PROJECT 2: Report on Analysis

I. Problem Statement and Description

Heart disease remains a significant global health concern, necessitating accurate prediction models for early detection and intervention. This project addresses the need for a reliable predictive model to assess an individual's risk of heart disease based on various medical attributes. By leveraging machine learning techniques, we aim to develop a tool that healthcare professionals can use to identify individuals at higher risk, enabling timely preventive measures and interventions.

Using a dataset comprising demographic information, medical history, and clinical measurements, we will employ supervised learning algorithms to build a predictive model for heart disease. The model's performance will be evaluated through rigorous testing, ensuring its accuracy and reliability in classifying individuals into high-risk and low-risk groups. Ultimately, this project strives to empower healthcare providers with a valuable tool for proactive heart disease management, ultimately leading to improved patient outcomes and reduced healthcare burden.

II. Brief Description of the Three Methods Used:

Linear Regression:

- Linear Regression is a linear modeling technique that assumes a linear relationship between the independent variables and the dependent variable.
- It estimates the parameters (coefficients) of a linear equation to best fit the observed data points, minimizing the sum of squared differences between predicted and actual values.
- Linear Regression is versatile and can handle both continuous and categorical independent variables, making it widely applicable in various fields such as economics, finance, and social sciences.

Logistic Regression:

- Logistic regression is a supervised learning algorithm used for binary classification. It models the probability that a given input belongs to a certain class using the logistic function.
- In multi-class classification problems like the Iris dataset, one-vs-all or one-vs-rest strategy can be employed where the logistic regression model is trained for each class against the others.

Decision Trees:

- Decision Trees are a non-linear classification method that recursively splits the data based on feature values to create a tree-like structure of decision rules.
- They partition the feature space into regions that are homogeneous with respect to the target variable.
- Decision Trees can handle both numerical and categorical data.

III. Experimental Results

Linear Regression

```
2]: from sklearn.linear_model import LinearRegression
    model = LinearRegression()
```

```
3]: model.fit(X_train, y_train)
```

```
3]: ▾ LinearRegression
    LinearRegression()
```

```
4]: model.score(X_test, y_test)
```

```
4]: 0.6151218749550655
```

```
5]: Lipredict = model.predict(X_test)
    Lipredict
```

```
5]: array([ 0.43745173,  0.52636761,  0.35230235, -0.037082  ,  0.0187627 ,
           -0.31586751,  1.02984678,  0.86795557,  0.63614529,  0.38200982,
            1.13657876,  1.11204565,  1.11322837,  0.79994219,  0.10102981,
            0.7202376 ,  0.7585017 ,  1.19622673, -0.14094057,  0.9331559 ,
            0.79031637,  0.20853058, -0.0821851 ,  0.23865322,  0.78060944,
            0.78605311,  0.62786287,  0.74544667, -0.11313589, -0.2034936 ,
            0.72326407,  0.32346871, -0.07627163, -0.19754573,  0.57506214,
            0.35939066,  0.68433538, -0.21906449,  0.57809215,  0.88043319,
            0.84114518,  0.64524237,  0.23256972,  0.22947682,  0.8250233 ,
            1.17276838,  0.00423519,  0.95483007,  0.80187234,  0.62215661,
            0.92880815,  1.01986567, -0.1977596 , -0.2002113 ,  0.32842826,
            0.72827237,  0.98224319, -0.00755103,  0.04336474,  0.56043273,
            0.08319025])
```

Logistic Regression

```
: from sklearn.linear_model import LogisticRegression
  model1 = LogisticRegression()
```

```
: model1.fit(X_train, y_train)
```

```
/Users/kavyakatumbaka/anaconda3/lib/python3.10/site-packages/sklearn/linear_model/_logistic.py:458: Conv
ing: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
```

```
: ▾ LogisticRegression
  LogisticRegression()
```

```
: model1.score(X_test, y_test)
```

```
: 0.8852459016393442
```

```
: Logpredict = model1.predict(X_test)
```

```
: Logpredict
```

```
: array([1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0,
         0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0,
         1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0])
```

Decision Trees

```
: from sklearn import tree
  model2 = tree.DecisionTreeClassifier()

: model2.fit(X_train, y_train)

:  ▾ DecisionTreeClassifier
   DecisionTreeClassifier()

: model2.score(X_test, y_test)

: 0.819672131147541

: dtpredict = model2.predict(X_test)

: print("Classification Report:\n", classification_report(y_test, dtpredict))

  Classification Report:
                precision    recall  f1-score   support

             0       0.76      0.89      0.82        28
             1       0.89      0.76      0.82        33

      accuracy                           0.82        61
     macro avg       0.83      0.83      0.82        61
  weighted avg       0.83      0.82      0.82        61


:
```

IV. Discussion of Results

- All three methods demonstrated effectiveness in predicting the likelihood of heart disease in individuals.
- Logistic regression emerged as the top performer among the methods, achieving an accuracy of 88.52%. Its relatively high accuracy suggests that it effectively captures the linear relationships between the medical attributes and the presence of heart disease.
- Decision trees closely followed logistic regression with an accuracy of 81.97%. Although slightly lower, decision trees still provided strong results, offering the advantage of interpretability due to their intuitive decision rules.
- Linear regression yielded an accuracy of 61.51%, which while lower than the other methods, still demonstrated predictive capability. However, its lower accuracy may indicate that the relationship between the independent variables and heart disease is not entirely linear, highlighting the importance of exploring more complex modeling techniques.
- The choice of method depends on various factors including the specific characteristics of the dataset, computational resources, and the interpretability of the model required for clinical decision-making. Logistic regression stands out for its high accuracy and straightforward interpretation, making it a suitable choice for practical applications in healthcare settings.