Heart Disease Prediction Web App – Project Documentation

Dataset Description

- **File**: `heart_problem.csv`
- **Content**: Medical records of patients with various health indicators such as:
  - `age, sex, anaemia, diabetes, high_blood_pressure, creatinine_phosphokinase, ejection_fraction`, etc.
- **Target Variable**: DEATH_EVENT (binary classification: 0 = survived, 1 = died)
- **Cleaning Steps**:
  - Duplicate rows removed
  - Missing values handled using `SimpleImputer`

ML Model Training

- **Notebook**: `heart_prob.ipynb`
- **Preprocessing**:
  - Encoding: OneHotEncoding for categorical variables
  - Scaling: StandardScaler
  - Feature selection using `SelectKBest` and `RFE`
- **Models Trained**:
  - Logistic Regression (final model used)
  - KNN, Decision Tree, Random Forest (for comparison)
- **Evaluation Metrics**:
  - Accuracy
  - Confusion Matrix
  - R^2 score, MSE, MAE (regression metrics were also tested)
- **Model Persistence**:
  - Saved using `joblib` as `logistic_model.pkl`

Authentication System

- **Framework**: Django's built-in auth system
- **User Interface**:
  - `signup.html, login.html, logout` views
- **Logic** (`views.py`):
  - User registration and login using `AuthenticationForm`
  - Custom `logout` function with redirection
- **Forms**:
  - `HeartProblemForm`: collects medical and personal data via form inputs

Integration Workflow

- **Backend**: Django views handle data processing and model predictions
- **Frontend**: HTML templates with Bootstrap-like styling
- **Prediction Flow**:
    1. User submits health data via form
    2. Backend loads `logistic_model.pkl`
    3. Features processed and fed into model
    4. Result rendered on `result.html`

⚠ Challenges Encountered

- **Data Quality**:
    - Needed to remove duplicates and handle missing values
- **Model Selection**:
    - Balancing performance with interpretability; logistic regression chosen for simplicity
- **Integration**:
    - Mapping form fields to model inputs required careful alignment
- **Deployment Concerns**:
    - Model file paths and CSV loading had to be handled carefully to avoid path issues