Hope Artificial Intelligence

Classification Assignment

Problem Statement or Requirement:

A requirement from the Hospital, Management asked us to create a predictive model which will predict the Chronic Kidney Disease (CKD) based on the several parameters. The Client has provided the dataset of the same.

1.) Identify your problem statement
Many input variables are there all are elements of blood , based its value need to predict CKD.
Stage 1: input is number so ML
Stage 2: input and output both are given so supervised learning
Stage 3: output is a categorical value so classification

2.) Tell basic info about the dataset (Total number of rows, columns)

399 rows × 25 columns

```
Column names :
'age', 'bp', 'sg', 'al', 'su', 'rbc', 'pc', 'pcc', 'ba', 'bgr', 'bu',
     'sc', 'sod', 'pot', 'hrmo', 'pcv', 'wc', 'rc', 'htn', 'dm', 'cad',
     'appet', 'pe', 'ane', 'classification'
```

3.) Mention the pre-processing method if you're doing any (like converting string to number – nominal data)

Using one hot encoding , called get_dummies function changed categorical values to number

4.) Develop a good model with good evaluation metric. You can use any machine learning algorithm; you can create many models. Finally, you have to come up with final model.

Logistic regression

```
n [79]: print(cfm_logi)

        [[49  2]
         [ 0 82]]
```

```
n [80]: print(clrpt_logi)
```

```
                 precision    recall  f1-score   support

             0       1.00      0.96      0.98        51
             1       0.98      1.00      0.99        82

      accuracy                           0.98       133
     macro avg       0.99      0.98      0.98       133
  weighted avg       0.99      0.98      0.98       133
```

Random forest

```
print(cfm)
```

```
[[51  0]
 [ 1 81]]
```

```
print(clrpt)
```

```
                 precision    recall  f1-score   support

             0       0.98      1.00      0.99        51
             1       1.00      0.99      0.99        82

      accuracy                           0.99       133
     macro avg       0.99      0.99      0.99       133
  weighted avg       0.99      0.99      0.99       133
```

```
: from sklearn.metrics import roc_auc_score
  score_logi=roc_auc_score(y_test,logi_clasi.predict_proba(x_test)[:,1])
  score_Rf=roc_auc_score(y_test,clasi.predict_proba(x_test)[:,1])
  print(score_logi,'\n',score_Rf)
  #randon forest is having scroe 99, logitic regression is having score 98
```

```
0.9985652797704447
 0.9998804399808704
```

5.) All the research values of each algorithm should be documented. (You can make tabulation or screenshot of the results.)

Documented above , used gridsearchcv to get the best model

 6.) Mention your final model, justify why u have chosen the same. Note: Mentioned points are necessary, kindly mail your document as well as .ipynb (code file) with respective name. ϖ ϖ Sub file name also should be properly named for Example (SVM_Ramisha_Assi-5.ipynb) Communication is important (How you are representing the document.) Kindly uploaded in the Github and Share it with us

I have chosen random forest as my final model .

Accuracy also 99% , roc_auc_score also 99%