IML Lab

Report 5:

Question 1:

Link to Collab:https://colab.research.google.com/drive/1UjFNV3VEP-MrhBP2FV6dggF7XIRKyBNP#scrollTo=rY0eDU_N8VFU

—-->>>

In this question,

1.Imported dataset using Pandas

2.Removed all the null values using ffill method

3.Plotted Histogram,Scatter Plot of Culmen Length ,Culmen Depth ,Flipper length and Body Mass using`df.hist( bins=50, figsize=(10, 10))`

And `from pandas.plotting import scatter_matrix`
`scatter_matrix( df, figsize=(20,20))`

4.Showed pie chart presentation of nominal values such as Species ,Island and Sex using `df[attribute].value_counts().plot(ax=ax,kind='pie').`

5.I have encoded ordinal/ nominal and categorical features into their numerical equivalent using one hot encoder as the data in the columns are nominal which can only be categorized not ranked . So instead of ordinal encoder I have used One Hot Encoder.

6.Splitted Data into Ratio -70:20:10 and cross validate the training of the data using validation set-

Now I have written procedure for the following 2 approaches-

With parameters-

i.max_depth

ii. Min_samples_leaf

We have checked the training set and validation set accuracy.

In this case,using the maximum depth approach We got more accuracy than the Min_sample_ leaf approach.

- *Pruning trees works nice for decision trees because it removes noise, but doing this within RF kills bagging which relies on it for having uncorrelated members during voting. Max depth is usually only a technical parameter to avoid recursion overflows while min sample in leaf is mainly for smoothing votes for regression .*
- *Here when Max depth =10 we tend to see overfitting of Data.*
- *At max depth =5 and 4, they are the best fitting .*
- *Accuracy :*

  Max Depth =5 :0.9223300970873787
  Max Depth =10 :0.8996763754045307
  Max Depth=4:0.9061488673139159

THe BEST ACCURACY AT: Max depth =5 ANd min_sample_leaf=2

—---------------------------Thank You—----------------------------------