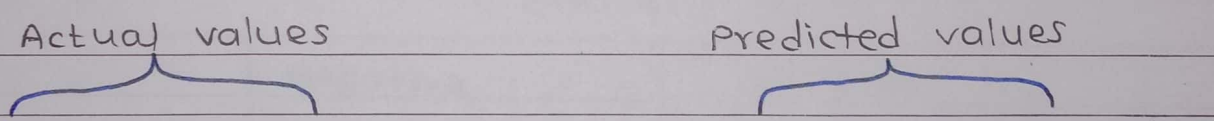
Assignment No. 5.2

Aim:- Data analytics 2 compute confusion matrix to find TP, FP, TN, FN, accuracy, error rate, precision recall on the given set.

Theory :- When we get data, after data cleaning, pre-processing & wrangling, the first step we do is to feed it to an outstanding model & get output in probabilities. But how can we measure the effectiveness of our model. Better the effectiveness, better the performance & that is exactly what we want. And it is where the confusion matrix comes into the limelight. Confusion matrix is a performance measurement for machine learning classification.

what is confusion matrix & why you need it?
Precision, specificity, accuracy & most importantly AUC-ROC curves



Actual vs predicted values

How to calculate confusion matrix for a 2-class classification problem?

Y	Y pred	output for threshold 0.6	Recall	Precision	Accuracy
0	0.5	0			
1	0.9	1			
0	0.7	1	1/2	2/3	4/7
1	0.7	1			
1	0.3	0			
0	0.4	0			
0	0.5	0			
0					

confusion matrix

$$\text{Recall} = \frac{TP}{TP + FN}$$

precision:-

$$\text{precision} = \frac{TP}{TP + FP}$$

Accurarcy:-

From the classes how many of them predicted correctly.

$$\text{F-measure} = \frac{2 * \text{Recall} * \text{precision}}{\text{Recall} + \text{precision}}.$$

what is confusion matrix:-

It is a matrix of size 2x2 for binary classificatn with actual values on one axis & predicted on another.

		Actual: Negative	Actual: positive
prediction	Negative	True Negative	false negative
	positive	false positive	True positive

confusion matrix.

The confusing terms in the confusion matrix are: true positive, true negative, false negative & false positive with an example.

Example:-

	Negative	positive
Negative	60	8
positive	22	10

confusion matrix for tumor detectn.

True positive:- model correctly predicts the positive class.

True Negative (TN):- correctly predicts negative class.

false positive (FP):- gives wrong predictn ~~of~~ negative class

False Negative (FN):- wrongly predicts positive classes.

$$TPR = \frac{TP}{TP + FN}$$

$$FNR = \frac{FN}{TP + FN}$$

$$TNR = \frac{TN}{TN + FP}$$

$$FPR = \frac{FP}{TN + FP}$$

Example 1:- credit card & fraud detectn

		Actual: Fair Transactn	Actual: Fraud Transactn
predicted:	fair	TN	FN
	fraud	FP	TP

Example 2: spam detection.

	Not spam	spam
Not spam	TN	FN
spam.	FP	TP

$$F1 \text{ score} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{Recall}}} = \frac{2*(\text{precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

$$F_\beta = (1+\beta^2) * \frac{(\text{precision} * \text{Recall})}{(\beta^2 * \text{precision}) + \text{Recall}}$$

Beta represents how many times recall is more imp than precision.

confusion matrix - An overview with python & R.

Introductn:-

To develop ML classificatn model, we collect data, explore, pre-process & clean it. after that we apply classificatn techniques.

confusion matrix: Definition :-

It is used to judge the performance of classifier on test dataset. confusion matrix is also termed as error matrix. it contains counts of correct & incorrect values.

Terminologies :-

1) TP 2) TN 3) FP 4) FN

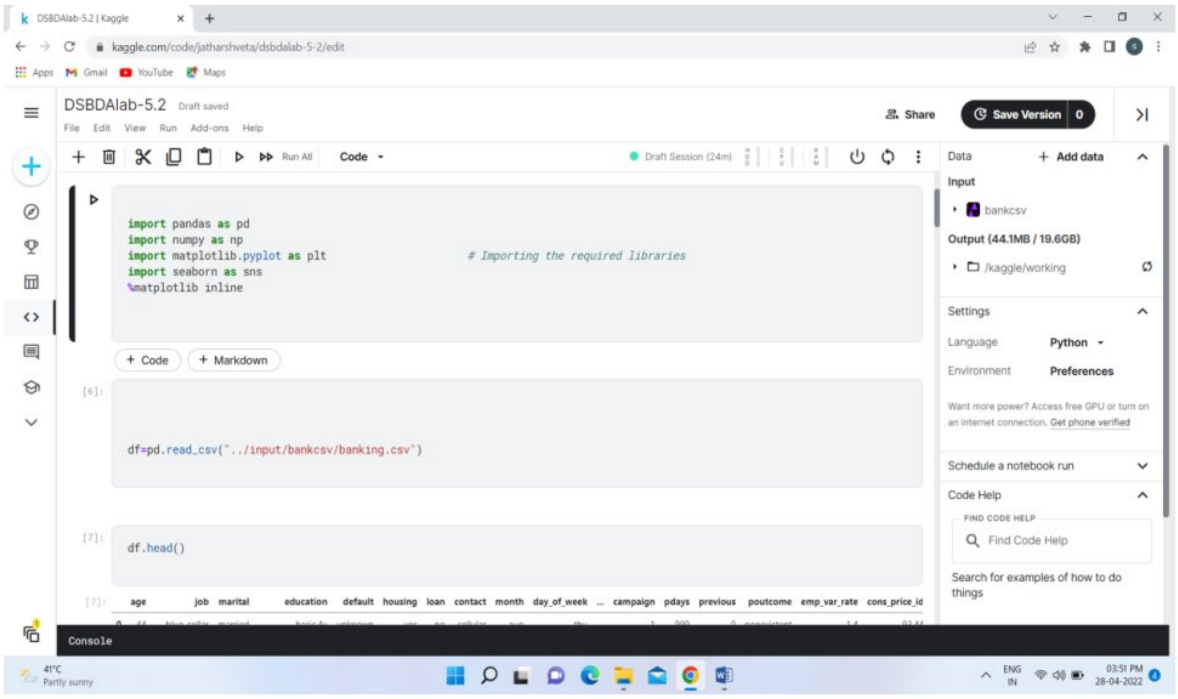
TP - both predicted & actual are positive

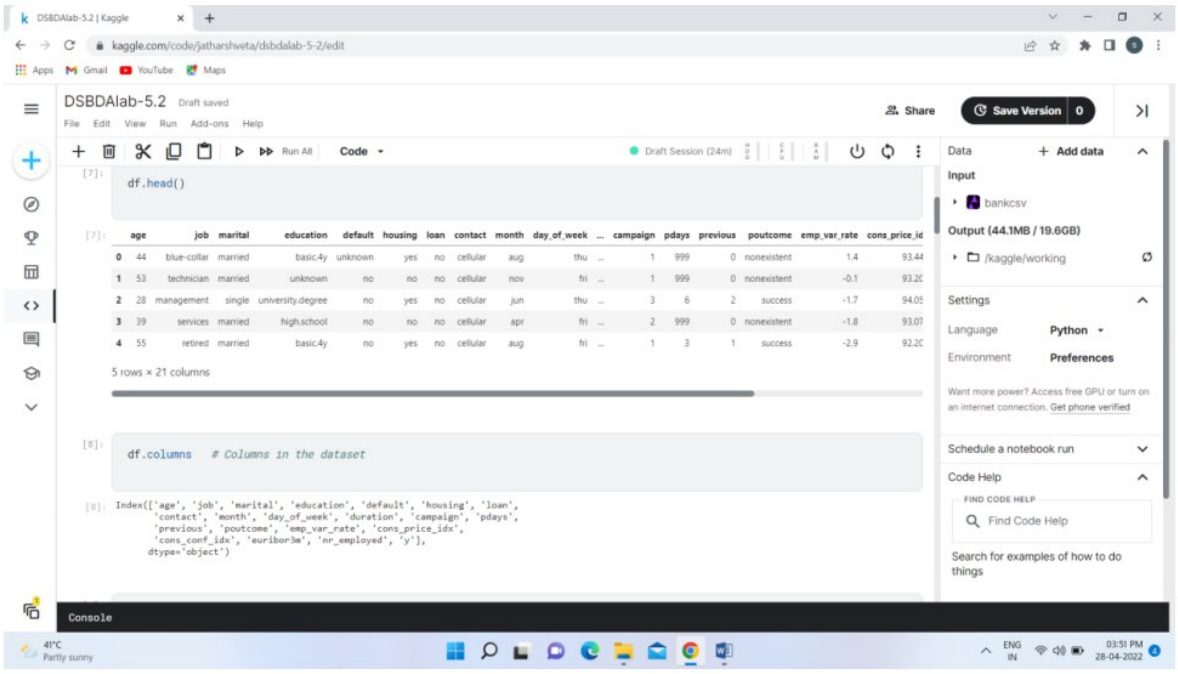
TN - both actual & predicted are negative.

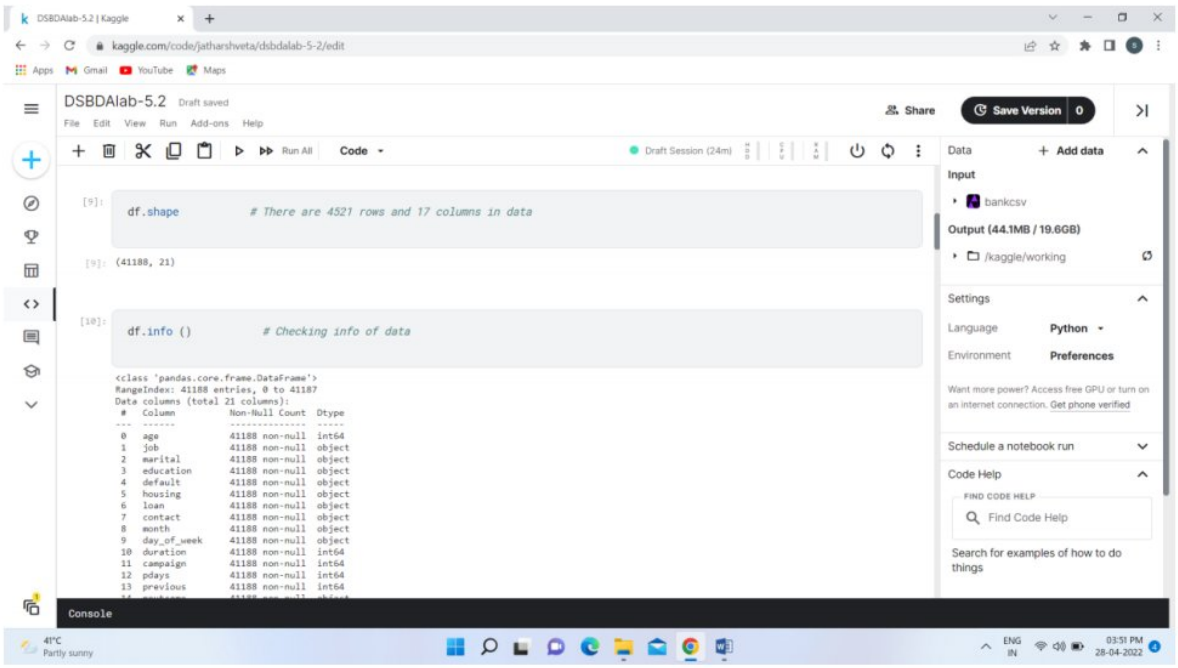
FP - actual value is negative but predicted +ve.

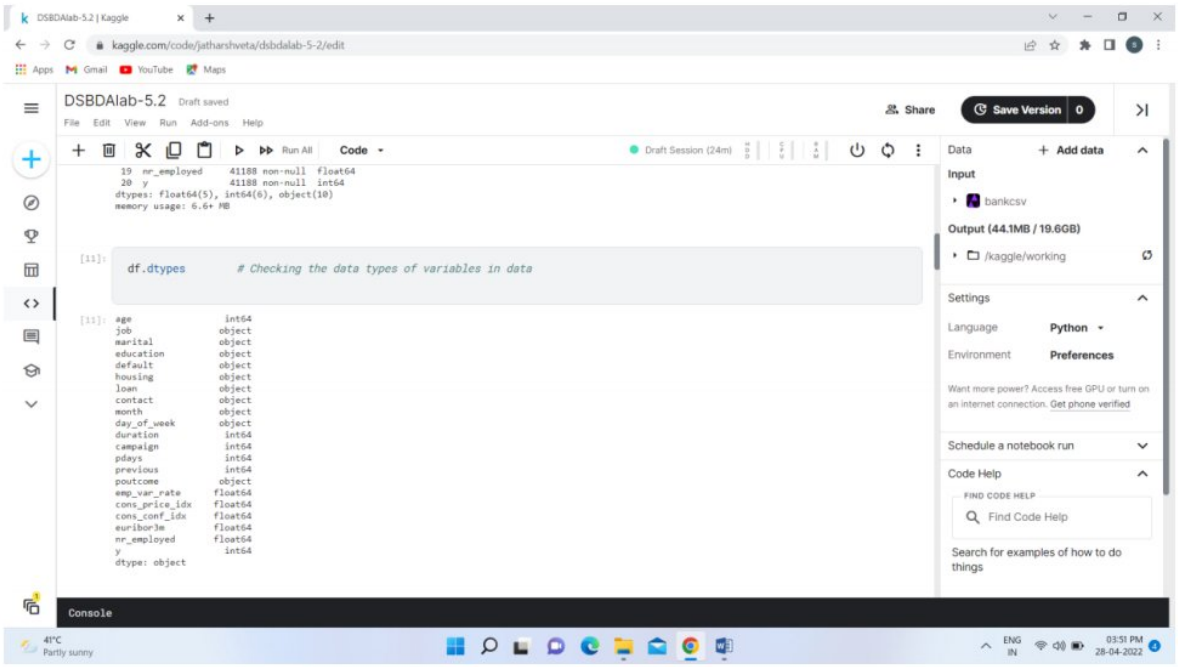
FN - actual value is positive but predicted -ve.

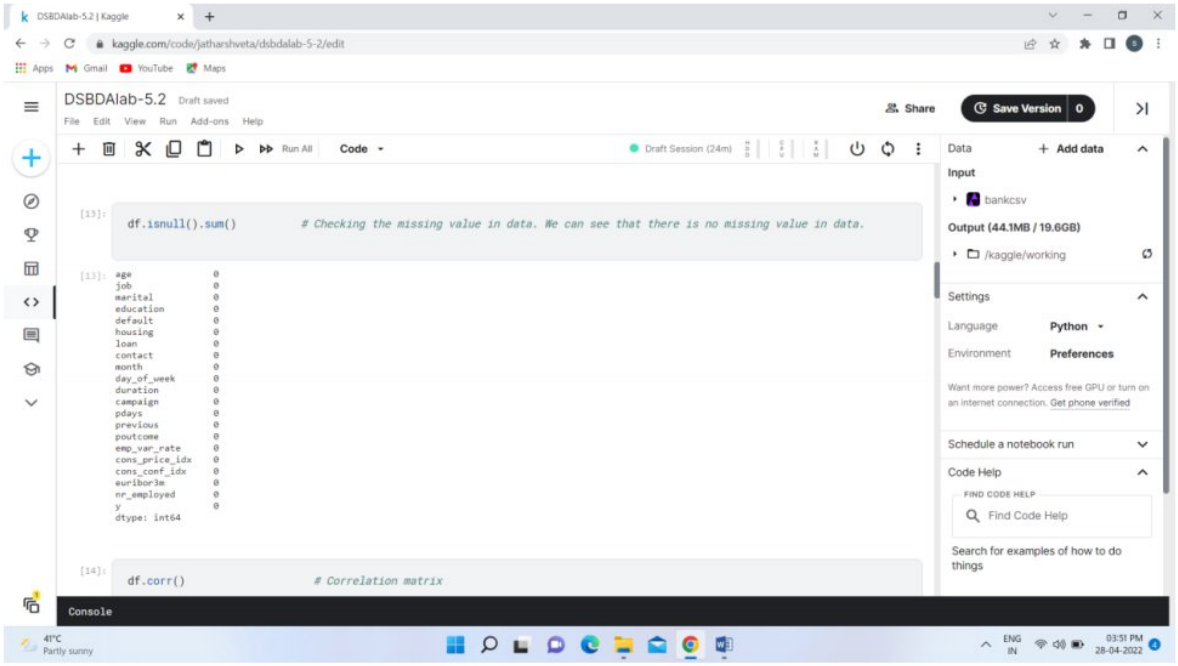
Conclusion :- confusion matrix, precision, recall & F1 score provides better insights into the predictn as compared to accuracy performance metrics. Applicatn of precision, recall & F1 score is informatn retrieval. & many more.

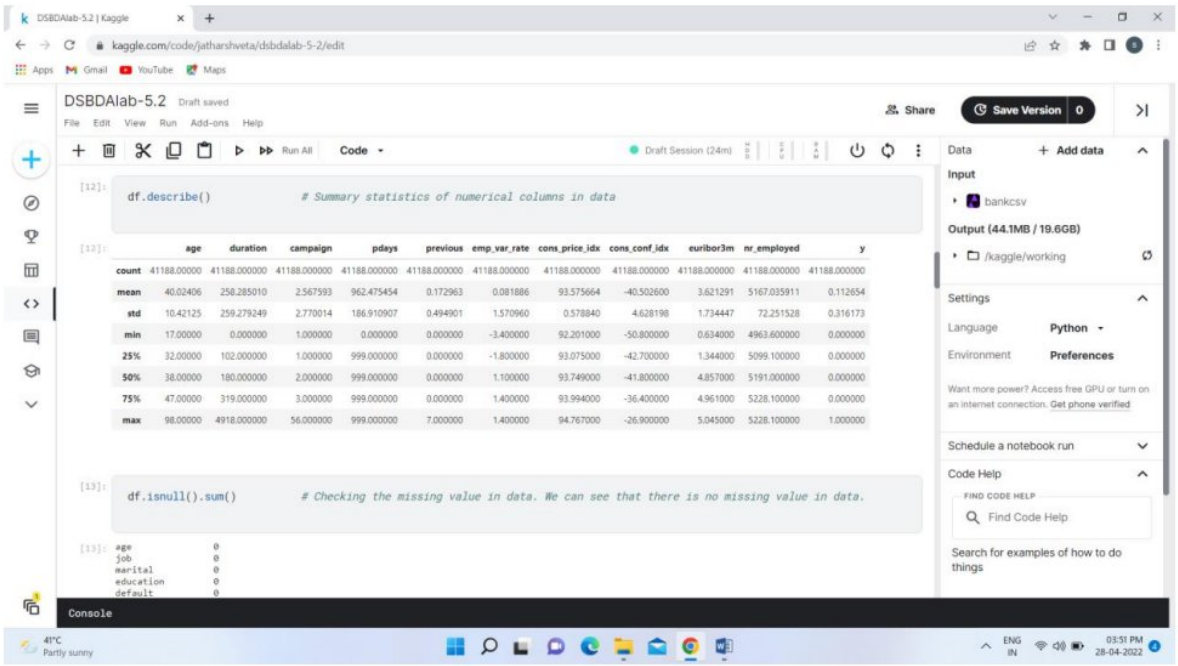


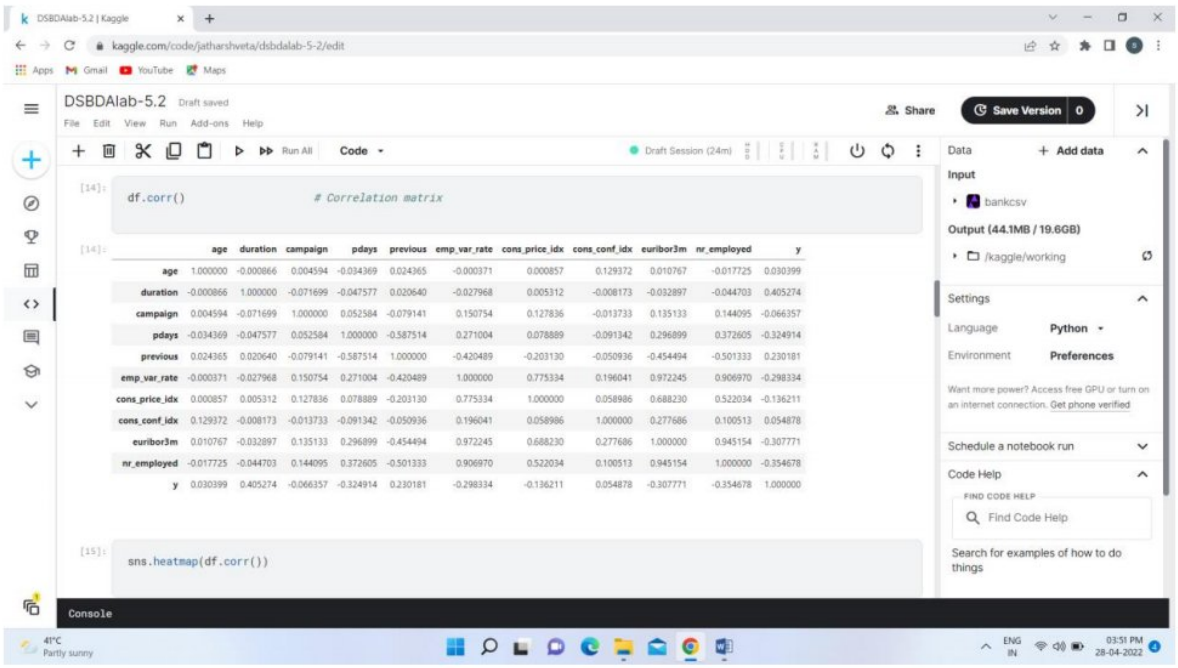


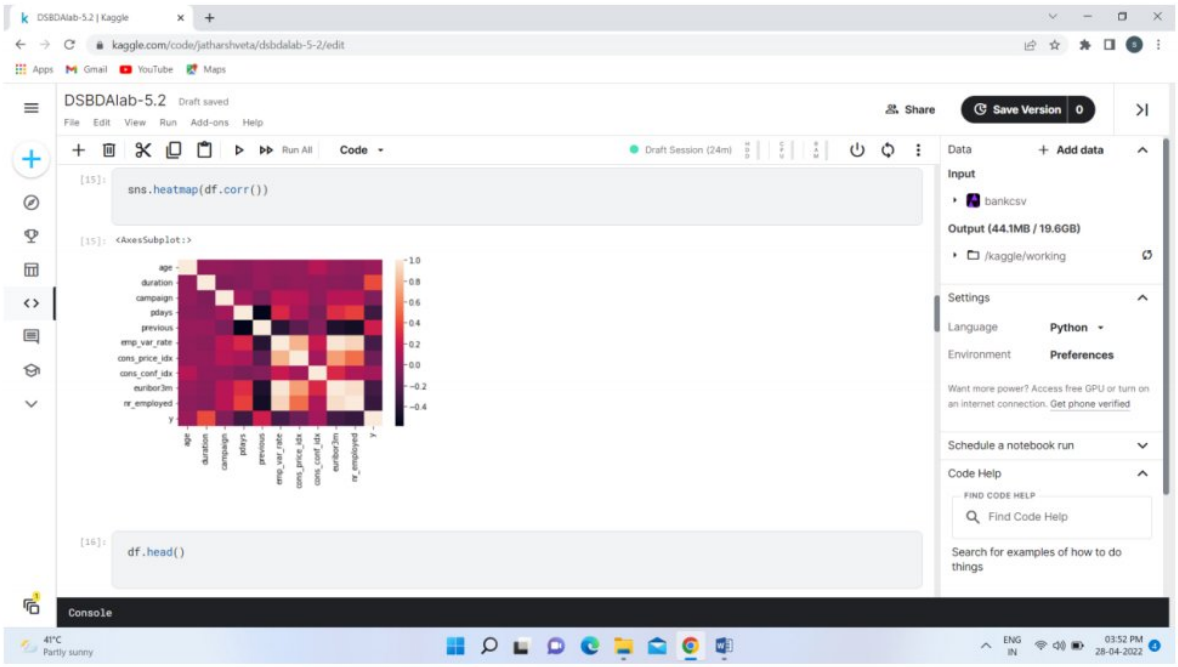


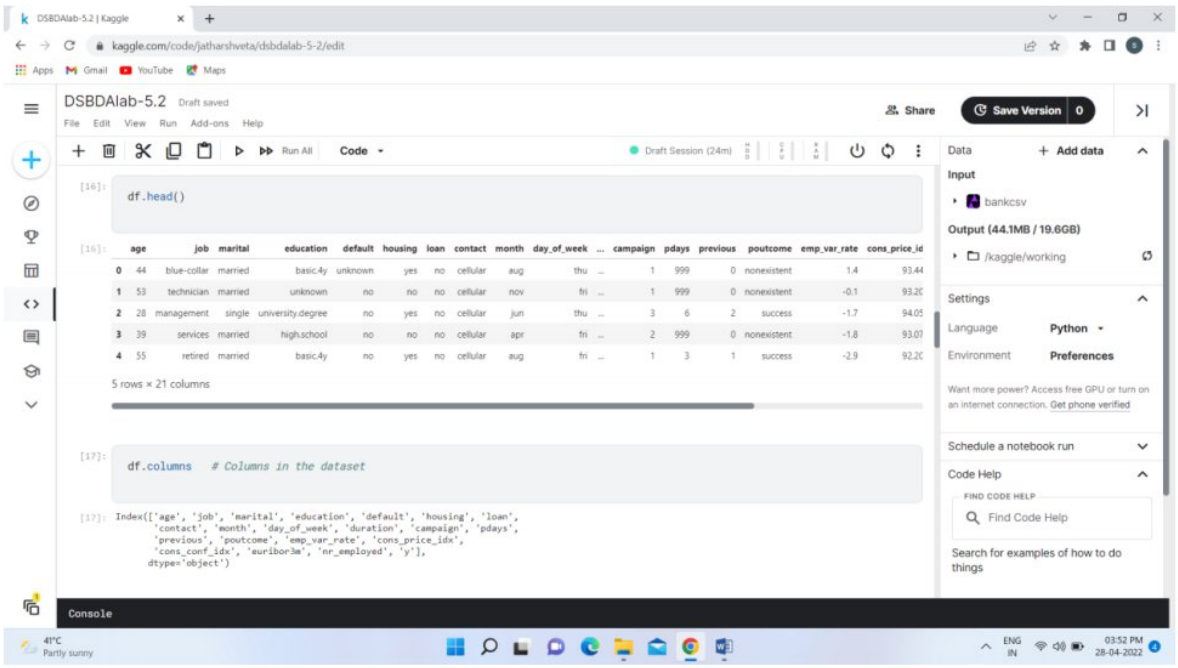


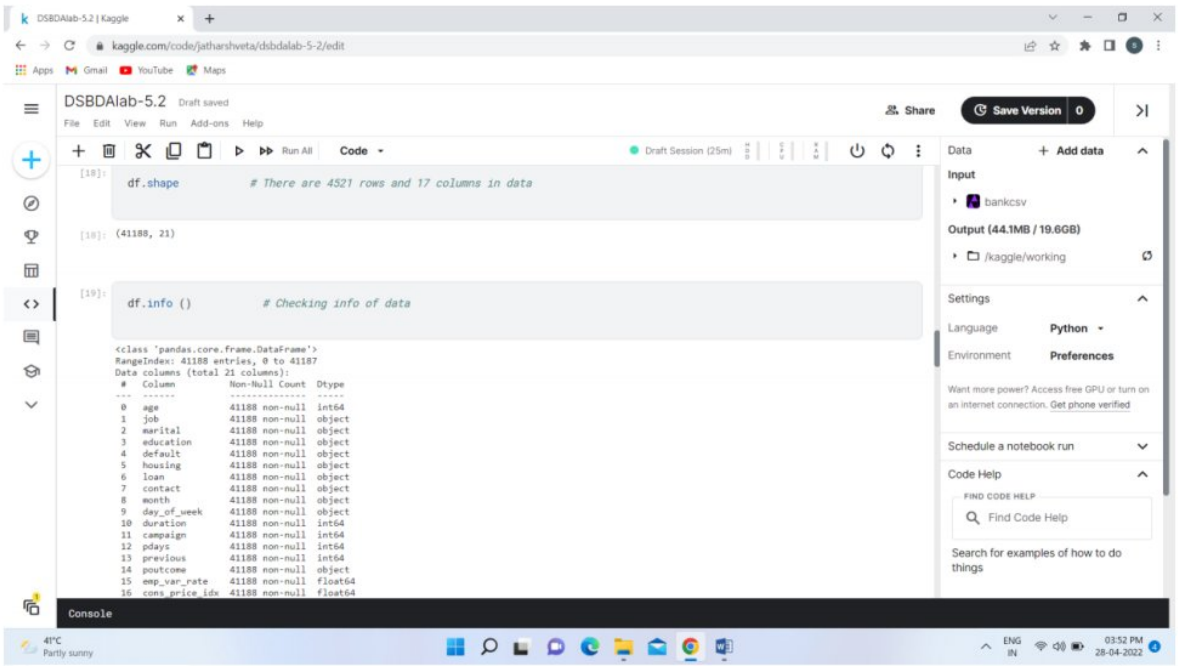


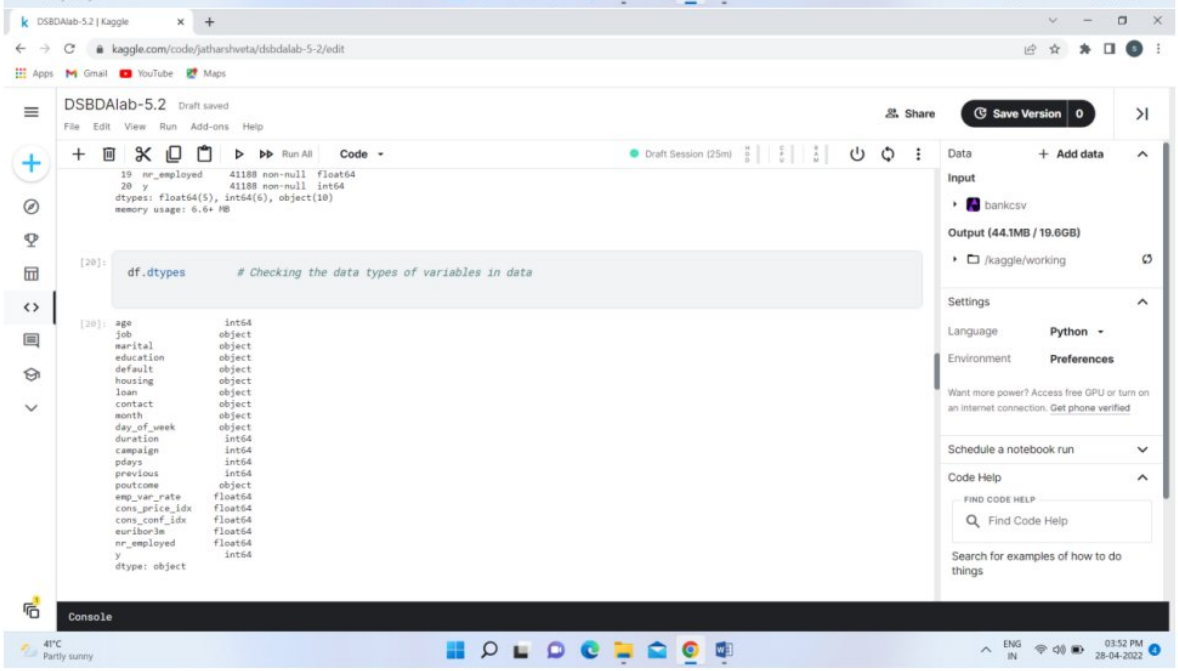


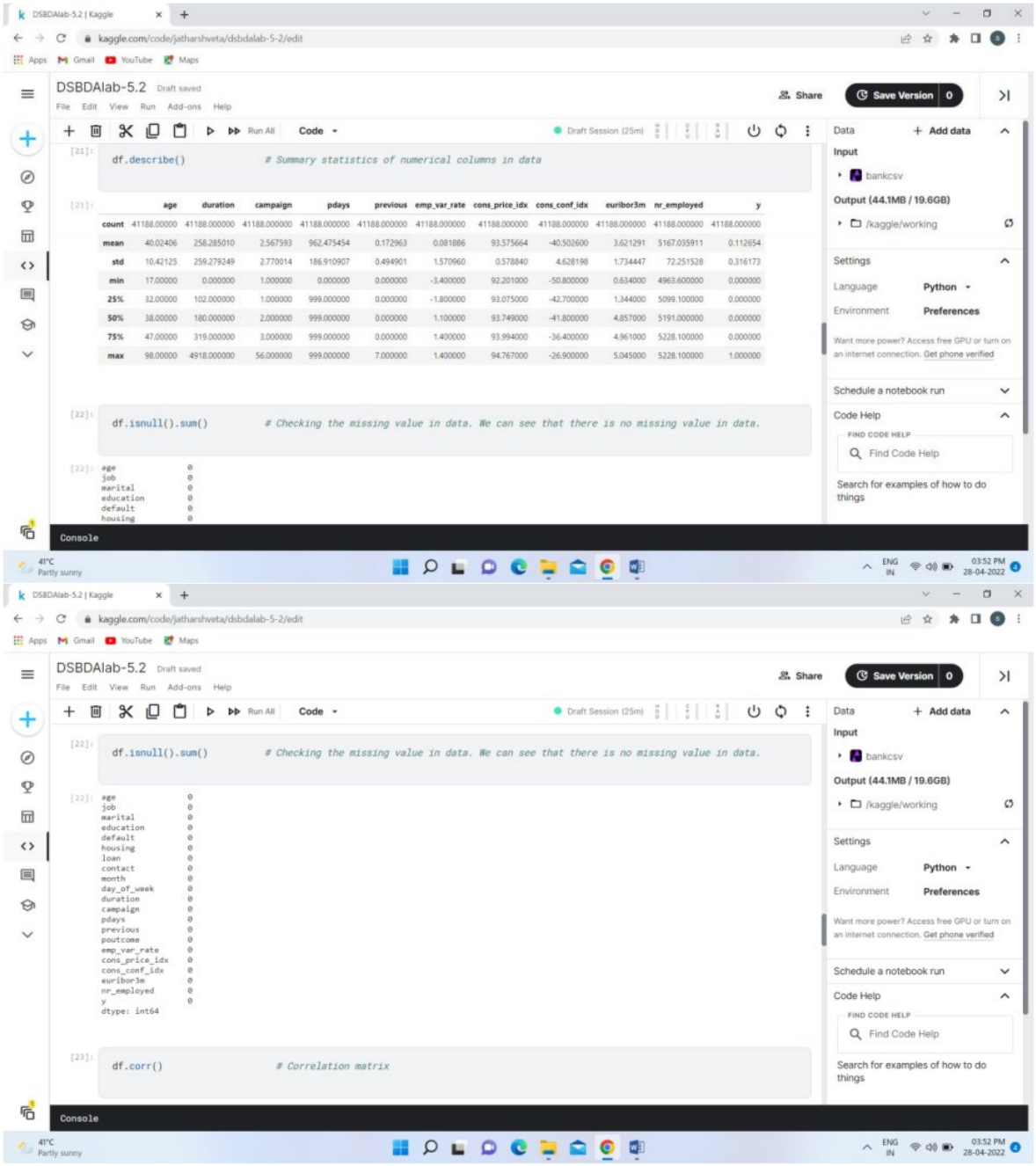












DSBDAlab-5.2 | Kaggle

kaggle.com/code/jatharshveta/dsbdalab-5-2/edit

Apps Gmail YouTube Maps

DSBDAlab-5.2 Draft saved

File Edit View Run Add-ons Help

Share Save Version 0

Run All Code Draft Session (25m)

[23]:
```
df.corr()                    # Correlation matrix
```

[23]:

	age	duration	campaign	pdays	previous	emp_var_rate	cons_price_idx	cons_conf_idx	euribor3m	nr_employed	y
age	1.000000	-0.000866	0.004594	-0.034369	0.024365	-0.000371	0.000857	0.129372	0.010767	-0.017725	0.030399
duration	-0.000866	1.000000	-0.071699	-0.047577	0.020640	-0.027968	0.005312	-0.008173	-0.032897	-0.044703	0.405274
campaign	0.004594	-0.071699	1.000000	0.052584	-0.079141	0.150754	0.127836	-0.013733	0.135133	0.144095	-0.066357
pdays	-0.034369	-0.047577	0.052584	1.000000	-0.587514	0.271004	0.078889	-0.091342	0.296899	0.372605	-0.324914
previous	0.024365	0.020640	-0.079141	-0.587514	1.000000	-0.420489	-0.203130	-0.050936	-0.454494	-0.501333	0.230181
emp_var_rate	-0.000371	-0.027968	0.150754	0.271004	-0.420489	1.000000	0.775334	0.196041	0.972245	0.906970	-0.298334
cons_price_idx	0.000857	0.005312	0.127836	0.078889	-0.203130	0.775334	1.000000	0.058986	0.688230	0.522034	-0.136211
cons_conf_idx	0.129372	-0.008173	-0.013733	-0.091342	-0.050936	0.196041	0.058986	1.000000	0.277686	0.100513	0.054878
euribor3m	0.010767	-0.032897	0.135133	0.296899	-0.454494	0.972245	0.688230	0.277686	1.000000	0.945154	-0.307771
nr_employed	-0.017725	-0.044703	0.144095	0.372605	-0.501333	0.906970	0.522034	0.100513	0.945154	1.000000	-0.354678
y	0.030399	0.405274	-0.066357	-0.324914	0.230181	-0.298334	-0.136211	0.054878	-0.307771	-0.354678	1.000000

[24]:
```
sns.heatmap(df.corr())
```

Data + Add data

Input

bankcsv

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

Language Python

Environment Preferences

Want more power? Access free GPU or turn on an internet connection. Get phone verified

Schedule a notebook run

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

Console

41°C Partly sunny

ENG IN 03:52 PM 28-04-2022

DSBDAlab-5.2 | Kaggle

kaggle.com/code/jatharshveta/dsbdalab-5-2/edit

Apps Gmail YouTube Maps

DSBDAlab-5.2 Draft saved

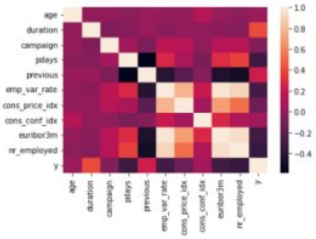
File Edit View Run Add-ons Help

Share Save Version 0

Run All Code Draft Session (25m)

[24]:
```
sns.heatmap(df.corr())
```

[24]: <AxesSubplot:>



[25]:
```
sns.countplot(y='job', data= df)
sns.countplot(x='marital', data= df)
sns.countplot(x='y', data= df)
```

Data + Add data

Input

bankcsv

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

Language Python

Environment Preferences

Want more power? Access free GPU or turn on an internet connection. Get phone verified

Schedule a notebook run

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

Console

41°C Partly sunny

ENG IN 03:52 PM 28-04-2022

DSBDAlab-5.2 | Kaggle

kaggle.com/code/jatharshveta/dsbdalab-5-2/edit

Apps Gmail YouTube Maps

DSBDAlab-5.2 Draft saved

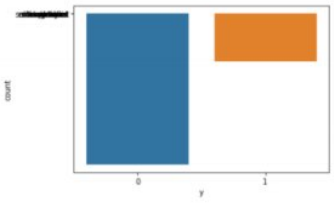
File Edit View Run Add-ons Help

Share Save Version 0

Run All Code Draft Session (25m)

[25]:
```
sns.countplot(y='job', data= df)
sns.countplot(x='marital', data= df)
sns.countplot(x='y', data= df)
```

[25]:
```
<AxesSubplot:xlabel='y', ylabel='count'>
```



[26]:
```
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder
```

Console

Data + Add data

Input

bankcsv

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

Language Python

Environment Preferences

Want more power? Access free GPU or turn on an internet connection. Get phone verified

Schedule a notebook run

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

41°C Partly sunny

ENG IN 03:53 PM 28-04-2022

DSBDAlab-5.2 | Kaggle

kaggle.com/code/jatharshveta/dsbdalab-5-2/edit

Apps Gmail YouTube Maps

DSBDAlab-5.2 Draft saved

File Edit View Run Add-ons Help

Share Save Version 0

Run All Code Draft Session (25m)

[26]:
```
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder
from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
from sklearn.metrics import accuracy_score,confusion_matrix
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
```

[36]:
```
le = preprocessing.LabelEncoder()
df.job = le.fit_transform(df.job)
df.job
```

[36]:
```
0        1
1        9
2        4
3        7
4        5
        ..
41183    5
41184    3
41185    0
41186    9
41187    8
```

Console

Data + Add data

Input

bankcsv

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

Language Python

Environment Preferences

Want more power? Access free GPU or turn on an internet connection. Get phone verified

Schedule a notebook run

Code Help

FIND CODE HELP

Find Code Help

Search for examples of how to do things

41°C Partly sunny

ENG IN 03:53 PM 28-04-2022