AI-Powered Heart Disease Assessment App

Unmatched Accuracy with Cutting-Edge Machine Learning Models

University of San Diego

Data Wrangling and Pre-Processing

- **Features selecting**
  - Selected only the features that impacting heart disease.
- **Data Cleaning**:
  - Handled missing values using Distribution-Based Imputation to preserve data distribution and variability.
  - Converted numerical variables to categorical.
- **Feature Engineering**:
  - Created new features such as BMI categories, age groups, and health status indicators.

Exploratory Data Analysis

- **Data Visualization**:
  - Analyzed distributions of categorical data using bar plots against target variable.
  - Identified key features impacting heart disease risk.
- **Data Encoding**:
  - Encoding categorical features into numerical using Catboost to improve modeling and assess linear and non-linear relationships between features and target variables

Modeling

- **Model Evaluation**:
  - Evaluated 9 models to determine the best performing algorithm for heart disease risk prediction.
  - Selected LightGBM for its efficiency, speed, and high performance with large datasets.
  - Chose EasyEnsembleClassifier for its ability to handle class imbalance effectively.
- **Final Model**:
  - Combined EasyEnsembleClassifier as a wrapper with LightGBM as the base estimator.
  - This combination outperformed both models individually, improving recall for the minority class (heart disease patients) from 24% to 81.3%.
- **Model Training**:
  - Split data into training and testing sets.
  - Tuned hyperparameters using OPTUNA, further enhancing recall metrics.
- **Model Evaluation**:
  - Focused on predicting true positives (heart disease patients), emphasizing recall and using ROC AUC to balance sensitivity and specificity.

App Development

- **Streamlit Framework**:
  - Utilized Streamlit to create an interactive and user-friendly web app.
  - Configured the app layout and design for better user experience using Python, HTML and CSS
- **Model Integration**:
  - Loaded the trained machine learning model and encoder.
  - Implemented functions to preprocess user input and predict heart disease risk.
- **User Interaction**:
  - Designed input forms for users to enter their health information.
  - Displayed personalized risk scores and actionable recommendations based on model predictions.