The due date for submitting this assignment has passed.
Due on 2023-08-27, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

You have last submitted on: 2023-08-27, 12:58 IST

Notes:

Maximum marks : 80

This exam consists of a CLASSIFICATION problem.
The target is the 'Heart_Disease' column.
Random state should be taken as 64 wherever applicable.
For NAT type of question if nothing is mentioned, Enter the answer accurately upto 3 decimal places.
The dataset is already preprocessed, i.e. no missing values and numerical and categorical features are scaled/encoded accordingly.
Shape of the dataset is (12499, 19) [Note if you are not getting this shape that means your data has not been uploaded correctly to the colab]
Sklearn assumes class 1 as positive class. Therefore, precision, recall are calculated for class 1.
After every 5 questions, please click the submit button to save the answers.
Please make sure that you keep submitting intermittently to save your answers. We will be marking your score as 0 if your submission is not received by the Portal

Click here to view the sklearn library reference

1) Click here to view the dataset *0 points*

Which dataset are you using for this exam?

- V1
- V2 (selected)
- V3
- V4
- V5

Yes, the answer is correct.
Score: 0
Accepted Answers:
V2

2) Break the dataset into features(X) and label (y), where the column Heart_Disease goes to y and the rest of the columns go to X. What proportion of data points belongs to label 1?

0.587

Yes, the answer is correct.
Score: 3
Accepted Answers:
(Type: Range) 0.54, 0.62

3 points

3) Split the dataset into train and test dataset into the 80:20 ratio while keeping random_state =64. How many examples are there in training dataset?

9999

Yes, the answer is correct.
Score: 3
Accepted Answers:
(Type: Numeric) 9999

3 points

4) Take LogisticRegression estimator with following parameters for training:
Use lbfgs as solver
Set random state to be equal to 64
Tolerance for stopping criteria to be 1e-4
inverse of regularization parameter, C = 0.1
Maximum number of iterations taken for the solvers to converge to be 100
Enter the f1 score for the given model using test set(X_test, y_test)

0.865

Yes, the answer is correct.
Score: 4
Accepted Answers:
(Type: Range) 0.8 , 0.9

4 points

5) What is the intercept (bias term in the decision function) learnt by the above model?

-0.089

Yes, the answer is correct.
Score: 4
Accepted Answers:
(Type: Range) -0.11,-0.05

4 points

6) **(Common Instructions for Q6 and Q7)**

Take LogisticRegression(random_state = 64) estimator with GridSearchCV. Hyperparameter tuning to be done over the following parameters:

solver as 'lbfgs' or 'sag'

Maximum number of iterations taken for the solver to converge to be [100, 200, 500]

value of inverse reularization parameter C to be [0.01, 0.1, 1, 10]

Use cross validation = 4

Use the best model from above hyper parameter tuning process to answer following questions:

Enter the value of C of the best estimator you got after training with GridSearchCV.

10

Yes, the answer is correct.
Score: 4
Accepted Answers:
(Type: Numeric) 10

4 points

7) Enter the value of the precision on the test set using the best model:

0.855

Yes, the answer is correct.
Score: 4
Accepted Answers:
(Type: Range) 0.8, 0.9

4 points

8) Use SGDClassifier on the training dataset (X_train and y_train) to train the model. Use the following parameters:
log_loss is the loss function to be used
apply ridge regularization,
maximum number of iterations is 10
constant learning rate of 0.01,
regularization rate value is 0.001,
Take random_state=64.
Set warm_start as False
Note : Please ignore the convergence warning.

Using above model, calculate and write the correct value of f1_score for the test set.

0.856

Yes, the answer is correct.
Score: 6
Accepted Answers:
(Type: Range) 0.8, 0.9

6 points

9) Use Gridsearchcv with KNeighborsClassifier estimator, f1 as scoring parameter, cv= 5. Consider [5,7,9,11] as K values to be examined.
Consider following parameters for KNeighborsClassifier:

Use Euclidean distance metric
Keep other parameter values as default values.
What is the best value of K you obtained using the above instructions?

11

Yes, the answer is correct.
Score: 4
Accepted Answers:
(Type: Numeric) 11

4 points

10) **(Common Instructions for Q10 and Q11)**
Take DecisionTreeClassifier(random_state = 64) estimator with GridSearchCV. Hyperparameter tuning to be done over the following parameters:

Criterion as 'entropy' or 'gini'

Splitter as 'random' or 'best'

Minimum number of samples per leaf as [6,8,10, 12, 14]

Maximum depth as [5,6, 7, 8]

Minimum number of samples per node to split as [16, 18, 20, 22]

Use cross validation = 3

Use the best model from above hyper parameter tuning process to answer following questions:

What is the depth of the best estimator?

7

Yes, the answer is correct.
Score: 5
Accepted Answers:
(Type: Numeric) 7

5 points

11) Find the number of leaves in the best estimator.

8

No, the answer is incorrect.
Score: 0
Accepted Answers:
(Type: Numeric) 101

5 points

12) **(Common Instructions for Q12, Q13 and Q14)** *3 points*
Suppose we train a decision tree classifier using the following two approaches:

Approach 1. Do not grow the full size tree by setting the parameters as:

max depth of the tree to be 10.
minimum samples in a node to split to be 50.
Approach 2. Use cost complexity pruning by setting the cost complexity parameter as 0.01.

NOTE:

Use random state to be 64 in both approaches.
Use entropy criterion in both approaches.

Which approach gives the better accuarcy on the test dataset?

- Approach 1 (selected)
- Approach 2

Yes, the answer is correct.
Score: 3
Accepted Answers:
Approach 1

13) What is the difference (absolute difference) in True Positives (TP) between these two approaches on test datasets?

73

Yes, the answer is correct.
Score: 5
Accepted Answers:
(Type: Numeric) 73

5 points

14) Find the sum of number of nodes of both trees (trees obtained using approach 1 and appoach 2).

358

Yes, the answer is correct.
Score: 5
Accepted Answers:
(Type: Numeric) 358

5 points

15) **(Common Instructions for Q15, and Q16)**
Take a RandomForestClassifier() model with following hyperparameter values and tune it using GridsearchCV.
Use n_estimators as [10,20,30]
Use max_features as ['sqrt', 'log2']
Use min_impurity_decrease as [0.001, 0.01, 0.1]
Take cv value= 5 and random_state = 64

Calculate the total number of misclassified samples by the best estimator for the test data.

393

Yes, the answer is correct.
Score: 5
Accepted Answers:
(Type: Numeric) 393

5 points

16) Find the value of min_impurity_decrease of the best estimator?

0.001

Yes, the answer is correct.
Score: 5
Accepted Answers:
(Type: Numeric) 0.001

5 points

17) **(Common Instructions for Q17,Q18)**

Take an adaboost model with following hyperparameter values and tune it using GridsearchCV.
Use n_estimators as [20,30, 50]
Use learning_rate as [0.1, 0.5,1,2]
Take cv value= 5 and random state = 64

Find the accuracy on the test dataset using the best estimator.

0.8736

Yes, the answer is correct.
Score: 5
Accepted Answers:
(Type: Range) 0.83, 0.91

5 points

18) How many class zero examples in the test dataset are correctly classified by the best estimator?

837

Yes, the answer is correct.
Score: 5
Accepted Answers:
(Type: Numeric) 837

5 points

19) Apply GridSearchCV using the support vector machine (SVM) classifier on the training dataset X_train, y_train and calculate the best value of C and kernel from the values below. *5 points*
use below hyperparameter values to tune the model
`kernel':['linear', 'rbf'],
'C':[1, 10]`)
Take cv =3 and random state to be 64
Which of the following options represent the best parameters?

- {'C': 1, 'kernel': 'rbf'}
- {'C': 10, 'kernel': 'rbf'}
- {'C': 1, 'kernel': 'linear'} (selected)
- {'C': 10, 'kernel': 'linear'}

Yes, the answer is correct.
Score: 5
Accepted Answers:
{'C': 1, 'kernel': 'linear'}