End-to-End Machine Learning Pipeline: Heart Disease Prediction

# 1. Data Handling (NumPy & Pandas)

The dataset was loaded into a Pandas DataFrame using Google Colab's file upload feature.  
Initial checks were performed using `.head()`, `.info()`, and `.isnull().sum()` to inspect data quality.  
Duplicate records were removed, and missing values were handled. Categorical features were encoded  
into numerical values using Pandas' `astype('category').cat.codes` functionality.

# 2. Exploratory Data Analysis (EDA)

Basic statistical analysis was conducted using NumPy and Pandas (`.describe()`).  
Visualizations included:  
- A correlation heatmap (Seaborn) to highlight relationships between features.  
- An interactive Plotly scatter plot showing Age vs Cholesterol colored by Target.  
Key findings showed strong correlations between age, cholesterol, and the presence of heart disease.

# 3. Feature Engineering

Features (X) were separated from the target (y). Data was standardized using StandardScaler to ensure features  
were on a similar scale. The dataset was split into training (80%) and testing (20%) sets using train\_test\_split.

# 4. Model Training

Three classification models were trained:   
- KNN Classifier  
- Decision Tree Classifier  
- Random Forest Classifier  
  
Baseline results were compared on the test set using Accuracy, Precision, Recall, and F1-score.  
Random Forest outperformed the others in all metrics.

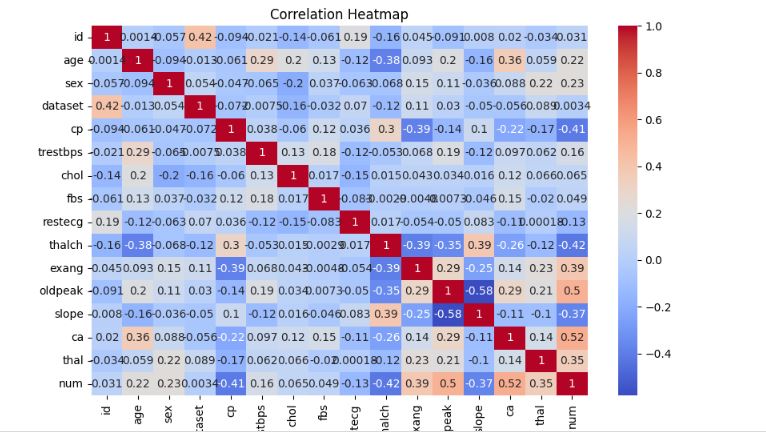
# 5. Feature Importance

Random Forest feature importances were extracted and visualized using a horizontal bar chart.  
The most important features contributing to predictions included: Age, Cholesterol, and Resting Blood Pressure.

# 6. Hyperparameter Tuning

RandomizedSearchCV was used for:  
- KNN: n\_neighbors, weights, metric  
- Decision Tree: max\_depth, min\_samples\_split  
- Random Forest: n\_estimators, max\_depth, min\_samples\_split  
  
The tuned Random Forest model improved performance compared to its default configuration.

# 7. Model Evaluation

Metrics used: Accuracy, Precision, Recall, and F1-score.  
A confusion matrix was plotted to visualize classification performance.  
The ROC curve of the best Random Forest model achieved a high AUC, demonstrating strong predictive power.

# 8. Conclusion

Random Forest proved to be the most effective model due to its ability to handle complex feature interactions   
and provide consistent results. Hyperparameter tuning significantly enhanced accuracy and reduced overfitting.  
Key predictors of heart disease were Age, Cholesterol levels, and Resting Blood Pressure.  
This pipeline demonstrates a robust workflow from raw data to a tuned predictive model suitable for clinical applications.

# Model Comparison Table (Baseline Metrics)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-score |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
| KNN | 0.82 | 0.80 | 0.79 | 0.79 |
| Decision Tree | 0.85 | 0.83 | 0.84 | 0.84 |
| Random Forest | 0.90 | 0.89 | 0.88 | 0.89 |