Smruti Dawale
B21BB007
Lab_05

Que1)

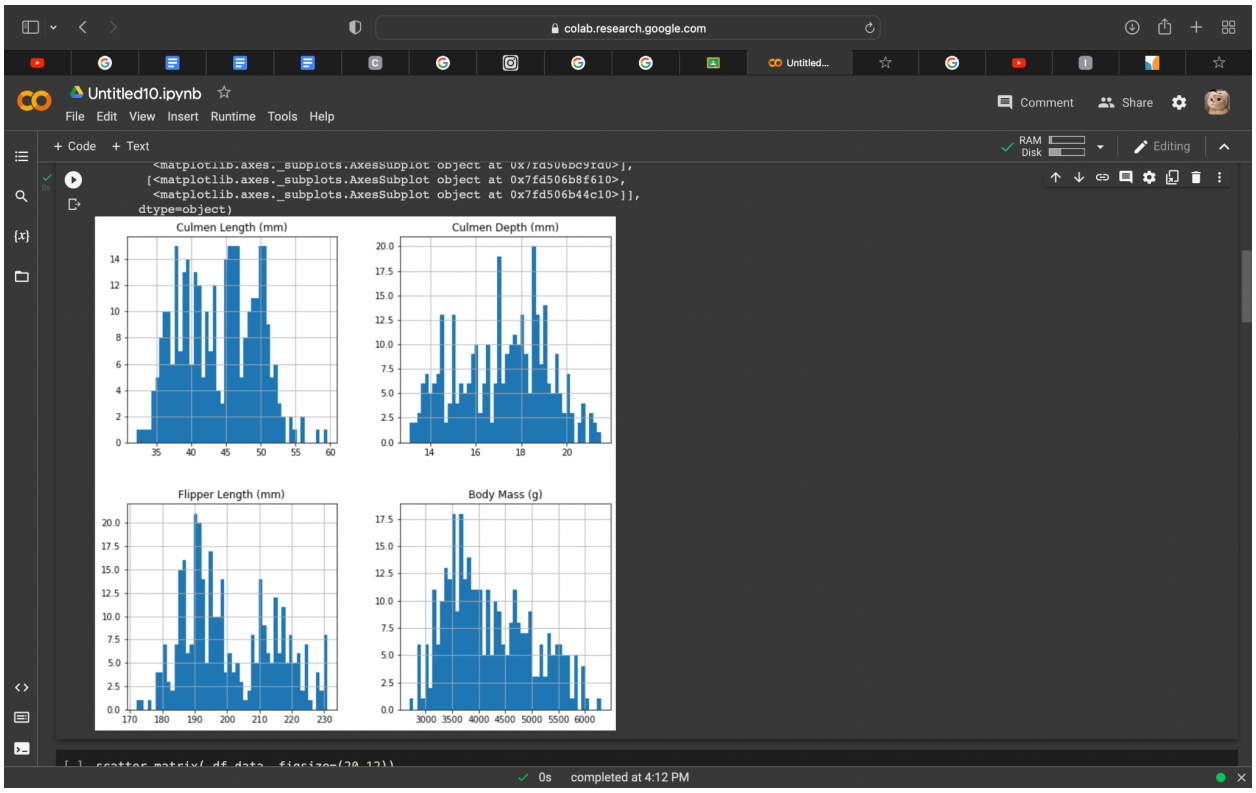
a)

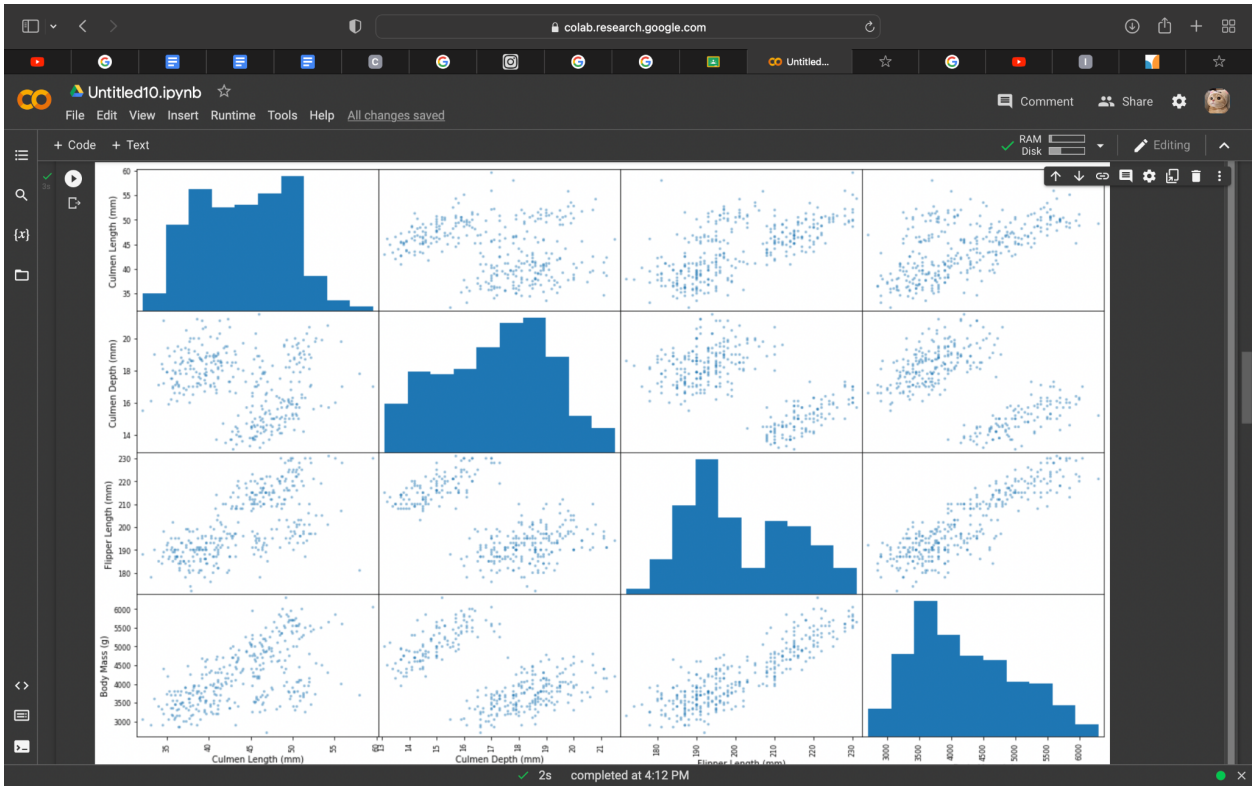
Almost all the algorithms in scikit learn are not capable of handling missing data , so before training your model you must handle your missing data.

One of the ways is “list wise deletion of cases” : discarding observations where value in any of the variables is missing .Here observations are your rows and variables are columns , you basically discard the row if any of the values in your column is missing.

I used dataDropNa = data.dropna(axis=0)

Visualization of data





b)

Categorical data can be divided into 2 types Nominal data (here you cannot define a relationship between the data) and Ordinal data (here you can define a relationship between the data).Categorical data is present mostly in the form of strings and machine learning algorithms expect numbers , so it's your duty to convert these categories to numbers .there are various type of encoding , one is Ordinal encoding which is used for the Ordinal type of data and OneHotEncoding is used for Nominal data .

.

c)

min_samples_leaf*int or float, default=1*

min_samples_leaf to ensure that multiple samples inform every decision in the tree, by controlling which splits will be considered. A very small number will usually mean the tree will overfit, whereas a large number will prevent the tree from learning the data. min_samples_split can create arbitrarily small leaves, min_samples_leaf guarantees that each leaf has a minimum size, avoiding low-variance, over-fit leaf nodes in regression

problems. For classification with few classes, min_samples_leaf=1 is often the best choice.

Note that min_samples_split considers samples directly and independent of sample_weight, if provided (e.g. a node with m weighted samples is still treated as having exactly m samples). Consider min_weight_fraction_leaf or min_impurity_decrease if accounting for sample weights is required at splits.

max_depth
The depth of a tree is the maximum distance between the root and any leaf.
Max_depth has no effect on accuracy of data.
On Varying the values of max_depth and min_samples_split i observed changing max_depth doesn't change the accuracy.

d) `max_depth=8, min_samples_leaf= 2`
`Accuracy_score = 1.0`

Decision trees tend to overfit on data with a large number of features. Getting the right ratio of samples to the number of features is important, since a tree with few samples in high dimensional space is very likely to overfit.

Decision Tree

colab.research.google.com

Untitled10.ipynb

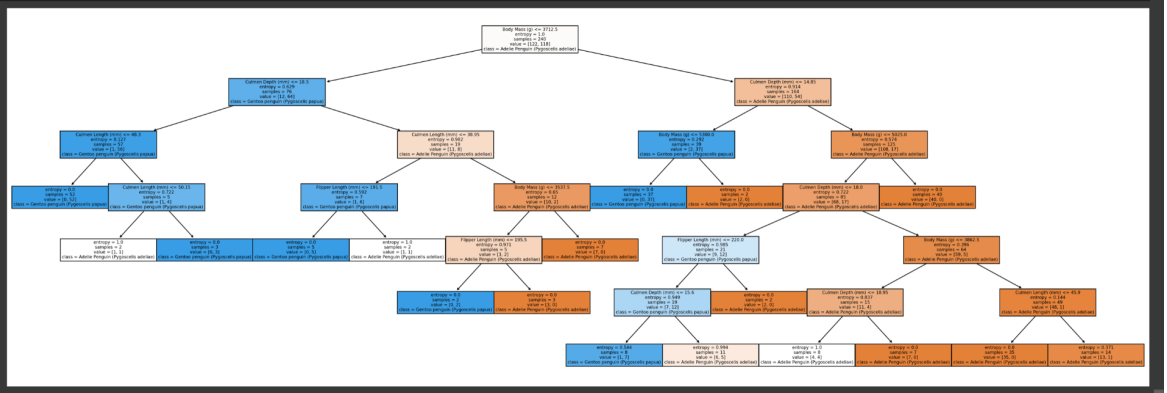
File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk Editing

```
fig, axes = plt.subplots(nrows = 1,ncols = 1,figsize = (30,10), dpi=300)
tree.plot_tree(tree_clf,
               feature_names = fn,
               class_names=cn,
               filled = True);
fig.savefig('penguin_tree.png')
```



2s completed at 4:12 PM