FLIKT TECHNOLOGY WEB SOLUTIONS

Flikt Technology web solution

Project Report :- Heart Disease Prediction & Analysis Web Application

Submitted to :- Vanshika ji

Submitted by :- Devansh Kumar

1. Executive Summary

This report details the end-to-end development of a machine learning-powered web application designed for the analysis and prediction of heart disease. The project encompasses a full data science workflow, including data preprocessing, exploratory data analysis (EDA), model training with hyperparameter optimization, and deployment via an interactive Streamlit dashboard. The primary outcome is a robust **Random Forest Classifier** that predicts patient risk with approximately **96% accuracy**. The final application serves as a practical tool for data-driven risk assessment, providing users with actionable insights and personalized health recommendations based on their clinical and lifestyle inputs.

2. Introduction

Cardiovascular diseases are a leading cause of mortality globally, making early and accurate risk identification a critical public health objective. This project addresses this challenge by leveraging machine learning to analyze complex patient data and identify high-risk individuals proactively. The goal is to translate a predictive model into a user-friendly tool that can aid in educational and preliminary screening contexts, promoting health awareness and timely medical consultation.

3. System Architecture & Dataset

3.1. Technical Architecture

The system is architected as a Python-based web application with two core components:

- **Machine Learning Backend:** A pre-trained Scikit-learn model (best_heart_disease_model.joblib) handles prediction logic. It is supported by a data preprocessing pipeline to ensure input data is correctly formatted.
- **Streamlit Frontend:** An interactive user interface that provides modules for data visualization and real-time risk prediction.

3.2. Dataset Description (heart_disease_dataset.csv)

The model was trained on a comprehensive heart disease dataset containing anonymized patient records. The dataset is characterized by a mix of numerical and categorical features, with **"Heart Disease"** (0 for No, 1 for Yes) serving as the binary target variable. Key features include:

Category	Features
Demographic	Age, Gender

Clinical Metrics	Cholesterol, Blood Pressure, Heart Rate, Blood Sugar
Behavioral Factors	Smoking, Alcohol Intake, Exercise Hours, Stress Level
Medical History	Family History, Diabetes, Obesity, Exercise Induced Angina

The dataset was found to be well-balanced, making it suitable for training a classification model without significant class imbalance issues.

4. Data Processing & Analysis

4.1. Data Preprocessing Pipeline

A standardized preprocessing pipeline was developed to prepare the data for modeling:

1. **Missing Value Imputation:** Null values in the Alcohol Intake feature were imputed with the string "None" to treat them as a distinct category.
2. **Categorical Encoding:** sklearn.preprocessing.LabelEncoder was applied to transform all categorical features (e.g., Gender, Smoking) into a numerical format suitable for the model.
3. **Feature Scaling:** All numerical features were standardized using sklearn.preprocessing.StandardScaler, which normalizes the data by removing the mean and scaling to unit variance. This prevents features with larger scales from dominating the model training process.

4.2. Exploratory Data Analysis (EDA)

The web application's visualization module provides several interactive EDA plots generated using Plotly and Seaborn, including:

- **Univariate Analysis:** Histograms to show the distribution of key numeric features (Age, Cholesterol).
- **Bivariate Analysis:** Count plots to explore the relationship between categorical predictors and the heart disease outcome.
- **Multivariate Analysis:** A correlation heatmap to visualize the linear relationships between all numerical features.

5. Predictive Modeling

5.1. Model Selection & Training

A **Random Forest Classifier** was selected as the final model due to its high performance and robustness.

The model was trained on an 80/20 train-test split of the preprocessed data.

5.2. Hyperparameter Tuning

GridSearchCV was employed to systematically identify the optimal hyperparameters for the Random Forest model, maximizing its predictive power.

5.3. Performance Evaluation

The model's performance on the unseen test set was excellent, validating its reliability:

- **Overall Accuracy:** ~96%
- **Macro F1-Score:** ~0.96
- **Precision & Recall:** The model demonstrated a strong balance of precision and recall for both positive and negative classes, indicating it is effective at correctly identifying patients with and without heart disease.

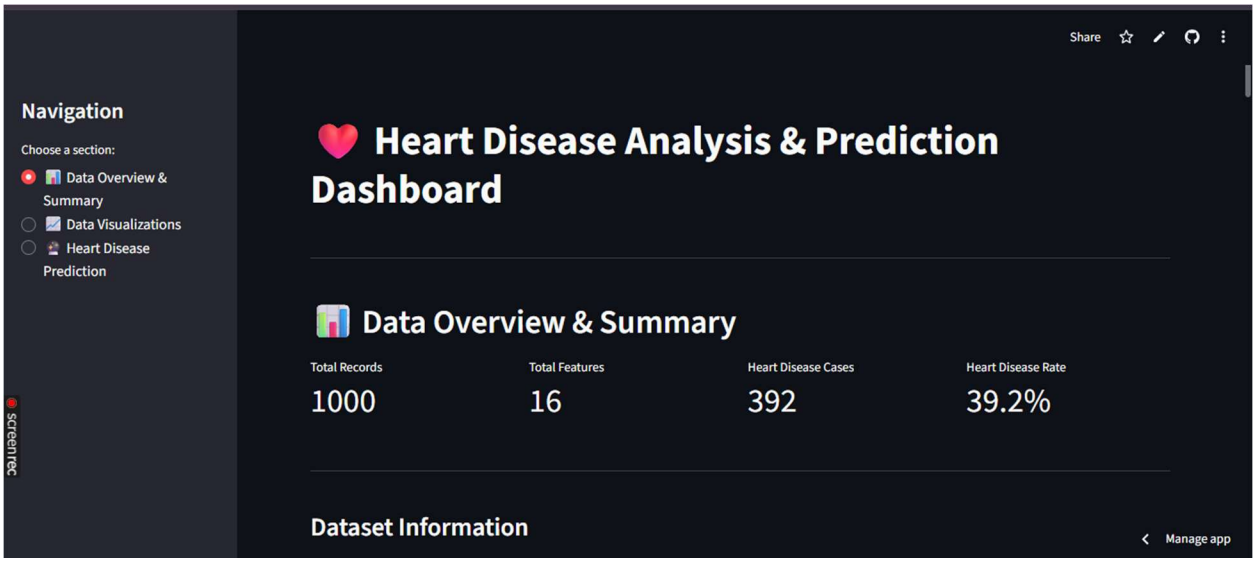
The finalized, trained model was serialized to best_heart_disease_model.joblib for deployment.

6. Deployed Application & User Interface

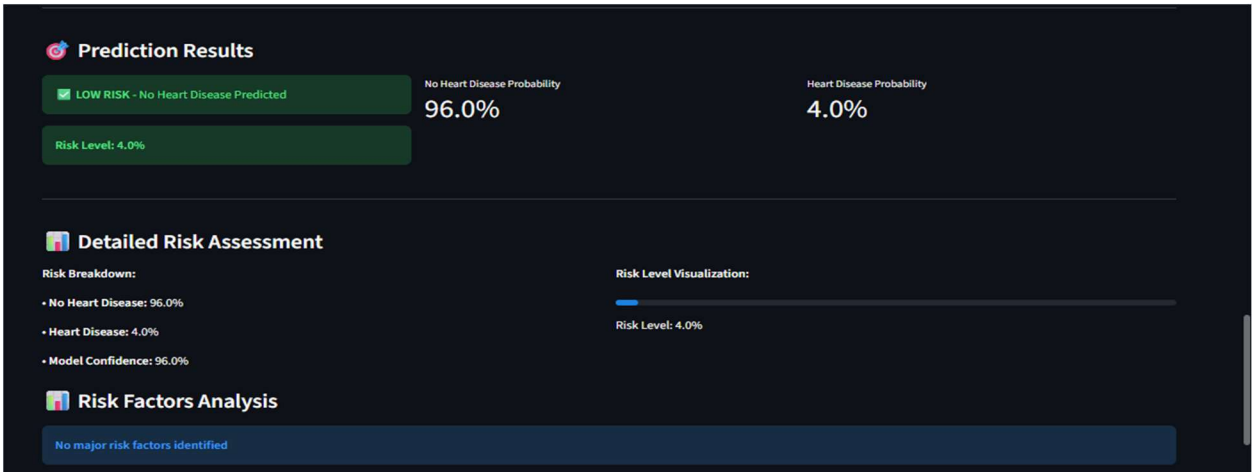
The application is deployed as an interactive Streamlit dashboard featuring:

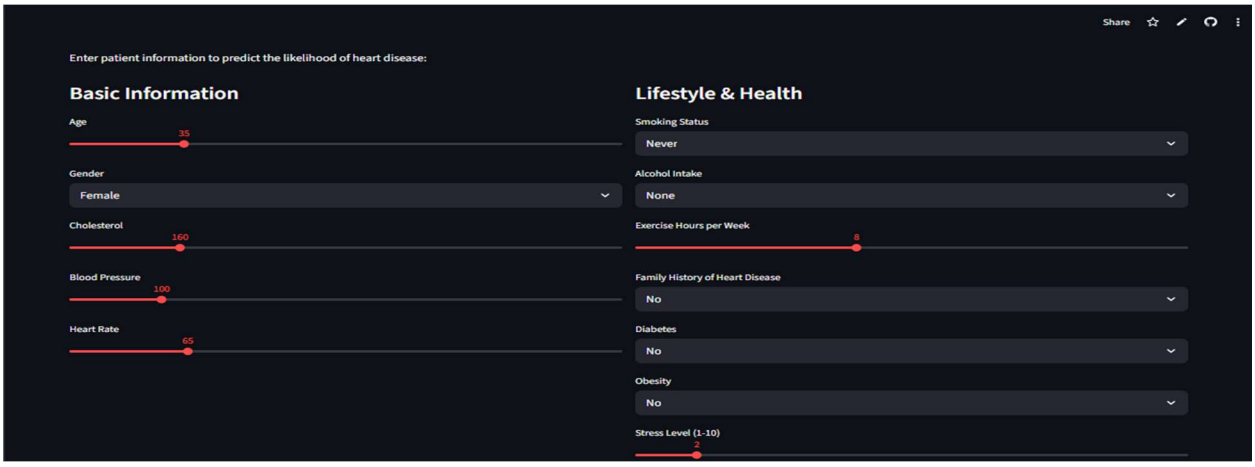
- **Modular Navigation:** A sidebar allows users to switch between the project overview, data visualizations, and the prediction engine.

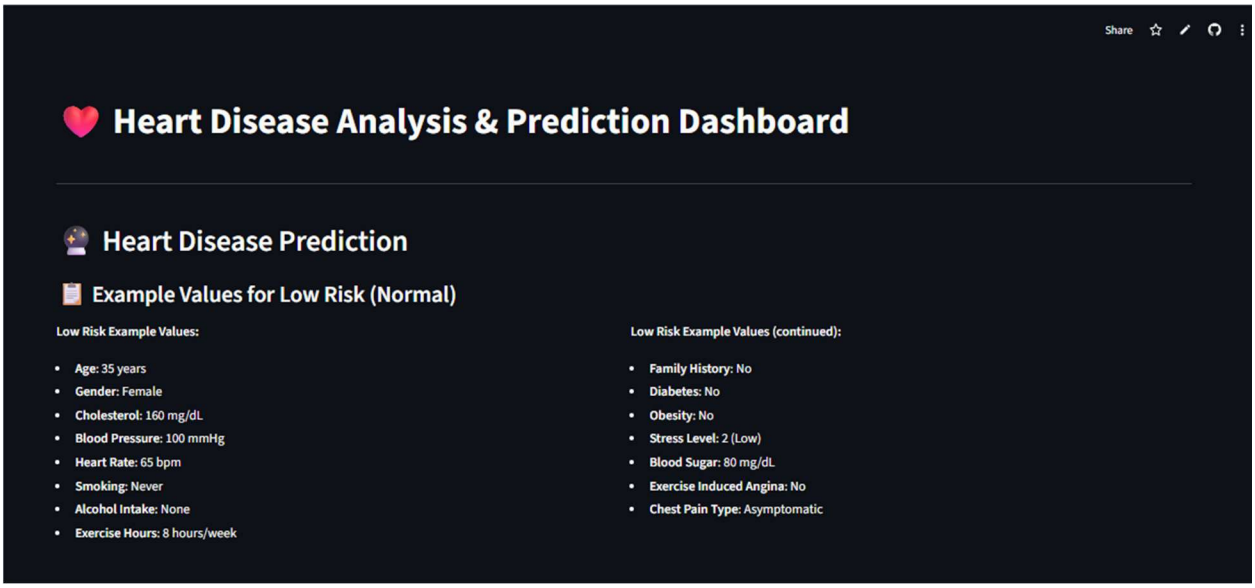
	Age	Gender	Cholesterol	Blood Pressure	Heart Rate	Smoking	Alcohol Intake	Exercise Hours	Family History	Diabetes	Obesity	S
0	75	Female	228	119	66	Current	Heavy	1	No	No	Yes	
1	48	Male	204	165	62	Current	None	5	No	No	No	
2	53	Male	234	91	67	Never	Heavy	3	Yes	No	Yes	
3	69	Female	192	90	72	Current	None	4	No	Yes	No	
4	62	Female	172	163	93	Never	None	6	No	Yes	No	
5	77	Male	309	110	73	Never	None	0	No	Yes	Yes	
6	64	Female	211	105	86	Former	Heavy	8	Yes	Yes	Yes	
7	60	Female	208	148	83	Never	Moderate	4	No	Yes	Yes	
8	37	Female	317	137	66	Current	Heavy	3	No	Yes	Yes	
9	63	Male	204	141	68	Former	Heavy	8	No	Yes	No	



- **Dynamic Prediction Form:** A user-friendly form with sliders and dropdown menus for inputting patient data.
- **Results Interpretation:** The application outputs:
  - A clear risk classification ("Low Risk" or "High Risk").
  - A list of identified risk factors based on the user's inputs (e.g., "High Blood Pressure," "Obesity").
  - Actionable, context-aware recommendations (e.g., suggesting a consultation for high-risk individuals).



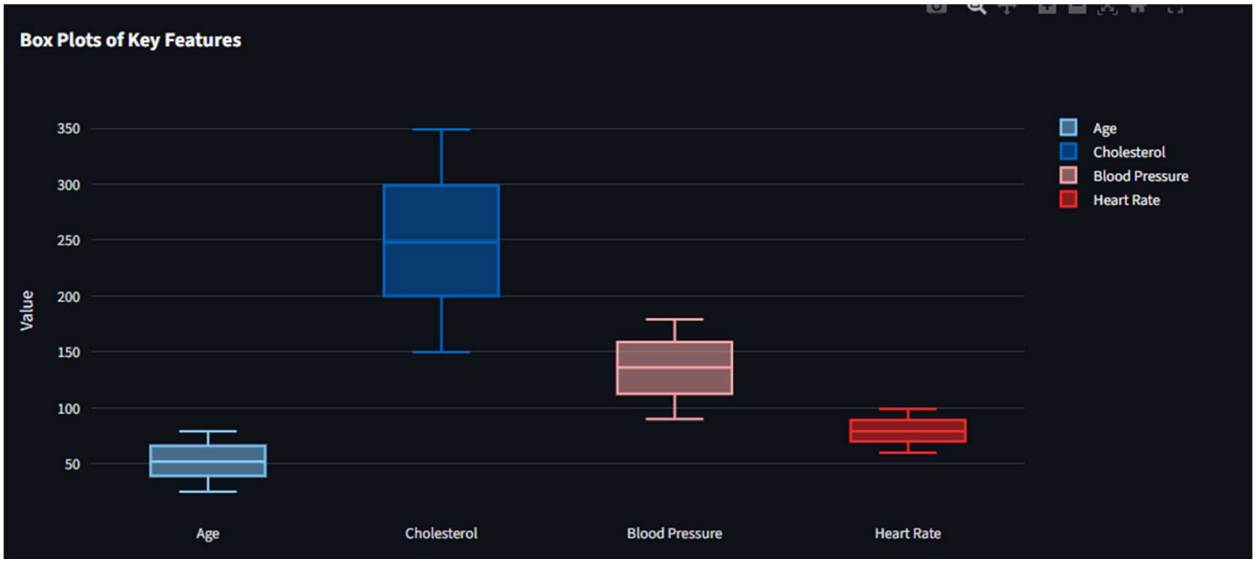




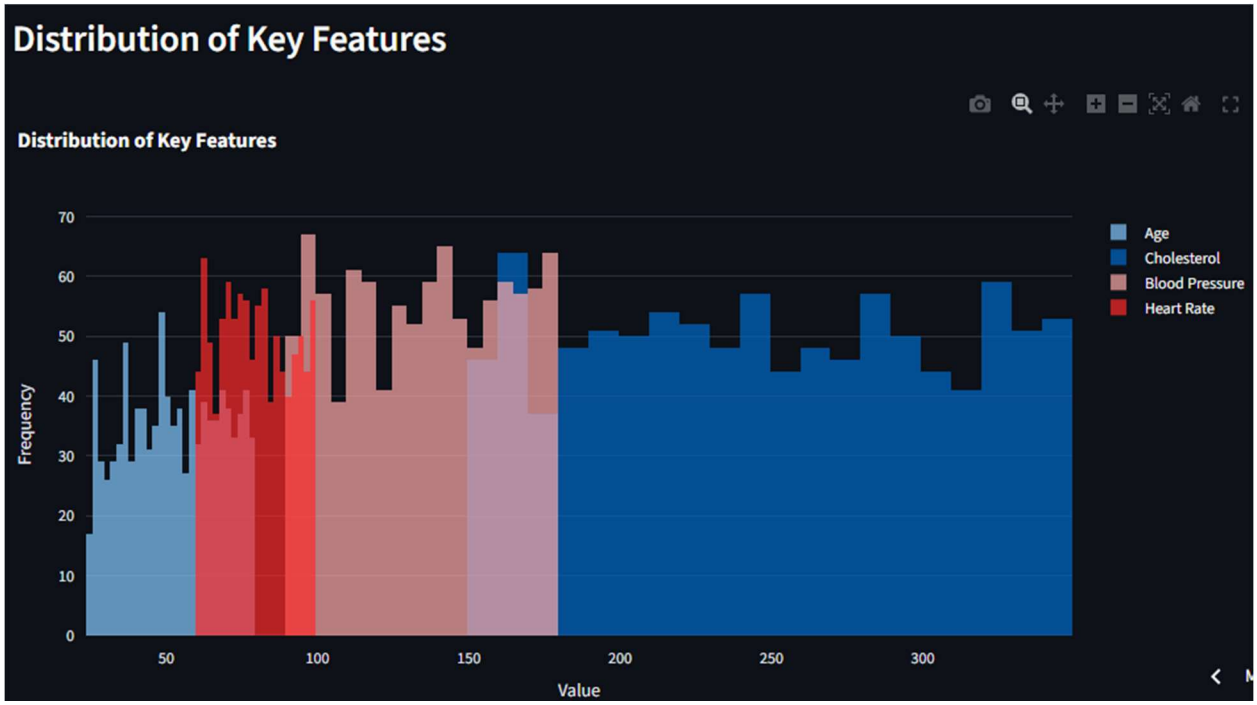
- **Demonstration Scenarios:** Pre-filled buttons for "Low Risk" and "High Risk" examples to showcase the tool's functionality.

7. Conclusion

This project successfully demonstrates the creation of a full-stack data science application, from raw data to a deployed, interactive tool. The Heart Disease Prediction Web Application stands as a robust proof-of-concept that effectively integrates a high-accuracy machine learning model with a user-centric interface. It serves as a valuable resource for educational purposes and highlights the potential of AI in promoting proactive health management.




Correlation Heatmap of Numerical Features
x: Blood Pressure
y: Age
color: 0.002092796
Age
Cholesterol
Blood Pressure
Heart Rate
Exercise Hours
Stress Level
Blood Sugar
Heart Disease
1
0.8
0.6
0.4
0.2
0
screenrec


Distribution of Key Features
Distribution of Key Features
Frequency
Value
Age
Cholesterol
Blood Pressure
Heart Rate
0
10
20
30
40
50
60
70
50
100
150
200
250
300