Heart Disease Prediction: A Complete Data Science Project

1. Executive Summary

This project presents a comprehensive machine learning solution for predicting the presence of heart disease in patients. The pipeline covers all key stages of a data science project, from data preprocessing and model training to deployment in a user-friendly web application. The final model, a hyperparameter-tuned Random Forest classifier, demonstrates strong predictive performance, making it a reliable tool for preliminary health screening.

2. Project Pipeline

Data Preprocessing

The initial dataset was loaded and cleaned to handle missing values and prepare the features. This step included:

- Handling missing values by filling them with a specific value or removing the corresponding rows.
- Dropping irrelevant columns to reduce noise.
- Encoding categorical features into a numerical format suitable for machine learning models.

Feature Engineering & Selection

To improve model efficiency and performance, two methods were explored for feature selection:

- **Manual Feature Selection:** Columns were carefully selected based on their relevance to the target variable (num).
- **Principal Component Analysis (PCA):** This dimensionality reduction technique was applied to identify the most significant features and reduce the feature space. The manually selected features proved to be more effective for this specific problem, and were used in the final pipeline.

Model Training & Evaluation

Multiple supervised learning models were trained on the preprocessed data, including **Decision Tree** and **Random Forest**. The performance of each model was evaluated using a key metric:

- **F1-Score:** This metric was chosen because it provides a balanced measure of a model's precision and recall, which is crucial for medical prediction tasks where both false positives and false negatives are critical. The **Random Forest Classifier** consistently outperformed the other models and was chosen for the final solution.

Hyperparameter Tuning

To optimize the Random Forest model's performance, **GridSearchCV** was used. This method systematically tests various combinations of hyperparameters to find the optimal set that yields the best possible performance on unseen data. The optimized model was then saved as a .pkl file for future use.

3. Deployment & User Interface

A functional web application was developed using **Streamlit**, providing a simple and intuitive user interface. This app allows a user to:

- **Input** their personal health data (age, cholesterol, etc.).
- Receive an **instant prediction** on the likelihood of having heart disease.
- **Visualize** their data point relative to the training dataset, allowing for a better understanding of their position within the health data distribution. The app was successfully deployed using **Ngrok**, making it publicly accessible for live demonstration and use.

4. Final Deliverables

This project successfully delivered the following key artifacts:

- ✅ **Cleaned dataset** with selected features.
- ✅ **Trained supervised models** and performance evaluation metrics.
- ✅ **Hyperparameter optimized model** saved in a .pkl format.
- ✅ **Functional Streamlit UI** for real-time predictions.

- ✅ **GitHub repository** with all source code and documentation (README.md, requirements.txt).
- ✅ **Publicly accessible live app** via a Ngrok link.

Author: Omnia Mohamed Ghazy