```
#IMPORTING LIBRARIES
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
from sklearn.naive_bayes import MultinomialNB
from sklearn.multiclass import OneVsRestClassifier
from sklearn import metrics
from sklearn.metrics import accuracy_score
from pandas.plotting import scatter_matrix
from sklearn.neighbors import KNeighborsClassifier
from sklearn import metrics
```

```
#READING INPUT CSV FILE
import pandas as pd
path="/content/drive/MyDrive/deeplearning/UpdatedResumeDataSet.csv"
a=pd.read_csv(path)
```

```
#BASIC INFORMATION FROM DAT SET
a.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 962 entries, 0 to 961
Data columns (total 2 columns):
 #   Column    Non-Null Count  Dtype 
---  ------    --------------  ----- 
 0   Category  962 non-null    object
 1   Resume    962 non-null    object
dtypes: object(2)
memory usage: 15.2+ KB
```

```
a.shape
```

```
(962, 2)
```

```
a.index
```

```
RangeIndex(start=0, stop=962, step=1)
```

```
a.columns
```

```
Index(['Category', 'Resume'], dtype='object')
```

```
a.describe
```

```
<bound method NDFrame.describe of          Category                                             Resume
0    Data Science  Skills * Programming Languages: Python (pandas...
1    Data Science  Education Details \r\nMay 2013 to May 2017 B.E...
2    Data Science  Areas of Interest Deep Learning, Control Syste...
3    Data Science  Skills â¢ R â¢ Python â¢ SAP HANA â¢ Table...
4    Data Science  Education Details \r\n MCA   YMCAUST,  Faridab...
..            ...                                                ...
957       Testing  Computer Skills: â¢ Proficient in MS office (...
958       Testing  â Willingness to accept the challenges. â ...
959       Testing  PERSONAL SKILLS â¢ Quick learner, â¢ Eagerne...
960       Testing  COMPUTER SKILLS & SOFTWARE KNOWLEDGE MS-Power ...
961       Testing  Skill Set OS Windows XP/7/8/8.1/10 Database MY...

[962 rows x 2 columns]>
```

```
a.isna().sum()
```

```
Category    0
Resume      0
dtype: int64
```

```
#ADDING COLUMN TO DATA SET
a['clean resume']=''
```

```
a.shape
```

```
(962, 3)
```

```
a.head(5)
```

	Category	Resume	clean resume
0	Data Science	Skills * Programming Languages: Python (pandas...	
1	Data Science	Education Details \r\nMay 2013 to May 2017 B.E...	
2	Data Science	Areas of Interest Deep Learning, Control Syste...	
3	Data Science	Skills â□¢ R â□¢ Python â□¢ SAP HANA â□¢ Table...	
4	Data Science	Education Details \r\n MCA YMCAUST, Faridab...	

```
print(a['Category'])
```

```
0      Data Science
1      Data Science
2      Data Science
3      Data Science
4      Data Science
           ...
957         Testing
958         Testing
959         Testing
960         Testing
961         Testing
Name: Category, Length: 962, dtype: object
```

```
q=(a['Category'].unique())
print(q)
```

```
['Data Science' 'HR' 'Advocate' 'Arts' 'Web Designing'
 'Mechanical Engineer' 'Sales' 'Health and fitness' 'Civil Engineer'
 'Java Developer' 'Business Analyst' 'SAP Developer' 'Automation Testing'
 'Electrical Engineering' 'Operations Manager' 'Python Developer'
 'DevOps Engineer' 'Network Security Engineer' 'PMO' 'Database' 'Hadoop'
 'ETL Developer' 'DotNet Developer' 'Blockchain' 'Testing']
```

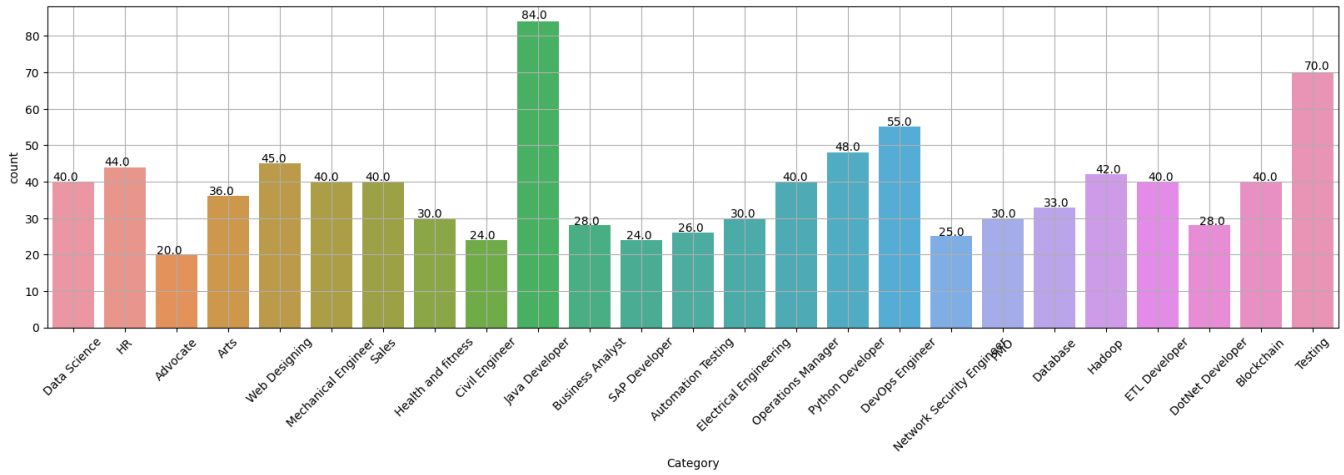
```
print(len(q))
```

```
25
```

```
print(a['Category'].value_counts())
```

```
Java Developer               84
Testing                      70
DevOps Engineer              55
Python Developer             48
Web Designing                45
HR                           44
Hadoop                       42
Blockchain                   40
ETL Developer                40
Operations Manager           40
Data Science                 40
Sales                        40
Mechanical Engineer          40
Arts                         36
Database                     33
Electrical Engineering       30
Health and fitness           30
PMO                          30
Business Analyst             28
DotNet Developer             28
Automation Testing           26
Network Security Engineer    25
SAP Developer                24
Civil Engineer               24
Advocate                     20
Name: Category, dtype: int64
```

```
#PLOTTING THE DATA
import seaborn as sns
plt.figure(figsize=(20,5))
plt.xticks(rotation=45)
ax=sns.countplot(x="Category", data=a)
for p in ax.patches:
    ax.annotate(str(p.get_height()), (p.get_x() * 1.01 , p.get_height() * 1.01))
plt.grid()
```




```
#REMOVING UNWANTED CHARACTERS FROM STRING
import re
def cleanResume(resumeText):
    resumeText = re.sub('http\S+\s*', ' ', resumeText)  # remove URLs
    resumeText = re.sub('RT|cc', ' ', resumeText)  # remove RT and cc
    resumeText = re.sub('#\S+', '', resumeText)  # remove hashtags
    resumeText = re.sub('@\S+', '  ', resumeText)  # remove mentions
    resumeText = re.sub('[%s]' % re.escape("""!"#$%&'()*+,-./:;<=>?@[\]^_`{|}~"""), ' ', resumeText)  # remove punctuations
    resumeText = re.sub(r'[^\x00-\x7f]',r' ', resumeText)
    resumeText = re.sub('\s+', ' ', resumeText)  # remove extra whitespace
    return resumeText

a['clean resume'] =a.Resume.apply(lambda x: cleanResume(x))
print(a['clean resume'])
```

```
0      Skills Programming Languages Python pandas num...
1      Education Details May 2013 to May 2017 B E UIT...
2      Areas of Interest Deep Learning Control System...
3      Skills R Python SAP HANA Tableau SAP HANA SQL ...
4      Education Details MCA YMCAUST Faridabad Haryan...
                             ...
957    Computer Skills Proficient in MS office Word B...
958     Willingness to a ept the challenges Positive ...
959    PERSONAL SKILLS Quick learner Eagerness to lea...
960    COMPUTER SKILLS SOFTWARE KNOWLEDGE MS Power Po...
961    Skill Set OS Windows XP 7 8 8 1 10 Database MY...
Name: clean resume, Length: 962, dtype: object
```

```
a.head()
```

	Category	Resume	clean resume
0	Data Science	Skills * Programming Languages: Python (pandas...	Skills Programming Languages Python pandas num...
1	Data Science	Education Details \r\nMay 2013 to May 2017 B.E...	Education Details May 2013 to May 2017 B E UIT...
2	Data Science	Areas of Interest Deep Learning, Control Syste...	Areas of Interest Deep Learning Control System...
3	Data Science	Skills â□¢ R â□¢ Python â□¢ SAP HANA â□¢ Table...	Skills R Python SAP HANA Tableau SAP HANA SQL ...
4	Data Science	Education Details \r\n MCA YMCAUST, Faridab...	Education Details MCA YMCAUST Faridabad Haryan...

```
data=a.copy()
print(data.shape)
```

```
(962, 3)
```

```
from sklearn.preprocessing import LabelEncoder
var_mod = ['Category']
le = LabelEncoder()
for i in var_mod:
    data[i] = le.fit_transform(data[i])
```

```
data.head()
```

	Category	Resume	clean resume
0	6	Skills * Programming Languages: Python (pandas...	Skills Programming Languages Python pandas num...
1	6	Education Details \r\nMay 2013 to May 2017 B.E...	Education Details May 2013 to May 2017 B E UIT...

```
data['Category'].value_counts()
```

```
15    84
23    70
8     55
20    48
24    45
12    44
13    42
3     40
10    40
18    40
6     40
22    40
16    40
1     36
7     33
11    30
14    30
19    30
4     28
9     28
2     26
17    25
21    24
5     24
0     20
Name: Category, dtype: int64
```

```
data.isna()
```

	Category	Resume	clean resume
0	False	False	False
1	False	False	False
2	False	False	False
3	False	False	False
4	False	False	False
...	...	...	...
957	False	False	False
958	False	False	False
959	False	False	False
960	False	False	False
961	False	False	False

962 rows × 3 columns

```
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from scipy.sparse import hstack

requiredText = data['clean resume'].values
requiredTarget = data['Category'].values
print(requiredText)
print(requiredTarget)
```

```
 'Skills Exceptional communication and networking skills Su essful working in a team environment as well as independently Ability
 'CORE COMPETENCIES Maintain processes to ensure project management documentation reports and plans are relevant a urate and comp
 'AREA OF EXPE ISE PROFILE Around 10 plus years proven experience with best global brand Wipro with below expertise PMO ITIL Mana
 'Skills Exceptional communication and networking skills Su essful working in a team environment as well as independently Ability
 'CORE COMPETENCIES Maintain processes to ensure project management documentation reports and plans are relevant a urate and comp
 'AREA OF EXPE ISE PROFILE Around 10 plus years proven experience with best global brand Wipro with below expertise PMO ITIL Mana
 'Skills Exceptional communication and networking skills Su essful working in a team environment as well as independently Ability
 'CORE COMPETENCIES Maintain processes to ensure project management documentation reports and plans are relevant a urate and comp
 'AREA OF EXPE ISE PROFILE Around 10 plus years proven experience with best global brand Wipro with below expertise PMO ITIL Mana
 'Skills Exceptional communication and networking skills Su essful working in a team environment as well as independently Ability
 'CORE COMPETENCIES Maintain processes to ensure project management documentation reports and plans are relevant a urate and comp
 'AREA OF EXPE ISE PROFILE Around 10 plus years proven experience with best global brand Wipro with below expertise PMO ITIL Mana
 'Skills Exceptional communication and networking skills Su essful working in a team environment as well as independently Ability
 'CORE COMPETENCIES Maintain processes to ensure project management documentation reports and plans are relevant a urate and comp
 'AREA OF EXPE ISE PROFILE Around 10 plus years proven experience with best global brand Wipro with below expertise PMO ITIL Mana
 'Skills Exceptional communication and networking skills Su essful working in a team environment as well as independently Ability
 'CORE COMPETENCIES Maintain processes to ensure project management documentation reports and plans are relevant a urate and comp
 'AREA OF EXPE ISE PROFILE Around 10 plus years proven experience with best global brand Wipro with below expertise PMO ITIL Mana
 'Skills Exceptional communication and networking skills Su essful working in a team environment as well as independently Ability
 'TECHNICAL EXPE ISE DB Languages SQL Database Tools SQL Server 2014 2017 Postgresql 9 5 9 6 Oracle 11gR2 Operating Systems Redha
 'Technical Expertise Operating Systems Microsoft Window Server 2003 2008 2008 R2 2012 Database Technologies SQL Server Sybase AS
 'TECHNICAL SKILLS Operating Systems MS Windows Server 2012 2008 XP Software and Tools MS LiteSpeed Idera SQL Safe SSMS Upgrade A
 'SKILLSET Oracle DBA MySQL MARIADB PostgreSQL Database Administration ITSKILLS SQL Oracle 10g 11g MYSQL MariaDB postgreSQL Windo
 'Education Details January 2016 BSc Mumbai Maharashtra Mumbai University January 2013 H S C Maharashtra Board January 2011 S S C
 'TECHNICAL SKILL Operating System LINUX Windows Server 2012 R2 Windows 98 Windows 2000 XP Tools Utility Packages SQL Loader SQL
 'Technical Skills Databases Oracle RDBMS 10g 11g 12c Technology utilities Data Pump RMAN Data guard ASM RAC Golden Gate Tools OC
 'Software Skills RDBMS MS SQL SERVER 2000 2005 2008 2012 2014 Operating Systems WINDOWS XP 7 WINDOWS SERVER 2008 12 Fundamentals
 'Areas of Expertise Oracle Databases 12c 11g 10g Weblogic 12c 11g Grid Infrastructure RMAN ASM Middleware OIM OAM SOA Shell Scri
 'Education Details May 2011 to May 2014 Bachelor of science Information technology Mumbai Maharashtra Mumbai university Oracle D
 'TECHNICAL SKILLS SQL Oracle v10 v11 v12 R programming Python linear regression machine learning and statistical modelling techn
 'TECHNICAL EXPE ISE DB Languages SQL Database Tools SQL Server 2014 2017 Postgresql 9 5 9 6 Oracle 11gR2 Operating Systems Redha
 'Technical Expertise Operating Systems Microsoft Window Server 2003 2008 2008 R2 2012 Database Technologies SQL Server Sybase AS
 'TECHNICAL SKILLS Operating Systems MS Windows Server 2012 2008 XP Software and Tools MS LiteSpeed Idera SQL Safe SSMS Upgrade A
 'SKILLSET Oracle DBA MySQL MARIADB PostgreSQL Database Administration ITSKILLS SQL Oracle 10g 11g MYSQL MariaDB postgreSQL Windo
 'Education Details January 2016 BSc Mumbai Maharashtra Mumbai University January 2013 H S C Maharashtra Board January 2011 S S C
 'TECHNICAL SKILL Operating System LINUX Windows Server 2012 R2 Windows 98 Windows 2000 XP Tools Utility Packages SQL Loader SQL
 'Technical Skills Databases Oracle RDBMS 10g 11g 12c Technology utilities Data Pump RMAN Data guard ASM RAC Golden Gate Tools OC
 'Software Skills RDBMS MS SQL SERVER 2000 2005 2008 2012 2014 Operating Systems WINDOWS XP 7 WINDOWS SERVER 2008 12 Fundamentals
 'Areas of Expertise Oracle Databases 12c 11g 10g Weblogic 12c 11g Grid Infrastructure RMAN ASM Middleware OIM OAM SOA Shell Scri
 'Education Details May 2011 to May 2014 Bachelor of science Information technology Mumbai Maharashtra Mumbai university Oracle D
 'TECHNICAL SKILLS SQL Oracle v10 v11 v12 R programming Python linear regression machine learning and statistical modelling techn
 'TECHNICAL EXPE ISE DB Languages SQL Database Tools SQL Server 2014 2017 Postgresql 9 5 9 6 Oracle 11gR2 Operating Systems Redha
```

```
word_vectorizer = TfidfVectorizer(
    sublinear_tf=True,
    stop_words='english')
word_vectorizer.fit(requiredText)
WordFeatures = word_vectorizer.transform(requiredText)
X_train,X_test,y_train,y_test = train_test_split(WordFeatures,requiredTarget,random_state=42, test_size=0.2,
                                                 shuffle=True, stratify=requiredTarget)
print(X_train.shape)
print(X_test.shape)
```

```
(769, 7351)
(193, 7351)
```

```
#ALGORITHM WORKING
clf = OneVsRestClassifier(KNeighborsClassifier())
clf.fit(X_train, y_train)
prediction = clf.predict(X_test)
print('Accuracy of KNeighbors Classifier on training set: {:.2f}'.format(clf.score(X_train, y_train)))
print('Accuracy of KNeighbors Classifier on test set:     {:.2f}'.format(clf.score(X_test, y_test)))
```

```
Accuracy of KNeighbors Classifier on training set: 0.99
Accuracy of KNeighbors Classifier on test set:     0.98
```

```
#TESTING OF DATA:
import re

def clResume(resumeText):
    # Remove URLs
    resumeText = re.sub('http\S+\s*', ' ', resumeText)

    # Remove RT and cc
    resumeText = re.sub('RT|cc', ' ', resumeText)

    # Remove hashtags
    resumeText = re.sub('#\S+', '', resumeText)

    # Remove mentions
    resumeText = re.sub('@\S+', '  ', resumeText)

    # Remove punctuations (except for special characters like ., !, ?)
    resumeText = re.sub('[^A-Za-z0-9.,!?]+', ' ', resumeText)
```

```
    # Remove non-ASCII characters
    resumeText = re.sub(r'[^\x00-\x7F]+', ' ', resumeText)

    # Remove extra whitespace
    resumeText = re.sub('\s+', ' ', resumeText).strip()

    return resumeText


# Assuming you have new resume data in a variable called 'new_resumes'
new_resumes = [
    "Experienced@ data analyst with $skills in Python and SQL.",
    "Graphic designer specializing in digital# art and illustration.",
    "Customer service representative with excellent communication skills.",
]

# Clean and preprocess the new data
cleaned_new_resumes = [clResume(resume) for resume in new_resumes]

print(cleaned_new_resumes)
```

```
['Experienced data analyst with skills in Python and SQL.', 'Graphic designer specializing in digital art and illustration.', 'Custo
```

```
# Transform the cleaned new data using the same TF-IDF vectorizer
from sklearn.feature_extraction.text import CountVectorizer

new_resume_features = word_vectorizer.transform(cleaned_new_resumes)

# Use the trained classifier to make predictions on the new data
new_predictions = clf.predict(new_resume_features)

# Map the predicted labels back to their original categories using LabelEncoder
predicted_categories = le.inverse_transform(new_predictions)
print("1",new_resume_features)
print("2",new_predictions)
print("3",predicted_categories)
```

```
1   (0, 6239)   0.30926867582777434
  (0, 6080)     0.22775251583612832
  (0, 5266)     0.3845964091548369
  (0, 2490)     0.6201293311877528
  (0, 1777)     0.26157238387037235
  (0, 484)      0.5015842704314555
  (1, 2005)     0.5830382722449241
  (1, 1934)     0.48696502235449235
  (1, 604)      0.6503317923190767
  (2, 6080)     0.21822806802118672
  (2, 5965)     0.3595394738060956
  (2, 5546)     0.6113176711756693
  (2, 2452)     0.4954114635665883
  (2, 1738)     0.29725030122476
  (2, 1399)     0.3400122383090542
2 [20 24  2]
3 ['Python Developer' 'Web Designing' 'Automation Testing']
```

```
# Print the predicted categories for the new resumes
for resume, category in zip(new_resumes, predicted_categories):
    print(f"Resume: {resume}")
    print(f"Predicted Category: {category}")
    print()
```

```
Resume: Experienced data analyst with skills in Python and SQL.
Predicted Category: Python Developer

Resume: Graphic designer specializing in digital art and illustration.
Predicted Category: Web Designing

Resume: Customer service representative with excellent communication skills.
Predicted Category: Automation Testing
```