2. Heart Disease Prediction (heart.ipynb)

- **Problem Statement**: The project aims to predict the presence or absence of heart disease in patients based on various medical attributes. This is a binary classification problem.
- **Dataset Used**
  :
  - The dataset was loaded from heart.csv.
  - The target variable is 'target' (likely 0 for no disease, 1 for disease).
  - Features include 'age', 'sex', 'cp' (chest pain type), 'trestbps' (resting blood pressure), 'chol' (serum cholesterol), 'fbs' (fasting blood sugar), 'restecg' (resting electrocardiographic results), 'thalach' (maximum heart rate achieved), 'exang' (exercise induced angina), 'oldpeak' (ST depression induced by exercise relative to rest), 'slope' (the slope of the peak exercise ST segment), 'ca' (number of major vessels colored by fluoroscopy), and 'thal' (thalassemia type).

- **Methodology and Approach**
  :
  - **Data Preprocessing**
    :
    - Checked for missing values using isnull().sum(); no missing values were found.
    - Checked for duplicated rows and removed them.
    - No explicit feature scaling was performed in the final model training pipeline shown, though it's often beneficial for distance-based or gradient-based algorithms.

  - **Model Training**
    :
    - The data was split into training (70%) and testing (30%) sets.
    - Several classification models were trained and evaluated:
      - Logistic Regression (LogisticRegression)
      - Support Vector Classifier (SVC)
      - K-Nearest Neighbors (KNeighborsClassifier)
      - Decision Tree Classifier (DecisionTreeClassifier)
      - Random Forest Classifier (RandomForestClassifier)
      - Gradient Boosting Classifier (GradientBoostingClassifier)

    - XGBoost Classifier (XGBClassifier)

- **Model Evaluation**
  :
  - Models were primarily evaluated using accuracy scores.
  - Classification reports and confusion matrices were generated for each model.
  - ROC AUC scores and ROC curves were also computed and plotted for each model.

- **Results and Conclusion**
  :
  - The models achieved varying accuracies, with Random Forest showing 90.11%, Logistic Regression 86.81%, SVC 87.91%, K-Nearest Neighbors 89.01%, Decision Tree 78.02%, Gradient Boosting 84.62%, and XGBoost 85.71%.
  - Based on the accuracy and ROC AUC scores, the Random Forest classifier and K-Nearest Neighbors performed strongly.
  - The notebook concludes by selecting Random Forest as a strong candidate and saves this model to heart_disease_model.pkl.