**Capstone Project: Clinical Decision Support Systems (CDSS) for Heart Disease Prediction Using Logistic Regression**

**1. Introduction**

**1.1. Background**

Heart disease remains one of the leading causes of death worldwide. Early detection can significantly improve patient outcomes. Clinical Decision Support Systems (CDSS) play an important role in aiding clinicians with data-driven insights, enabling them to make informed decisions. This project focuses on developing a Python-based CDSS that utilizes Logistic Regression (LR), Random Forest, Neural Networks, data analysis and visualization techniques. LR and Random Forest are simple yet effective machine learning models, to predict the likelihood of heart disease using the UCI Heart Disease dataset available on Kaggle.

**1.2. Objectives**

The objective of this project is to create a CDSS that:  
- Analyzes patient data from the UCI Heart Disease dataset.  
- Utilizes Logistic Regression, Random Forest and Neural networks to predict the presence of heart disease.  
- Visualizes the prediction results and key features contributing to the decision.   
- Provides clinicians with actionable insights to support their decision-making process.

**2. Dataset Overview**

**2.1. Source**

The dataset used in this project is the UCI Heart Disease dataset, which is publicly available on Kaggle. The dataset contains 303 records with 14 features that describe various patient attributes, including demographic and medical information.

**2.2. Features**

The key features in the dataset include: (These features I’m planning to use in my project)  
- Age: Age of the patient.  
- Sex: Gender of the patient (Male, Female)  
- Chest Pain Type (cp): Type of chest pain experienced (4 categories).  
- Resting Blood Pressure (trestbps): Resting blood pressure (in mm Hg).  
- Serum Cholesterol (chol): Serum cholesterol level in mg/dl.  
- Max Heart Rate (thalach): Maximum heart rate achieved.  
- Exercise-Induced Angina (exang): Whether exercise-induced angina is present (1 = yes; 0 = no).  
- ST Depression Induced by Exercise (oldpeak): ST depression relative to rest.  
- Target: Diagnosis of heart disease (Need to figure out the target variable(threshold value) for each feature)

**2.3. Data Preprocessing**

The dataset was preprocessed to ensure it was suitable for machine learning tasks:  
- Handling Missing Values: Missing data were either imputed or removed.  
  
**3. Machine Learning Model:**

**3.1 Logistic Regression**

**3.2. Why Logistic Regression?**

Logistic Regression is ideal for binary classification problems like heart disease prediction. It is simple to implement, interpretable, and provides a probability score for the predicted outcome between two data factors making it suitable for clinical decision support.

**3.3. Model Implementation**

The Logistic Regression model will be implemented using Python's Scikit-learn library. The process involved:  
1. Training: Splitting the dataset into training (80%) and testing (20%) sets.  
2. Evaluation: Assessing the model's performance using metrics such as accuracy, precision, recall, and F1-score. (I need to learn more about how each component works for my project through ML course work)

**3.4. Random Forest**

**3.5 Why Random Forest?**

Random Forest is chosen for the heart disease dataset because it is mostly used for the non-linear relationships between patient features such as cholesterol, blood pressure, age, and exercise-induced angina. It uses multiple decision trees to improve prediction accuracy. Other performances can be measured by using some metrics like accuracy, precision, recall and F1 score.

**3.6 Model Implementation**

Implementation is done by randomly selecting features such as cholesterol, age, and blood pressure to build multiple decision trees. Each tree will independently predict the presence of heart disease, and the final prediction will be based on the majority vote of the trees.

**3.7 Neural Networks  
  
3.8 Why Neural Networks?**

This model captures the complex interactions between multiple patient attributes, such as how age combined with cholesterol levels affects heart disease risk. Neural Networks can improve prediction accuracy by learning these metrics patterns, still interpreting the risk remains challenging in clinical decision making.

**Model Implementation**

Neural Networks are all made-up of layers of connected units called neurons which help to learn metrics or patterns in data. The features process the data through layers to understand the relationships and outputs whether a person has heart disease.

**4. System Design and Implementation**

**4.1. Data Analysis and Visualization**

Detailed analysis and visualization of the dataset need to be conducted to understand the relationships between different features and the target variable (threshold value) for each key feature, so that the predicted factors can be compared with each target variable and decision can be taken not on one factor, but on multiple factors. These multiple factors help for predicting heart diseases early rather than advanced stages. After data analysis, the values are then visualized to understand the patterns of each behavior. Techniques like bar charts, scattered plots, correlation matrices, and histograms were used to explore these relationships.

**4.2. Decision Support Module**

The CDSS uses the trained Logistic Regression, Random Forest and Neural networks models to provide predictions on heart disease. It outputs a probability score along with key features influencing the decision, helping clinicians to understand the underlying factors.

**5. Insights to Clinicians**

This project will develop a CDSS for heart disease prediction using Logistic Regression and other ML models. The system provides interpretable predictions that can assist clinicians in early detection and management of heart disease. Interpretable models like LR and Decision trees allow clinicians to see how each feature (age, cholesterol, etc.) influences the prediction, enabling them to trust and act upon the situations and results. By providing not just predictions but explanations of how certain patient factors (such as high cholesterol or blood pressure) lead to higher risks of heart disease, which helps for early detection and management. The visualization and neural networks will help for accurate predictions because of the results from specific variables. With further refinement, the CDSS can become a valuable tool in clinical settings.  
  
**6. Challenges:**

1. Even though the UCI heart disease dataset is a popular one, it has some imbalanced data like more patients are there with no heart disease than those with heart disease.

2. Selection of features is very important. Combining factors such as BP and Cholesterol are very useful but comparing them with other features is challenging.

**7. Conclusions and Discussion**

**Milestone 2:** After implementing all three models, we will start comparing each metric for each model, and then we will perform cross-validation to evaluate the robustness of each model. So, these comparisons will show whether Logistics Regression, Random Forest, or Neural networks perform better for heart disease prediction.

After analysis and visualization, proper conclusions need to be made, and future directions need to be documented.

**8. Reference articles**

1. Ayon, S. I., Islam, M. M., & Hossain, M. R. (2022). Coronary artery heart disease prediction: A comparative study of computational intelligence techniques. IETE Journal of Research, 68(4), 2488-2507.

2. Alfadli, K. M., & Almagrabi, A. O. (2023). Feature-limited prediction on the UCI Heart Disease dataset. Computers, Materials & Continua, 74(3).

3. Rawson, T. M., Moore, L. S. P., Hernandez, B., et al. (2017). A systematic review of clinical decision support systems for antimicrobial management. Clinical Microbiology and Infection, 23(8), 524-532.

4. Jiang, F., Jiang, Y., Zhi, H., et al. (2017). Artificial intelligence in healthcare: Past, present and future. Stroke and Vascular Neurology, 2(4), 230-243.

5. Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., & Sun, J. (2016). Doctor AI: Predicting clinical events via recurrent neural networks. Journal of Machine Learning Research, 17(1), 2762-2781.

6. Saqlain, M., Mushtaq, I., Sufyan, M., et al. (2021). Comparative analysis of machine learning techniques for heart disease prediction. International Journal of Biomedical and Health Informatics, 8(2), 101-111.

7. Khatibi, T., Razavi, S. M., & Shakiba, B. (2020). Application of machine learning methods to predict cardiovascular disease using clinical data. Journal of Research in Medical Sciences, 25, 85.

8. Alaa, A. M., Yoon, J., Hu, S., & van der Schaar, M. (2017). Personalized risk scoring for critical care prognosis using mixtures of Gaussian processes. IEEE Transactions on Biomedical Engineering, 65(1), 207-218.

9.<https://www.researchgate.net/publication/371142044_Predicting_Heart_Disease_using_Logistic_Regression>

10. <https://www.nature.com/articles/s41598-024-58489-7>