DATA SCIENCE

PRESENTED BY M.RISHETHA

ROLL NO:21781A0590

SRI VENKATESWARA COLLEGE OF ENGINEERING & TECHNOLOGY(AUTONOMOUS)

SVCET

సాధనం సర్వత్ర సాధకం

CHITTOOR

INTRODUCTION

- What is a Data?

  Data is nothing but collection of facts.

- What is Data Science?

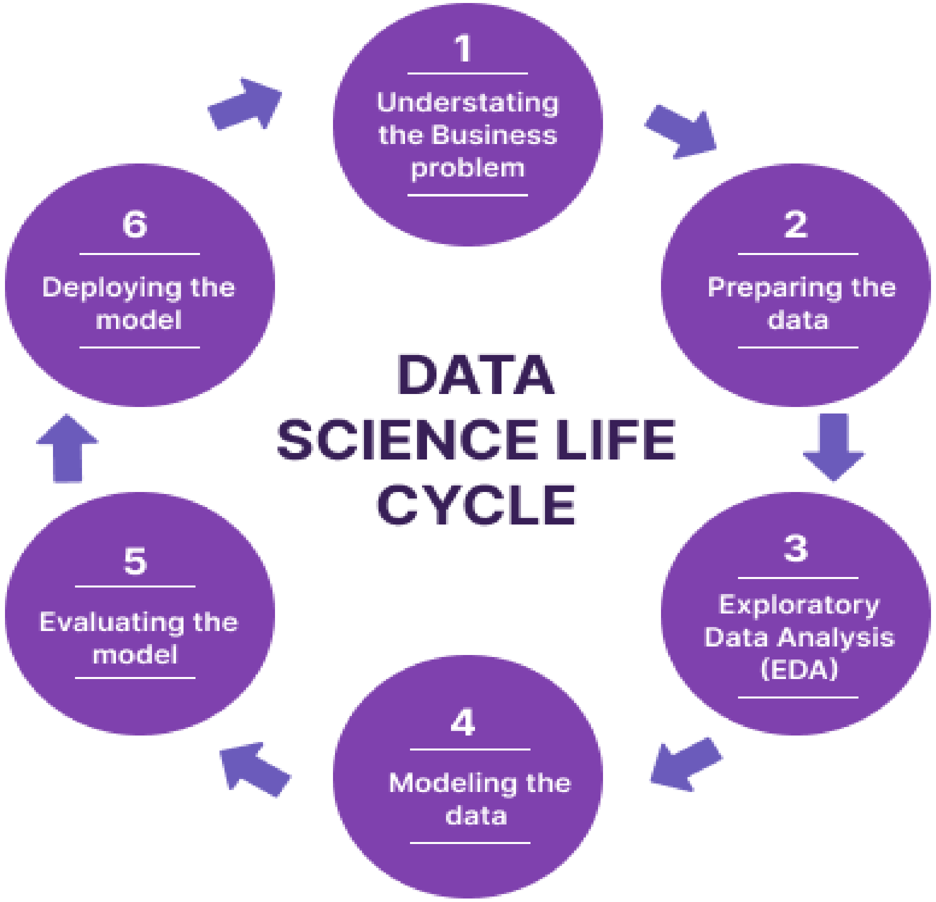
  Data Science is a process of using data to find solutions or to predict outcome for a problem statement.

Why is Data Science is important?

- Data Science is the branch of Artificial Intelligence or we can say it is the future of AI.
- Now a days every industry require Data Science is for improvement of business and customer satisfaction.
- The main purpose is to transfer the row data into valuable information.
- There are many sources of data such as texts, videos, images etc.
- For all these consequences we have to go for advanced analytical tools and algorithms to draw meaningful insights

Data Science Life cycle

- Data Science lifecycle is an extensive step-by-step guide that illustrates how machine learning and other analytical techniques can be used to generate insights and predictions from data to accomplish a business objective



Real World Applications of Data Science

- ❑ **In search engines**
- ❑ **In transport**
- ❑ **In finance**
- ❑ **In E-Commerce**
- ❑ **In Health Care**
- ❑ **Image Recognition**
- ❑ **Targeting Recommendation**
- ❑ **Airline Routing Planning**
- ❑ **Data Science in Gaming**
- ❑ **Medicine and Drug Development**
- ❑ **In Delivery Logistics**
- ❑ **Automobiles**

Role of Data Scientists

- Data Science identify the questions to understand the business problem
- Gather data from various sources, public data
- Process the data and convert it into suitable format for analysis
- Visualizing data
- Feed the data into algorithms or a statistical model
- Prepare the results and insights by deploying models into applications

Skills Required for Data Scientists

- Posses knowledge of Python programming, R language, SQL, Database, SAS and sometimes Java, Scala.
- Mathematical expertise, statistical thinking, technical acumen and multi-model communication skills, curious mind, creativity.
- Machine Learning Algorithms such as Regression, clustering, Decision Tree, Support Vector Machines, Naïve Bayes etc.

Job Roles for Data Scientist

- Data Scientist
- Machine Learning Engineer
- Data Consultant
- Data Analyst

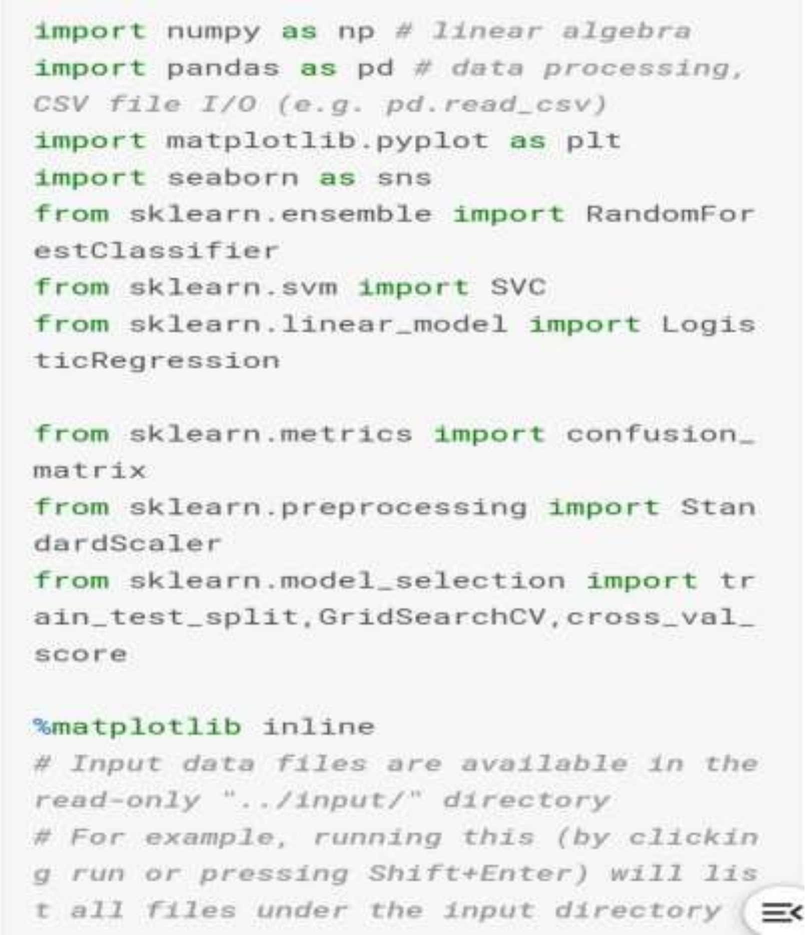
project

Project Description

Problem Statement **:** Create a classification model to predict whether credit risk is good or bad.

Context : As a banking, Financial institution is interest is to know the potential financial whereabouts of the customers in order to determine whether the credit risk associated with them is good or bad.

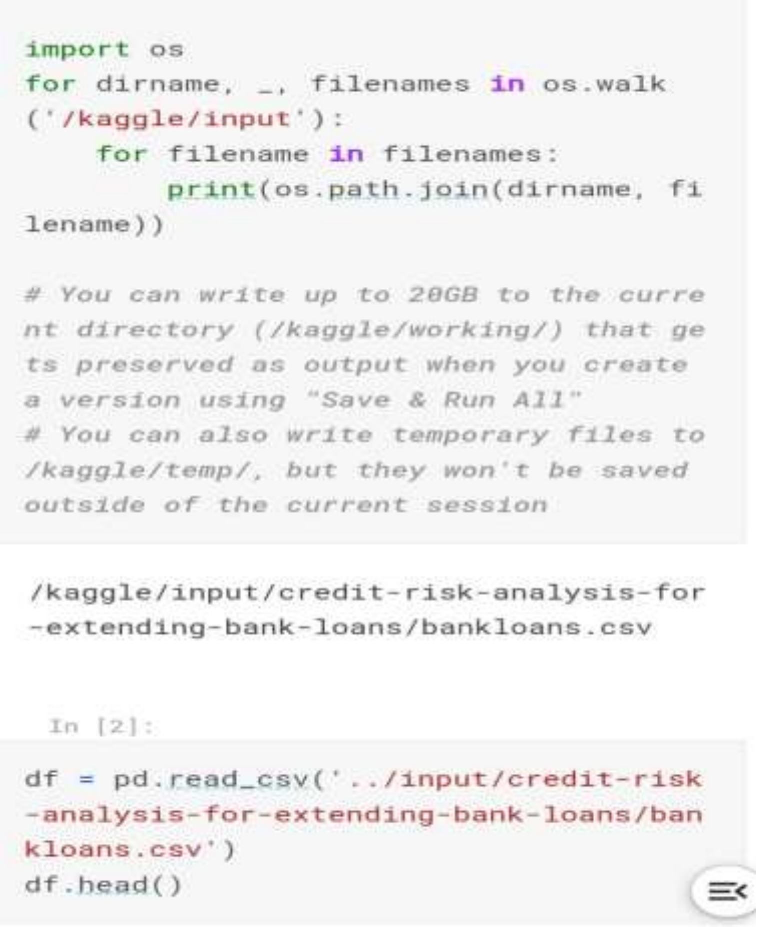
The dataset consists of 21 features of the customers.It could be used to predict the customer could be given credit. Many features require data cleaning.



```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression

from sklearn.metrics import confusion_matrix
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split,GridSearchCV,cross_val_score

%matplotlib inline
# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory
```



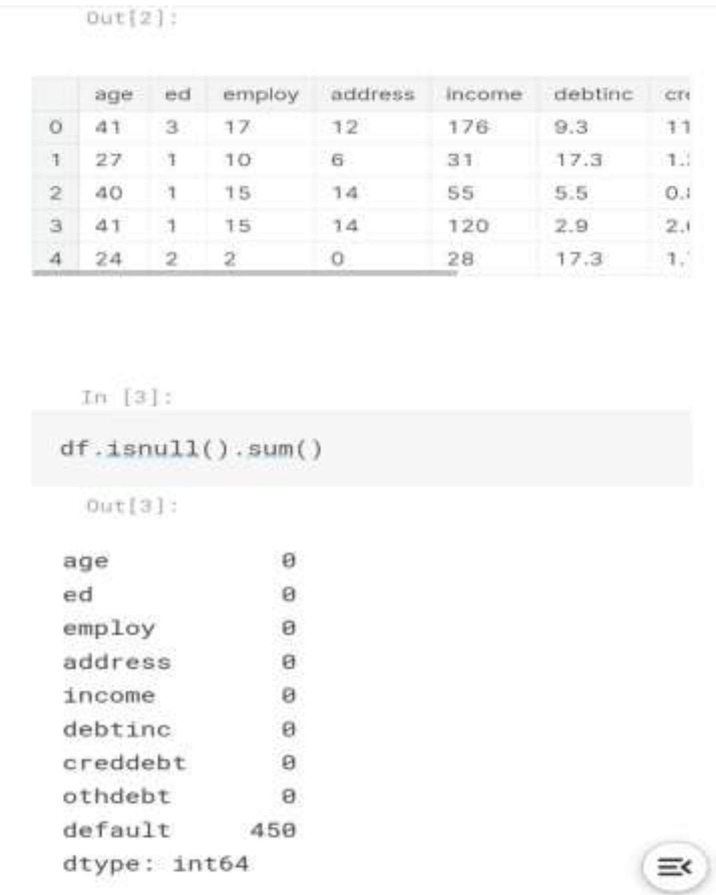
```
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```

```
/kaggle/input/credit-risk-analysis-for-extending-bank-loans/bankloans.csv
```

```
In [2]:
```

```
df = pd.read_csv('../input/credit-risk-analysis-for-extending-bank-loans/bankloans.csv')
df.head()
```



Out[2]:

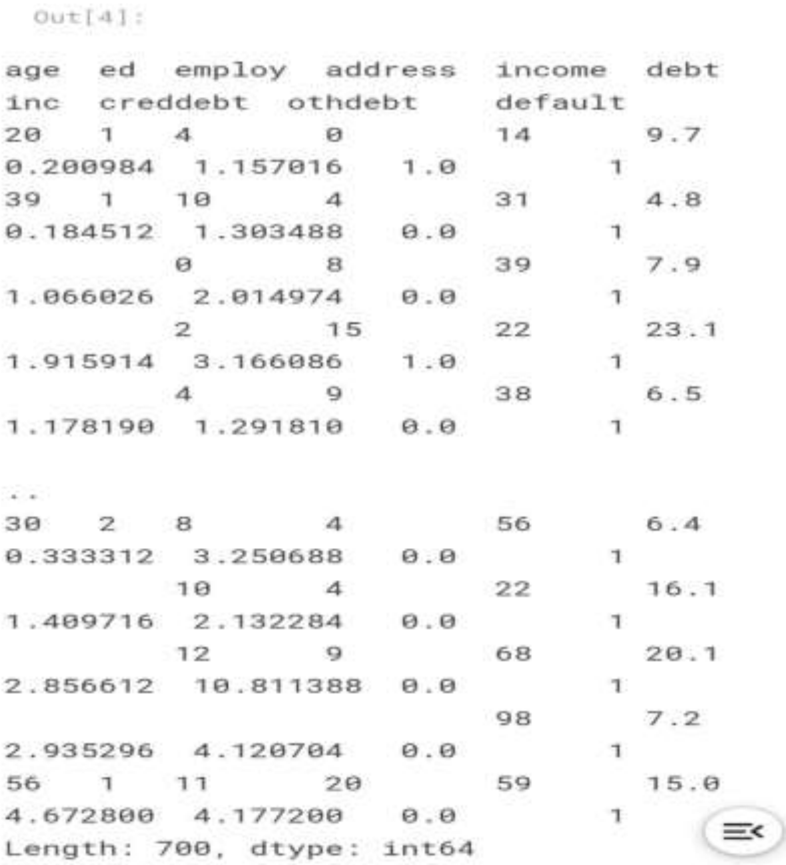
	age	ed	employ	address	income	debtinc	cr
0	41	3	17	12	176	9.3	11
1	27	1	10	6	31	17.3	1.
2	40	1	15	14	55	5.5	0.
3	41	1	15	14	120	2.9	2.
4	24	2	2	0	28	17.3	1.

In [3]:

```
df.isnull().sum()
```

Out[3]:

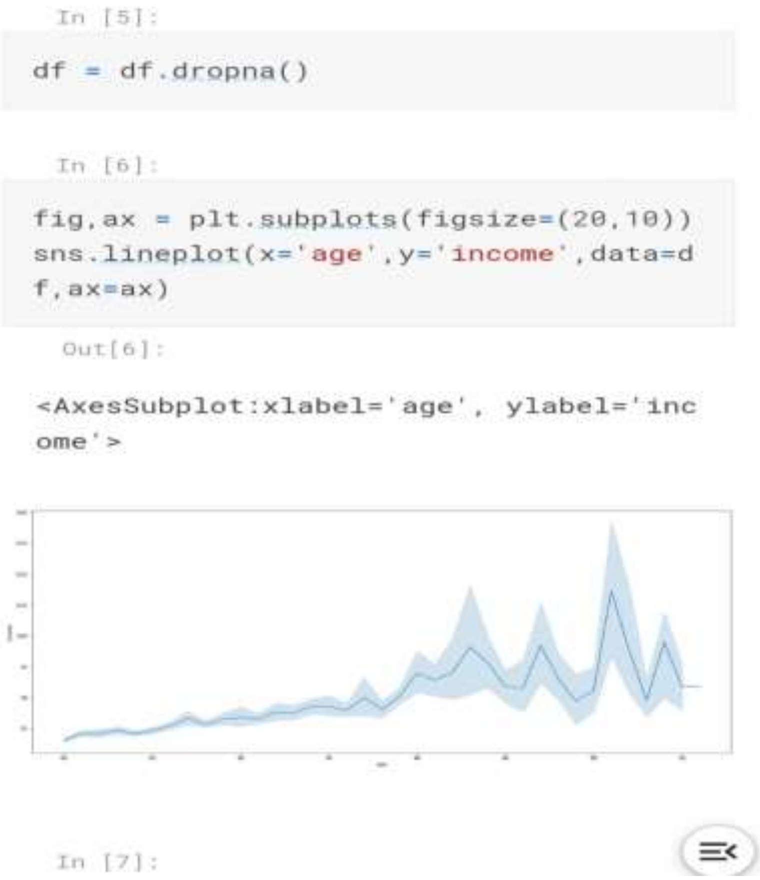
```
age           0
ed            0
employ        0
address       0
income        0
debtinc       0
creddebt      0
othdebt       0
default     450
dtype: int64
```



Out[4]:

```
age  ed  employ  address   income  debt
inc  creddebt  othdebt     default
20   1   4       0         14      9.7
0.200984  1.157016   1.0        1
39   1   10      4         31      4.8
0.184512  1.303488   0.0        1
         0       8         39      7.9
1.066026  2.014974   0.0        1
         2       15        22      23.1
1.915914  3.166086   1.0        1
         4       9         38      6.5
1.178190  1.291810   0.0        1


..
30   2   8       4         56      6.4
0.333312  3.250688   0.0        1
         10      4         22      16.1
1.409716  2.132284   0.0        1
         12      9         68      20.1
2.856612  10.811388  0.0        1
                           98      7.2
2.935296  4.120704   0.0        1
56   1   11      20        59      15.0
4.672800  4.177200   0.0        1
Length: 700, dtype: int64
```



In [5]:

```
df = df.dropna()
```

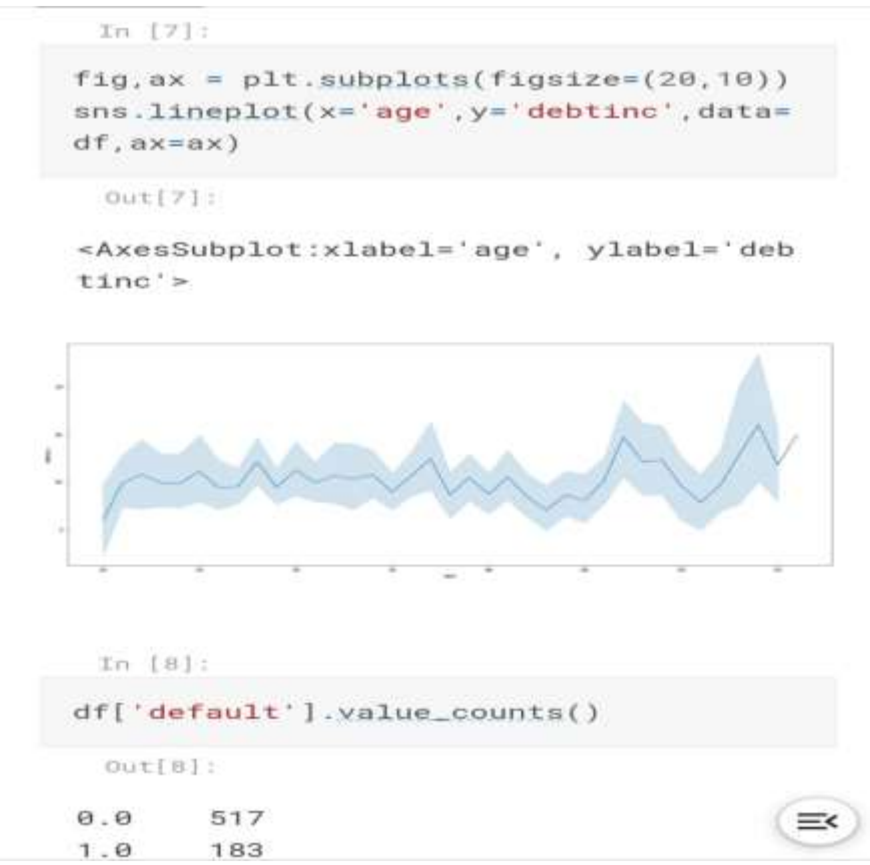
In [6]:

```
fig,ax = plt.subplots(figsize=(20,10))
sns.lineplot(x='age',y='income',data=d
f,ax=ax)
```

Out[6]:

```
<AxesSubplot:xlabel='age', ylabel='inc
ome'>
```

In [7]:



In [7]:

```
fig,ax = plt.subplots(figsize=(20,10))
sns.lineplot(x='age',y='debtinc',data=
df,ax=ax)
```

Out[7]:

```
<AxesSubplot:xlabel='age', ylabel='deb
tinc'>
```

In [8]:

```
df['default'].value_counts()
```

Out[8]:

```
0.0      517
1.0      183
```

thank you