Cardiovascular disease predictor using machine learning

Objective

To predict the likelihood of cardiovascular disease using machine learning models based on patient health data. This project aims to assist in early diagnosis and effective treatment by identifying high-risk individuals.

Dataset

- **Source**: Provided CSV file.
- **Features**:
    - Demographic: Age, Gender
    - Physical: Height, Weight
    - Vital Signs: Systolic and Diastolic blood pressure (ap_hi, ap_lo)
    - Lifestyle: Smoking, Alcohol intake, Physical activity
    - Biochemical: Cholesterol, Glucose levels
- **Target**: cardio (0: No disease, 1: Disease)

Project Workflow

1. **Data Preprocessing**:
    - **Missing Value Handling**: Verified no missing values in the dataset.
    - **Data Cleaning**: Removed anomalies such as extreme outliers in height, weight, and blood pressure.
    - **Feature Scaling**: Standardized continuous variables for uniformity.
    - **Feature Encoding**: Encoded categorical variables for machine learning compatibility.
2. **Exploratory Data Analysis (EDA)**:
    - **Visualizations**:
        - **Histograms**: Distribution of age, cholesterol, and glucose levels.
        - **Box Plots**: Outlier analysis for height, weight, and blood pressure.
        - **Correlation Matrix**:

- Identified relationships among features.
- Blood pressure, cholesterol, and glucose had significant correlation with cardio.

- **Key Insights**:
   - Individuals with higher cholesterol levels are more likely to have cardiovascular disease.
   - Age and blood pressure are critical factors influencing heart health.

3. **Model Development**:
   - **Machine Learning Models Used**:
      - Logistic Regression (LR).
      - Support Vector Machine (SVM).
      - K-Nearest Neighbors (KNN).
      - Decision Trees (DT).
      - Random Forest (RF).
   - **Evaluation Metrics**:
      - Accuracy, Precision, Recall, F1-Score, Confusion Matrix.
   - **Results**:
      - Logistic Regression: 72% accuracy.
      - Random Forest: 75% accuracy (Best performing model).
      - KNN: 68% accuracy.
      - SVM and Decision Trees also performed moderately well.
4. **Model Comparison**:
   - Random Forest outperformed other models in both accuracy and stability.
   - Its ensemble nature provides better generalization for unseen data.
5. **Final Model Deployment**:
   - Selected Random Forest as the final model.
   - Saved the trained model using Pickle for deployment in healthcare applications.

Future Recommendations

1. **Data Enrichment**:
   - Include additional features like medical history, genetic factors, and dietary habits for better predictions.
2. **Model Improvements**:
   - Experiment with advanced models like Gradient Boosting, XGBoost, or deep learning techniques.
3. **Real-World Testing**:
   - Test the model on real-time patient data for validation and further refinement.
4. **Deployment**:
   - Integrate the model into a web or mobile application for easy accessibility by healthcare professionals.

Results

- **Accuracy of Final Model (Random Forest)**: **75%**
- **Key Findings**:
  - High cholesterol and glucose levels significantly increase cardiovascular risk.
  - Regular physical activity reduces the risk of heart disease.
  - Age and gender are notable predictors.

Conclusion

The project successfully developed a system capable of predicting cardiovascular disease with reasonable accuracy. The insights can help healthcare providers prioritize patients for early intervention and recommend lifestyle changes to mitigate risks.