Heart Attack Risk Prediction

PRESENTED BY

SREERASYA KOLLA

Dataset

- From Kaggle
- CSV Format with 25 Features and a Target (Heart Attack Risk)
- Features are combination of Categorical and Numeric data
- Target is Categorical data (1/0)

Data Cleaning (Preprocessing)

- Data is clean, no missing values
- No cleaning needed

Feature Engineering (Preprocessing)

- Feature Extraction
  - New Features named Systolic pressure and Diastolic pressure were extracted from the original feature Blood Pressure
- Feature Selection
  - Selected features that exhibit strong relationship with target variable
  - Correlation coefficient and chi square test statistics
  - 13 Features out of 25

Normalization (Preprocessing)

- **Scaled numerical features and encoded categorical features**
- **Min Max Scaler and One Hot Encoder**

Model Selection

- **Classification problem**
- **Logistic Regression, Random Forest Classifier, Decision Tree Classifier**

Model Training and Evaluation

- Split data 70% for training, 30% for testing
- Trained three models
- Confusion Matrix, Precision, Recall, Accuracy , F1 score
- Logistic Regression < Random Forest Classifier < Decision Tree Classifier
- But not so good results
    - Accuracy 53.8%, Precision 36%, Recall 39%, F1 Score 37%

Class weighting & Hyperparameter Tuning

- Retrained decision tree classifier with best parameters
  - Grid search
  - Best params - gini, max depth -5 , min samples split - 10
  - Precision improved to 40%
  - Recall reduced again (3%)
- Found class imbalance in dataset
- Increased majority class(0) twice and minority class(1) thrice
- Retrained decision tree classifier with new dataset and best parameters

Results

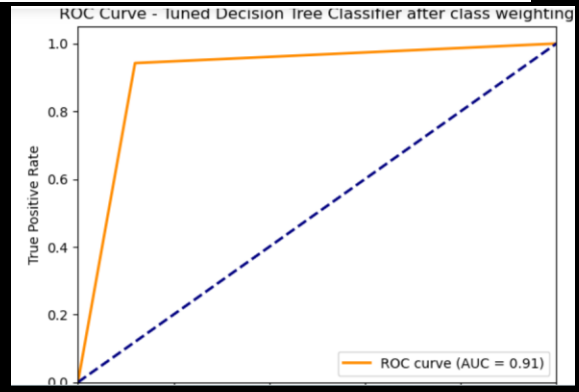
```
[[1689    2]
 [ 935    3]]
Evaluation Metrics for Logistic Regression:
Accuracy: 0.6435907189045265
Precision: 0.6
Recall: 0.0031982942430703624
F1 Score: 0.006362672322375397
```

```
[[1617   74]
 [ 902   36]]

Evaluation Metrics for Random Forest Classifier:
Accuracy: 0.6287561810574362
Precision: 0.32727272727272727
Recall: 0.03837953091684435
F1 Score: 0.06870229007633588
```

```
[[1047  644]
 [ 568  370]]

Evaluation Metrics for Decision Tree Classifier:
Accuracy: 0.5389882084442754
Precision: 0.36489151873767256
Recall: 0.39445628997867804
F1 Score: 0.3790983606557377
```

```
Best Hyperparameters:
{'criterion': 'gini', 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 10}
[[1652   39]
 [ 909   29]]

Evaluation Metrics for Tuned Decision Tree Classifier:
Accuracy: 0.6394066184861164
Precision: 0.4264705882352941
Recall: 0.03091684434968017
F1 Score: 0.05765407554671968
```

```
Confusion Matrix for Tuned Decision Tree Classifier after class weighting:
[[2983  404]
 [ 162 2651]]

Evaluation Metrics for Tuned Decision Tree Classifier after class weighting:
Accuracy: 0.9087096774193548
Precision: 0.8677577741407528
Recall: 0.9424102381798791
F1 Score: 0.9035446489434218
```



Conclusion

ABLE TO PREDICT HEART ATTACK RISK WITH ACCURACY, PRECISION AND RECALL OF 90.8%, 86.7%,94%.

Thank You!