**Title: Prediction of heart disease using LogisticRegression**

# **Abstract:**

Cardiovascular diseases remain a leading cause of mortality worldwide, necessitating effective predictive tools for timely diagnosis and intervention. This project employs machine learning techniques, specifically Logistic Regression, to predict the presence of heart disease based on a comprehensive dataset from 1988.

**About Dataset**

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "Heart Diseases" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

Attribute Information:

1. age
2. sex
3. chest pain type (4 values)
4. resting blood pressure
5. serum cholestoral in mg/dl
6. fasting blood sugar > 120 mg/dl
7. resting electrocardiographic results (values 0,1,2)
8. maximum heart rate achieved
9. exercise induced angina
10. oldpeak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment
12. number of major vessels (0-3) colored by flourosopy
13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

# **Methodology**

### Data Preprocessing:

Initial data exploration checks for missing values and duplicates. Incorrect values, such as 'M' and 'F' for gender, are replaced with numerical equivalents. The dataset is split into features (X) and the target variable (y). Categorical features are encoded, and a preprocessing pipeline is constructed.

Data Visualization:

Visualizations, including histograms, boxplots, and pairplots, aid in understanding feature distributions and relationships. Heatmaps illustrate the correlation matrix, guiding feature selection and model interpretation.

### Model Building:

Logistic Regression is chosen for its suitability in binary classification tasks. A preprocessing pipeline is created, integrating feature scaling and one-hot encoding.

### Model Training and Evaluation:

The dataset is split into training and testing sets. The Logistic Regression model is trained on the former and evaluated on the latter. Model scores, accuracy, and F1 score metrics are computed to assess performance.

### Results:

The Logistic Regression model achieves a commendable accuracy score of 86.4%, indicating its effectiveness in predicting heart disease. F1 score, a metric considering both precision and recall, is calculated at 0.88.

### Model Deployment (Simulation):

A simulation demonstrates the model's practical use. User input for health parameters is processed through the preprocessor, and the model predicts the likelihood of heart disease. This user-friendly interface enhances awareness and facilitates proactive health management.

# **Conclusion:**

The Logistic Regression model, trained on historical data, emerges as a valuable tool for predicting heart disease. Its transparency and interpretability make it accessible for both healthcare professionals and individuals seeking insights into their cardiac health.