: Older patients are more likely to have exercise-induced angina (Hypothesis 2 from [9]).

3. Predictive Modeling

- Logistic regression will be implemented to classify patients as having heart disease. Cross-validation will be used to tune hyperparameters and assess model performance.

4. Feature Selection and Dimensionality Reduction

- PCA and recursive feature elimination will help identify the most significant predictors, improving model performance and interpretability.

**Task Delegation**

Team Members:

- Alek, Daniel, Krishan: Conduct statistical analysis and upload final code to GitHub [11].

- Anvita, Aleksandra, Hardeep: Create visualizations and provide detailed explanations for each visual [11].

- Krishan, Anvita, Hardeep: Perform descriptive statistics and summarize key findings.

**Expected Outcomes**

*The outcomes of this project will include:*

- Exploratory Insights: Detailed visualizations and summary statistics offering insights into the key variables influencing heart disease.

- Statistical Validation: Hypotheses validated or refuted based on the data, helping identify significant factors related to heart disease.

- Predictive Models: The development of robust models (e.g., random forest, logistic regression) that accurately predict the likelihood of heart disease based on key clinical features.

- Feature Importance: Identification of the most important variables contributing to heart disease risk, such as cholesterol levels and resting blood pressure.

**Conclusion**

This project aims to comprehensively analyze heart disease risk factors, using statistical techniques and predictive modeling to draw meaningful conclusions. By testing hypotheses and building models based on real-world clinical data, we hope to contribute to heart disease prevention, enabling healthcare providers to take timely actions based on patient data. The expected outcome is a set of validated models and actionable insights that can guide future clinical interventions and research efforts.

**Assumptions:**

- The data will be free of significant missing values and follow a normal distribution for key variables.

- Predictive models can generalize to unseen patient data based on sufficient feature selection and dimensionality reduction efforts.