HEART FAILURE EARLY DETECTION

C R DEEEPAK KUMAR

1833009

M.Sc DECISION & COMPUTING SCIENCES

1. IMPORT ALL LIBRARIES

* matplotlib.pyplot
* pandas
* seaborn
* imbalanced-learn
* mlxtend
* keras
* tensorflow
* vecstack

1. UPLOAD DATASET IN NOTEBOOK FOR ANALYSIS
   * Using pd.read\_csv
2. CHECK FOR MISSING VALUES
   * Using df.isnull().sum()
3. CHECK FORMAT OF ALL VARIABLES
   * df.info()
4. NUMBER OF OBSERVATIONS PRESENT IN TARGET VARIABLE
   * Imbalance present in target variable detected.
5. DEFINE INDEPENDENT AND DEPENDENT VARAIBLES
   * All columns chosen except target variable – Independent.
   * Target (‘DEATH\_EVENT’) variable – Dependent.
6. SCALE ALL INDEPENDENT VARIABLES
   * To have a similar measure so that ML model can make an accurate Prediction.
7. SPLIT TRAINING AND TESTING DATASET
   * Split from Independent and Dependent variables so that it can used for Training, Testing, and Evaluation.
8. DEFINE XGBCLASSIFIER
   * From xgboost import the XGBClassifier Algorithm.
9. IMPORT OVERSAMPLING FROM IMBALANCED LEARN
   * Oversample the observations present on Target variable to balance the observations.
10. DEFINE RANDOM SEARCH ALGORITHM
    * From the sklearn.model\_selection import RandomizedSearchCV
11. DEFINE RANDOM FOREST PARAMETERS
    * n\_estimators
    * max\_depth
    * min\_samples\_split
    * min\_samples\_leaf
12. DEFINE RANDOM FOREST CLASSIFIER ALGORITHM
    * Using Library sklearn.ensemble import the Random Forest Algorithm
13. USING RANDOM SEARCH FIND BEST COST-EFFECTIVE RANDOM FOREST ALGORITHM FITTING THE UNBALANCED DATASET
    * Best Algorithm shows accuracy score as 93 %.
    * Unfortunately, False Negative values quite high therefore cannot be used in Real world to save lives.
14. DEFINE PARAMETERS OF XGBCLASSIFIER
    * colsample\_bytree
    * n\_estimators
    * max\_depth
15. USE GRID SEARCH TO FIND BEST XGBCLASSIFIER ALGORITHM.
16. USE RANDOM SEARCH TO FIND BEST RANDOM FOREST ALGORITHM FITTING THE BALANCED DATASET.
17. USE MLXTEND CLASSIFIER TO STACK ALL THE THREE MODELS FITTED ON BALANCED SETS
    * Overfitting detected moving onto next stacking algorithm
18. DEFINE ADABOOST CLASSIFIER ALGORITHM
    * Using Library sklearn.ensemble import the AdaBoostClassifier Algorithm
19. DEFINE ADABOOST PARAMETERS
    * n\_estimators
20. DEFINE THE BEST ADABOOST ALGORITHM USING GRID SEARCH
21. USE VECSTACK TO STACK ALL THE MODELS FITTED ON BALANCED DATASET
    * Best Algorithm shows accuracy score as 90%.
    * Fortunately, this model shows an improvement since the value of False negative has decreased while the value of False Positive has increased.
    * Therefore, it is an expensive Life saver model.
22. USING UNDERSAMPLING METHOD USED TO CHECK WHETHER IT GIVES A BETTER SCORE THAN THE OVERSAMPLING.
23. THE SCORE WAS 83% WHICH IS SIGNIFICANTLY LOWER THEREFORE UNDERSAMPLING METHOD DROPPED.
24. IN CONCLUSION
    * VECSTACK ALGORITHM IS THE BEST DUE TO ITS LOW FALSE NEGATIVE VALUES.