

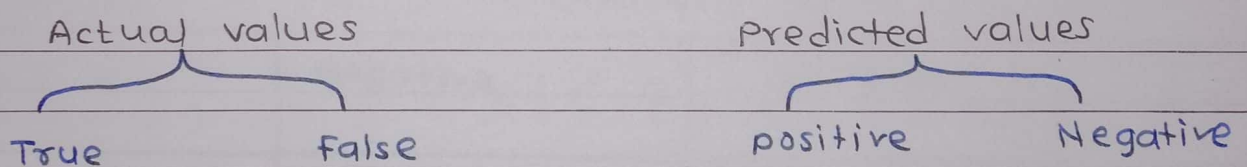
## Assignment No. 5.2

**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	1/2	2/3	4/7
1	0.9	1			
0	0.7	1			
1	0.7	1			
1	0.3	0			
0	0.4	0			
0	0.5	0			

FOR EDUCATIONAL USE

confusion matrix

FOR EDUCATIONAL USE

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

precision:-

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

Accuracy:-

From the classes how many of them predicted correctly.

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

what is confusion matrix:-

It is a matrix of size  $2 \times 2$  for binary classification with actual values on one axis & predicted on another.

		Actual	
		Negative	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	Fraud Transactn
predicted	fair	TN	FN
	fraud	FP	TP

Example 2: spam detection.

	Not spam	Spam
Not spam	TN	FN
spam	FP	TP

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

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

$\beta$  represents how many times recall is more imp than precision.

confusion matrix - An overview with python & R.

Introducn:-

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 <sup>contains</sup> 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 Draft saved

File Edit View Run Add-ons Help

+ Code - Draft Session (24m)

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

# Importing the required libraries

+ Code + Markdown

[6]:

```
df=pd.read_csv("../input/bankcsv/banking.csv")
```

[7]:

```
df.head()
```

[7]:

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	campaign	pdays	previous	poutcome	emp_var_rate	cons_price_id
--	-----	-----	---------	-----------	---------	---------	------	---------	-------	-------------	-----	----------	-------	----------	----------	--------------	---------------

Console

41°C Partly sunny

ENG IN 03:51 PM 28-04-2022

DSBDalab-5.2 Draft saved

File Edit View Run Add-ons Help

+ Code - Draft Session (24m)

[7]:

```
df.head()
```

[7]:

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	campaign	pdays	previous	poutcome	emp_var_rate	cons_price_id
0	44	blue-collar	married	basic4y	unknown	yes	no	cellular	aug	thu	...	1	999	0	nonexistent	1.4	93.44
1	53	technician	married	unknown	no	no	no	cellular	nov	fri	...	1	999	0	nonexistent	-0.1	93.20
2	28	management	single	university.degree	no	yes	no	cellular	jun	thu	...	3	6	2	success	-1.7	94.05
3	39	services	married	highschool	no	no	no	cellular	apr	fri	...	2	999	0	nonexistent	-1.8	93.07
4	55	retired	married	basic4y	no	yes	no	cellular	aug	fri	...	1	3	1	success	-2.9	92.20

5 rows x 21 columns

[8]:

```
df.columns # Columns in the dataset
```

[8]:

```
Index(['age', 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day_of_week', 'duration', 'campaign', 'pdays', 'previous', 'poutcome', 'emp_var_rate', 'cons_price_id', 'cons_conf_id', 'euribor3m', 'nr_employed', 'y'], dtype=object)
```

Console

41°C Partly sunny

ENG IN 03:51 PM 28-04-2022

DSBDAlab-5.2 | Kaggle

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

DSBDAlab-5.2 Draft saved

File Edit View Run Add-ons Help

+ Save Version 0

Run All Code

Draft Session (24m)

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

```
[9]: df.shape # There are 4521 rows and 17 columns in data

[9]: (41188, 21)

[10]: df.info() # Checking info of data

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column             Non-Null Count  Dtype  
---  --
 0   age                41188 non-null  int64  
 1   job                41188 non-null  object  
 2   marital            41188 non-null  object  
 3   education          41188 non-null  object  
 4   default            41188 non-null  object  
 5   housing            41188 non-null  object  
 6   loan               41188 non-null  object  
 7   contact            41188 non-null  object  
 8   month             41188 non-null  object  
 9   day_of_week        41188 non-null  object  
10   duration           41188 non-null  int64  
11   campaign           41188 non-null  int64  
12   pdays             41188 non-null  int64  
13   previous           41188 non-null  int64  
14   postcode           41188 non-null  int64  
15   emp_var_rate       41188 non-null  float64 
16   cons_price_idx     41188 non-null  float64 
17   cons_conf_idx      41188 non-null  float64 
18   euribor3m         41188 non-null  float64 
19   nr_employed        41188 non-null  float64 
20   y                 41188 non-null  int64  
dtypes: float64(5), int64(6), object(10)
memory usage: 6.6+ MB
```

Console

41°C Partly sunny

ENG IN 03:51 PM 28-04-2022

DSBDAlab-5.2 | Kaggle

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

DSBDAlab-5.2 Draft saved

File Edit View Run Add-ons Help

+ Save Version 0

Run All Code

Draft Session (24m)

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

```
[11]: df.dtypes # Checking the data types of variables in data

[11]: age                int64
job                object
marital            object
education          object
default            object
housing            object
loan               object
contact            object
month             object
day_of_week        object
duration           int64
campaign           int64
pdays            int64
previous           int64
postcode           object
emp_var_rate       float64
cons_price_idx     float64
cons_conf_idx      float64
euribor3m         float64
nr_employed        float64
y                 int64
dtype: object
```

Console

41°C Partly sunny

ENG IN 03:51 PM 28-04-2022

DSBDAlab-5.2 | Kaggle

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

DSBDAlab-5.2 Draft saved

File Edit View Run Add-ons Help

Run All Code

Draft Session (24m)

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

```
[13]: df.isnull().sum() # Checking the missing value in data. We can see that there is no missing value in data.
```

```
[13]: age      0
      job      0
      marital  0
      education  0
      default  0
      housing  0
      loan     0
      contact  0
      month    0
      day_of_week  0
      duration  0
      campaign  0
      pdays    0
      previous  0
      poutcome  0
      emp_var_rate  0
      cons_price_idx  0
      cons_conf_idx  0
      euribor3m  0
      nr_employed  0
      y         0
      dtype: int64
```

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

Console

4°C Partly sunny

ENG IN 03:51 PM 28-04-2022

DSBDAlab-5.2 | Kaggle

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

DSBDAlab-5.2 Draft saved

File Edit View Run Add-ons Help

Run All Code

Draft Session (24m)

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

```
[12]: df.describe() # Summary statistics of numerical columns in data
```

```
[12]:
```

	age	duration	campaign	pdays	previous	emp_var_rate	cons_price_idx	cons_conf_idx	euribor3m	nr_employed	y
count	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963	0.081886	93.575664	-40.502600	3.621291	5167.035911	0.112654
std	10.42125	259.279249	2.770014	166.910907	0.494901	1.570960	0.578840	4.628198	1.734447	72.251528	0.316173
min	17.000000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000	-50.800000	0.634000	4963.600000	0.000000
25%	32.000000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000	-42.700000	1.344000	5096.100000	0.000000
50%	38.000000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000	-41.800000	4.857000	5191.000000	0.000000
75%	47.000000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000	-36.400000	4.961000	5228.100000	0.000000
max	98.000000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000	-26.900000	5.045000	5228.100000	1.000000

```
[13]: df.isnull().sum() # Checking the missing value in data. We can see that there is no missing value in data.
```

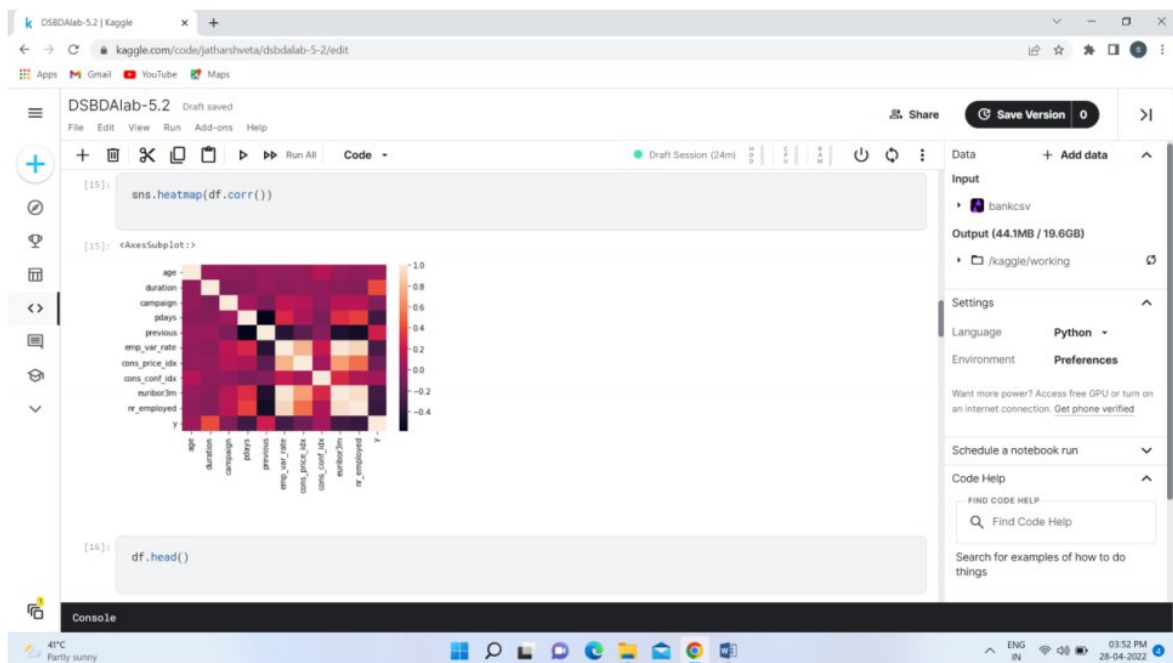
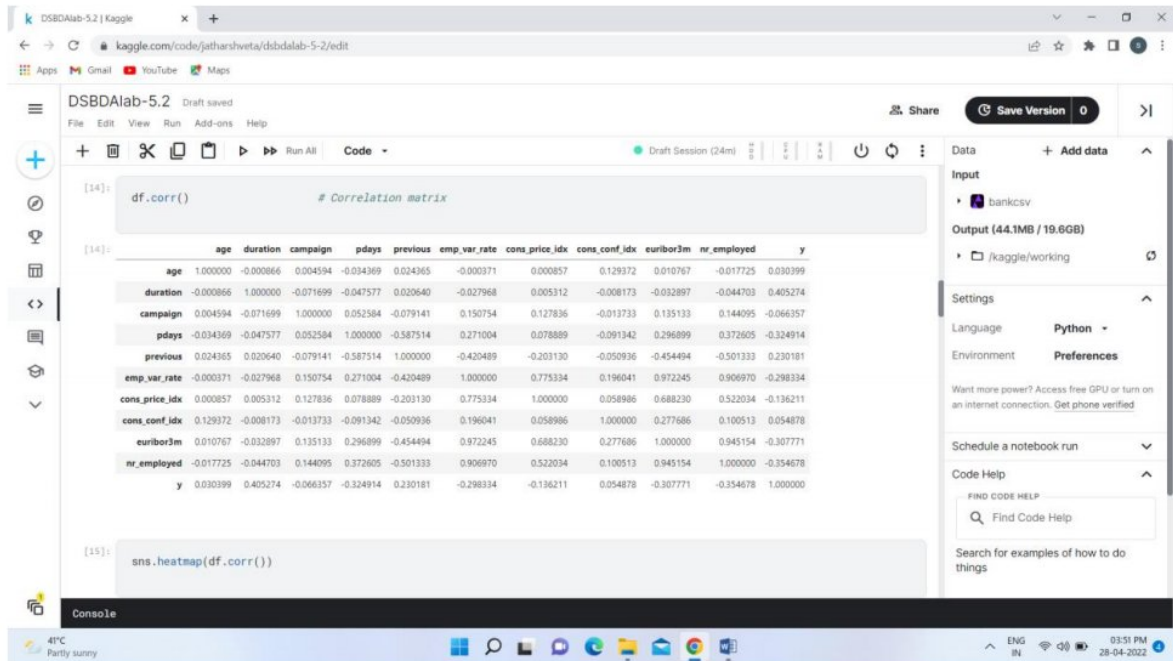
```
[13]: age      0
      job      0
      marital  0
      education  0
      default  0
```

Console

4°C Partly sunny

ENG IN 03:51 PM 28-04-2022





DSBDAlab-5.2 | Kaggle

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

AppsGmailYouTubeMaps

DSBDAlab-5.2Draft saved

FileEditViewRunAdd-onsHelp

ShareSave Version0

+[-]Run AllCode

Draft Session (24m)

[-]RunAllCode

[16]:

df.head()

[16]:

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	...	campaign	pdays	previous	poutcome	emp_var_rate	cons_price_id
0	44	blue-collar	married	basic4y	unknown	yes	no	cellular	aug	thu	...	1	999	0	nonevistent	1.4	93.44
1	53	technician	married	unknown	no	no	no	cellular	nov	fri	...	1	999	0	nonevistent	-0.1	93.20
2	28	management	single	university.degree	no	yes	no	cellular	jun	thu	...	3	6	2	success	-1.7	94.05
3	39	services	married	highschool	no	no	no	cellular	apr	fri	...	2	999	0	nonevistent	-1.8	93.07
4	55	retired	married	basic4y	no	yes	no	cellular	aug	fri	...	1	3	1	success	-2.9	92.20

5 rows × 21 columns

[17]:

df.columns # Columns in the dataset

[17]:

Index(['age', 'job', 'marital', 'education', 'default', 'housing', 'loan', 'contact', 'month', 'day\_of\_week', 'duration', 'campaign', 'pdays', 'previous', 'poutcome', 'emp\_var\_rate', 'cons\_price\_id', 'cons\_conf\_idx', 'euribor3m', 'nr\_employed', 'y'], dtype='object')

Input

bankcsv

Output (44.1MB / 19.6GB)

/kaggle/working

Settings

LanguagePython

EnvironmentPreferences

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

4°CPartly sunny

ENGIN

83:52 PM28-04-2022

DSBDalab-5.2 | Kaggle

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

DSBDalab-5.2 Draft saved

File Edit View Run Add-ons Help

Run All Code

Draft Session (25m)

Data + Add data

Save Version 0

Input

banks.csv

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

```
[18]: df.shape # There are 4521 rows and 17 columns in data
```

```
[18]: (41188, 21)
```

```
[19]: df.info() # Checking info of data
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41188 entries, 0 to 41187
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype  
---  -
 0   age                   41188 non-null  int64  
 1   job                   41188 non-null  object  
 2   marital               41188 non-null  object  
 3   education             41188 non-null  object  
 4   default               41188 non-null  object  
 5   housing               41188 non-null  object  
 6   loan                  41188 non-null  object  
 7   contact               41188 non-null  object  
 8   month                 41188 non-null  object  
 9   day_of_week           41188 non-null  object  
10   duration              41188 non-null  int64  
11   campaign              41188 non-null  int64  
12   pdays                 41188 non-null  int64  
13   previous              41188 non-null  int64  
14   poutcome              41188 non-null  object  
15   emp_var_rate          41188 non-null  float64 
16   cons_price_idx        41188 non-null  float64 
17   cons_conf_idx         41188 non-null  float64 
18   euribor3m             41188 non-null  float64 
19   nr_employed            41188 non-null  float64 
20   y                      41188 non-null  int64  
dtypes: float64(5), int64(6), object(10)
memory usage: 6.6+ MB
```

```
[20]: df.dtypes # Checking the data types of variables in data
```

```
age                int64
job                object
marital            object
education          object
default            object
housing            object
loan               object
contact            object
month              object
day_of_week        object
duration           int64
campaign           int64
pdays             int64
previous           int64
poutcome           object
emp_var_rate       float64
cons_price_idx     float64
cons_conf_idx      float64
euribor3m          float64
nr_employed        float64
y                  int64
dtype: object
```

Console

41°C Partly sunny

DSBDalab-5.2 | Kaggle

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

DSBDalab-5.2 Draft saved

File Edit View Run Add-ons Help

Run All Code

Draft Session (25m)

Data + Add data

Save Version 0

Input

banks.csv

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



Scanned with CamScanner

41°C Partly sunny 03:52 PM 28-04-2022

DSBDAlab-5.2 | Kaggle

File Edit View Run Add-ons Help

Code

```
[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)

Console

Share Save Version 0

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

DSBDAlab-5.2 | Kaggle

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

DSBDAlab-5.2 Draft saved

File Edit View Run Add-ons Help

+ Run All Code

Draft Session (25m)

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

```
[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

DSBDAlab-5.2 | Kaggle

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

DSBDAlab-5.2 Draft saved

File Edit View Run Add-ons Help

+ Run All Code

Draft Session (25m)

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

```
[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



DSBDAlab-5.2 | Kaggle

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

DSBDAlab-5.2 Draft saved

File Edit View Run Add-ons Help

+ Draft Session (28m)

[38]:

```
df.marital = le.fit_transform(df.marital)
df.marital
```

[38]:

```
0    1
1    1
2    2
3    1
4    1
..
41183 1
41184 1
41185 2
41186 1
41187 2
Name: marital, Length: 41188, dtype: int64
```

[39]:

```
df.default = le.fit_transform(df.default)
df.default
```

[39]:

```
0    1
1    0
2    0
3    0
4    0
..
41183 1
41184 1
41185 1
```

Console

41°C Partly sunny

ENG IN 03:53 PM 28-04-2022

DSBDAlab-5.2 | Kaggle

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

DSBDAlab-5.2 Draft saved

File Edit View Run Add-ons Help

+ Draft Session (28m)

[40]:

```
df.education = le.fit_transform(df.education)
df.education
```

[40]:

```
0    0
1    7
2    6
3    3
4    0
..
41183 3
41184 0
41185 6
41186 5
41187 3
Name: education, Length: 41188, dtype: int64
```

[41]:

```
df.housing = le.fit_transform(df.housing)
df.housing
```

[41]:

```
0    2
1    0
2    2
3    0
4    2
..
41183 0
41184 0
41185 2
```

Console

41°C Partly sunny

ENG IN 03:53 PM 28-04-2022

DSBDAlab-5.2 | Kaggle

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

DSBDAlab-5.2 Draft saved

File Edit View Run Add-ons Help

+ Draft Session (28m)

Code

```
[44]: df.loan = le.fit_transform(df.loan)
df.loan
```

```
[44]: 0 0
1 0
2 0
3 0
4 0
..
41183 2
41184 0
41185 2
41186 2
41187 0
Name: loan, Length: 41188, dtype: int64
```

```
[45]: df.contact = le.fit_transform(df.contact)
df.contact
```

```
[45]: 0 0
1 0
2 0
3 0
4 0
..
41183 1
41184 1
41185 1
41186 1
41187 1
```

Console

41°C Partly sunny

DSBDAlab-5.2 | Kaggle

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

DSBDAlab-5.2 Draft saved

File Edit View Run Add-ons Help

+ Draft Session (28m)

Code

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

DSBDAlab-5.2 | Kaggle

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

DSBDAlab-5.2 Draft saved

File Edit View Run Add-ons Help

+ Draft Session (28m)

Code

```
[47]: df.month = le.fit_transform(df.month)
df.month
```

```
[47]: 0 1
1 7
2 4
3 0
4 1
..
41183 4
41184 6
41185 6
41186 8
41187 6
Name: month, Length: 41188, dtype: int64
```

```
[48]: df.poutcome = le.fit_transform(df.poutcome)
df.poutcome
```

```
[48]: 0 1
1 1
2 2
3 1
4 2
..
41183 1
41184 1
41185 1
41186 1
41187 1
```

Console

41°C Partly sunny

DSBDAlab-5.2 | Kaggle

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

DSBDAlab-5.2 Draft saved

File Edit View Run Add-ons Help

+ Draft Session (28m)

Code

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

Scanned with CamScanner