product_matching_Results_Improve

September 16, 2022

[1]:
```
import pandas as pd
import numpy as np
import time

import re
#from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.metrics import f1_score
from sklearn.metrics import accuracy_score
from sklearn.metrics import precision_score
from sklearn.metrics import recall_score
from sklearn.metrics import confusion_matrix
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
import warnings
warnings.filterwarnings('ignore')
from sklearn import preprocessing
```

1 Converting Json data to DataFrame

[2]:
```
data = pd.read_json('computers_train_xlarge.json',lines = True)
data.to_csv('train.csv',index = False)
df = pd.read_csv('train.csv')
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 68461 entries, 0 to 68460
Data columns (total 20 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   id_left               68461 non-null  int64
 1   category_left         68461 non-null  object
 2   cluster_id_left       68461 non-null  int64
 3   id_right              68461 non-null  int64
 4   category_right        68461 non-null  object
 5   cluster_id_right      68461 non-null  int64
 6   label                 68461 non-null  int64
```

```
 7   pair_id                 68461 non-null  object
 8   brand_left              34233 non-null  object
 9   brand_right             34245 non-null  object
 10  description_left        47460 non-null  object
 11  description_right       48360 non-null  object
 12  keyValuePairs_left      18765 non-null  object
 13  keyValuePairs_right     20275 non-null  object
 14  price_left              11521 non-null  object
 15  price_right             11492 non-null  object
 16  specTableContent_left   20897 non-null  object
 17  specTableContent_right  22157 non-null  object
 18  title_left              68461 non-null  object
 19  title_right             68461 non-null  object
dtypes: int64(5), object(15)
memory usage: 10.4+ MB
```

[3]:
```
df.columns
```

[3]:
```
Index(['id_left', 'category_left', 'cluster_id_left', 'id_right',
       'category_right', 'cluster_id_right', 'label', 'pair_id', 'brand_left',
       'brand_right', 'description_left', 'description_right',
       'keyValuePairs_left', 'keyValuePairs_right', 'price_left',
       'price_right', 'specTableContent_left', 'specTableContent_right',
       'title_left', 'title_right'],
      dtype='object')
```

2 Data Cleaning : Removing Unwanted Columns

[4]:
```
df.drop(df.columns[[2,5,6,8,9,12,13,14,15,16,17]],axis=1,inplace =True)
df
```

[4]:

	id_left	category_left	id_right	\
0	2551242	Computers_and_Accessories	16272671	
1	16757469	Computers_and_Accessories	16476204	
2	232007	Computers_and_Accessories	16442945	
3	2066119	Computers_and_Accessories	12411100	
4	6656540	Computers_and_Accessories	2639431	
…	…	…	…	
68456	6493497	Computers_and_Accessories	16149764	
68457	17075265	Computers_and_Accessories	17346839	
68458	16408794	Computers_and_Accessories	3675781	
68459	13925964	Computers_and_Accessories	15659664	
68460	8308259	Computers_and_Accessories	16419137	

	category_right	pair_id	\
0	Computers_and_Accessories	2551242#16272671	

```
1      Computers_and_Accessories  16757469#16476204
2      Computers_and_Accessories    232007#16442945
3      Computers_and_Accessories   2066119#12411100
4      Computers_and_Accessories    6656540#2639431
…                            …                …
68456  Computers_and_Accessories   6493497#16149764
68457  Computers_and_Accessories  17075265#17346839
68458  Computers_and_Accessories   16408794#3675781
68459  Computers_and_Accessories  13925964#15659664
68460  Computers_and_Accessories   8308259#16419137

                                          description_left  \
0      "DDR4, 2666MHz, CL16, 1.2v, XMP 2.0, Lifetime …
1      "Description:2 x 72GB 2.5-inch Serial Attached…
2       "SDSDJ-1024 BXP 1GB 9p SD Class 2 Secure Digi…
3               "DISCO DURO INTERNO SOLIDO HDD SSD"@es
4                                                   NaN
…                                                   …
68456  "Description:5 x 300GB 2.5-inch Serial Attache…
68457                                               NaN
68458                                               NaN
68459  "Built to WD's highest standards of quality an…
68460  "61 cm 250 cd / m² 1920 x 1080 Pixeles 5 ms LE…

                                         description_right  \
0                                                   NaN
1       "Description:10 x 72GB 2.5-inch Serial Attach…
2       "Description:Genuine HPE 1GB FBD PC2-5300(2x5…
3                                                   NaN
4                                                   NaN
…                                                   …
68456   "Description:2 x 72GB 2.5-inch Serial Attache…
68457   "Cost-effective SSD featuring TurboWrite and …
68458  "Quad Core Technology, 3.6GHz clock speed, 8MB…
68459                                               NaN
68460  "Longitud diagonal: 24 "; Tamaño: 16:9; Tecnol…

                                            title_left  \
0       "Corsair Vengeance LPX Black 64GB (4x16GB) DD…
1       "DH0072BALWL HP 72-GB 3G 15K 2.5 DP SAS", "Nu…
2       "SanDisk SDSDJ-1024 BXP 1GB 9p SD Class 2 Sec…
3       "DISCO DURO INTERNO SOLIDO HDD SSD KINGSTON V…
4       "Corsair Vengeance LED 32GB (2 x 16GB) DDR4 D…
…                                                   …
68456   "DG0300FARVV HP 300-GB 6G 10K 2.5 DP SAS", "N…
68457   "Samsung - 840 EVO 250GB 2.5" Solid State Dri…
68458   "Socket H4 1151 - Coffee Lake Core i7-8700K 6…
```

```
68459   "WD Blue WD5000AZLX - hard drive 500 GB SATA …
68460   "Acer KA KA240H 24" Full HD TN Negro pantalla…

                                                title_right
0       "Corsair Vengeance LPX CMK64GX4M4A2666C16 - P…
1        "DH0072BALWL HP 72-GB 3G 15K 2.5 DP SAS" "Null"
2       "397409-B21 HP 1GB (2x512MB) PC2-5300 SDRAM" …
3       "DISCO DURO SSD Kingston Technology SSDNow V3…
4       "Corsair - Vengeance LPX 32GB (2 x 16GB) DDR4…
…                                                     …
68456     "Null" "512743-001 HP 72-GB 6G 15K 2.5 DP SAS"
68457              "SSD 750 EVO 2.5" SATA III 120GB "@en
68458   "7th Generation Intel® Core  i7 7700 3.6GHz S…
68459   "m rock ships"@en-US "M-ROCK Ships New Camera…
68460   "XL2430 MON ZOWIE 24 LED PANORAMI" PANORAMI |…

[68461 rows x 9 columns]
```

3 Data Analysis

[5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 68461 entries, 0 to 68460
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id_left            68461 non-null  int64
 1   category_left      68461 non-null  object
 2   id_right           68461 non-null  int64
 3   category_right     68461 non-null  object
 4   pair_id            68461 non-null  object
 5   description_left   47460 non-null  object
 6   description_right  48360 non-null  object
 7   title_left         68461 non-null  object
 8   title_right        68461 non-null  object
dtypes: int64(2), object(7)
memory usage: 4.7+ MB
```

[6]:
```
df['category_left'].value_counts()
```

[6]:
```
Computers_and_Accessories    64958
Other_Electronics             1115
Office_Products                844
Video_Games                    575
Camera_and_Photo               286
```

```
Musical_Instruments             269
Luggage_and_Travel_Gear         256
Tools_and_Home_Improvement      108
Cellphones_and_Accessories       50
Name: category_left, dtype: int64
```

[7]:
```
df['category_right'].value_counts()
```

[7]:
```
Computers_and_Accessories     66972
Other_Electronics               463
Video_Games                     405
Office_Products                 262
Luggage_and_Travel_Gear         122
Cellphones_and_Accessories      109
Camera_and_Photo                 74
Musical_Instruments              44
Tools_and_Home_Improvement       10
Name: category_right, dtype: int64
```

[8]:
```
df['id_left'].value_counts()
```

[8]:
```
14619165    43
8364399     42
5411511     41
8167053     40
4927241     40
            ..
6810432      1
16476204     1
10969778     1
5872903      1
4604957      1
Name: id_left, Length: 4287, dtype: int64
```

[9]:
```
(df['id_left'] < 0).sum()
```

[9]:
```
0
```

[10]:
```
df['id_right'].value_counts()
```

[10]:
```
7900893     116
12050629    107
7440444     106
6443065      97
13889537     97
            ...
5199342       1
```

```
16567684      1
16123988      1
13661066      1
7867351       1
Name: id_right, Length: 4144, dtype: int64
```

[11]:
```
df['description_left'].isnull().sum()
```

[11]:
```
21001
```

[12]:
```
df['title_left'].isnull().sum()
```

[12]:
```
0
```

4 Filling the Null values of Description Columns by Title

[13]:
```
df.description_left.fillna(df.title_left, inplace = True)
df['title_left']
```

[13]:
```
0         "Corsair Vengeance LPX Black 64GB (4x16GB) DD…
1         "DH0072BALWL HP 72-GB 3G 15K 2.5 DP SAS", "Nu…
2         "SanDisk SDSDJ-1024 BXP 1GB 9p SD Class 2 Sec…
3         "DISCO DURO INTERNO SOLIDO HDD SSD KINGSTON V…
4         "Corsair Vengeance LED 32GB (2 x 16GB) DDR4 D…
                              …
68456     "DG0300FARVV HP 300-GB 6G 10K 2.5 DP SAS", "N…
68457     "Samsung - 840 EVO 250GB 2.5" Solid State Dri…
68458     "Socket H4 1151 - Coffee Lake Core i7-8700K 6…
68459     "WD Blue WD5000AZLX - hard drive 500 GB SATA …
68460     "Acer KA KA240H 24" Full HD TN Negro pantalla…
Name: title_left, Length: 68461, dtype: object
```

[14]:
```
df['description_left'].isnull().sum()
```

[14]:
```
0
```

[15]:
```
df['description_right'].isnull().sum()
```

[15]:
```
20101
```

[16]:
```
df['title_right'].isnull().sum()
```

[16]:
```
0
```

[17]:
```
df.description_right.fillna(df.title_left, inplace = True)
df['description_right']
```

[17]:
```
0         "Corsair Vengeance LPX Black 64GB (4x16GB) DD…
1         "Description:10 x 72GB 2.5-inch Serial Attach…
2         "Description:Genuine HPE 1GB FBD PC2-5300(2x5…
3         "DISCO DURO INTERNO SOLIDO HDD SSD KINGSTON V…
4         "Corsair Vengeance LED 32GB (2 x 16GB) DDR4 D…
                              …
68456     "Description:2 x 72GB 2.5-inch Serial Attache…
68457     "Cost-effective SSD featuring TurboWrite and …
68458    "Quad Core Technology, 3.6GHz clock speed, 8MB…
68459     "WD Blue WD5000AZLX - hard drive 500 GB SATA …
68460    "Longitud diagonal: 24 "; Tamaño: 16:9; Tecnol…
Name: description_right, Length: 68461, dtype: object
```

[18]:
```
df['description_right'].isnull().sum()
```

[18]:
```
0
```

5 Comparing category left and category right

[19]:
```
df['category_match'] = np.where(df['category_left'] == df['category_right'],1,0)
df.head()

#1-Match / 0-Unmatch
```

[19]:
```
    id_left             category_left  id_right            category_right  \
0   2551242  Computers_and_Accessories  16272671  Computers_and_Accessories
1  16757469  Computers_and_Accessories  16476204  Computers_and_Accessories
2    232007  Computers_and_Accessories  16442945  Computers_and_Accessories
3   2066119  Computers_and_Accessories  12411100  Computers_and_Accessories
4   6656540  Computers_and_Accessories   2639431  Computers_and_Accessories

             pair_id                                    description_left  \
0   2551242#16272671  "DDR4, 2666MHz, CL16, 1.2v, XMP 2.0, Lifetime …
1  16757469#16476204  "Description:2 x 72GB 2.5-inch Serial Attached…
2    232007#16442945   "SDSDJ-1024 BXP 1GB 9p SD Class 2 Secure Digi…
3   2066119#12411100            "DISCO DURO INTERNO SOLIDO HDD SSD"@es
4    6656540#2639431   "Corsair Vengeance LED 32GB (2 x 16GB) DDR4 D…

                                  description_right  \
0   "Corsair Vengeance LPX Black 64GB (4x16GB) DD…
1   "Description:10 x 72GB 2.5-inch Serial Attach…
2   "Description:Genuine HPE 1GB FBD PC2-5300(2x5…
3   "DISCO DURO INTERNO SOLIDO HDD SSD KINGSTON V…
4   "Corsair Vengeance LED 32GB (2 x 16GB) DDR4 D…

                                       title_left  \
```

```
0   "Corsair Vengeance LPX Black 64GB (4x16GB) DD…
1   "DH0072BALWL HP 72-GB 3G 15K 2.5 DP SAS", "Nu…
2   "SanDisk SDSDJ-1024 BXP 1GB 9p SD Class 2 Sec…
3   "DISCO DURO INTERNO SOLIDO HDD SSD KINGSTON V…
4   "Corsair Vengeance LED 32GB (2 x 16GB) DDR4 D…
```

	title_right	category_match
0	"Corsair Vengeance LPX CMK64GX4M4A2666C16 - P…	1
1	"DH0072BALWL HP 72-GB 3G 15K 2.5 DP SAS" "Null"	1
2	"397409-B21 HP 1GB (2x512MB) PC2-5300 SDRAM" …	1
3	"DISCO DURO SSD Kingston Technology SSDNow V3…	1
4	"Corsair - Vengeance LPX 32GB (2 x 16GB) DDR4…	1

[20]:
```
df['category_match'].value_counts()
```

[20]:
```
1    65018
0     3443
Name: category_match, dtype: int64
```

[21]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 68461 entries, 0 to 68460
Data columns (total 10 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id_left            68461 non-null  int64
 1   category_left      68461 non-null  object
 2   id_right           68461 non-null  int64
 3   category_right     68461 non-null  object
 4   pair_id            68461 non-null  object
 5   description_left   68461 non-null  object
 6   description_right  68461 non-null  object
 7   title_left         68461 non-null  object
 8   title_right        68461 non-null  object
 9   category_match     68461 non-null  int32
dtypes: int32(1), int64(2), object(7)
memory usage: 5.0+ MB
```

[22]:
```
df.drop(df.index[df['category_match'] == 0],inplace = True)
```

[23]:
```
df['category_match'].value_counts()
```

[23]:
```
1    65018
Name: category_match, dtype: int64
```

[24]:
```
newdf=df.copy()
newdf.head()
```

[24]:

	id_left	category_left	id_right	category_right	pair_id	description_left	description_right	title_left	title_right	category_match
0	2551242	Computers_and_Accessories	16272671	Computers_and_Accessories	2551242#16272671	"DDR4, 2666MHz, CL16, 1.2v, XMP 2.0, Lifetime …	"Corsair Vengeance LPX Black 64GB (4x16GB) DD…	"Corsair Vengeance LPX Black 64GB (4x16GB) DD…	"Corsair Vengeance LPX CMK64GX4M4A2666C16 - P…	1
1	16757469	Computers_and_Accessories	16476204	Computers_and_Accessories	16757469#16476204	"Description:2 x 72GB 2.5-inch Serial Attached…	"Description:10 x 72GB 2.5-inch Serial Attach…	"DH0072BALWL HP 72-GB 3G 15K 2.5 DP SAS", "Nu…	"DH0072BALWL HP 72-GB 3G 15K 2.5 DP SAS" "Null"	1
2	232007	Computers_and_Accessories	16442945	Computers_and_Accessories	232007#16442945	"SDSDJ-1024 BXP 1GB 9p SD Class 2 Secure Digi…	"Description:Genuine HPE 1GB FBD PC2-5300(2x5…	"SanDisk SDSDJ-1024 BXP 1GB 9p SD Class 2 Sec…	"397409-B21 HP 1GB (2x512MB) PC2-5300 SDRAM" …	1
3	2066119	Computers_and_Accessories	12411100	Computers_and_Accessories	2066119#12411100	"DISCO DURO INTERNO SOLIDO HDD SSD"@es	"DISCO DURO INTERNO SOLIDO HDD SSD KINGSTON V…	"DISCO DURO INTERNO SOLIDO HDD SSD KINGSTON V…	"DISCO DURO SSD Kingston Technology SSDNow V3…	1
4	6656540	Computers_and_Accessories	2639431	Computers_and_Accessories	6656540#2639431	"Corsair Vengeance LED 32GB (2 x 16GB) DDR4 D…	"Corsair Vengeance LED 32GB (2 x 16GB) DDR4 D…	"Corsair Vengeance LED 32GB (2 x 16GB) DDR4 D…	"Corsair - Vengeance LPX 32GB (2 x 16GB) DDR4…	1

6 Applying Maching Numbers function to match the products Features

[25]:
```
def matching_numbers(title_right, title_left):

    title_right = set(re.findall(r'[0-9]+', description_right))
    title_left = set(re.findall(r'[0-9]+', title_left))
    union = title_right.union(title_left)
```

```
    intersection = title_right.intersection(title_left)

    if len(title_right)==0 and len(title_left) == 0:
        return 1
    else:
        return (len(intersection)/ len(union))
```

[ ]:

[26]:
```
def remove_spaces(text):
    text=text.strip()
    text=text.split()
    return ' '.join(text)

contraction = {'cause':'because',
               'aint': 'am not',
               'aren\'t': 'are not'}

def mapping_replacer(x,dic):
    for words in dic.keys():
        if ' ' + words + ' ' in x:
            x=x.replace(' '+ words +' ' ,' '+dic[words]+' ' )
    return x
```

[27]:
```
#Stemming, lemmetisation and tokenisation
import nltk
from nltk.tokenize import word_tokenize
from nltk.stem.wordnet import WordNetLemmatizer
from nltk.stem.lancaster import LancasterStemmer

nltk.LancasterStemmer
ls = LancasterStemmer()
lem = WordNetLemmatizer()
def lexicon_normalization(text):
    words = word_tokenize(text)


    # 1- Stemming
    words_stem = [ls.stem(w) for w in words]

    # 2- Lemmatization
    words_lem = [lem.lemmatize(w) for w in words_stem]
    return words_lem
```

[28]:
```
#Handling stopwords
from collections import Counter
def remove_stopword(text):
```

```
    stop_words = stopwords.words('english')
    stopwords_dict = Counter(stop_words)
    text = ' '.join([word for word in text.split() if word not in␣
 ↪stopwords_dict])
    return text
```

[29]:
```
#Removing links, brackets, numbers, punctuations etc.

def clean_text(text):
    '''Make text lowercase, remove text in square brackets,remove links,remove␣
 ↪punctuation
    and remove words containing numbers.'''
    text = str(text).lower()
    text = re.sub('\[.*?\]', ' ', text)
    text = re.sub('https?://\S+|www\.\S+', ' ', text)
    text = re.sub('<.*?>+', ' ', text)
    text = re.sub('_', ' ', text)
    text = re.sub('\n', '', text)
    #text = re.sub('\w*\d\w*', '', text)
    text = re.sub('\'','', text)
    text = re.sub('and', '', text)


    return text
```

[30]:
```
#Handling stopwords
from collections import Counter
def remove_stopword(text):
    stop_words = stopwords.words('english')
    stopwords_dict = Counter(stop_words)
    text = ' '.join([word for word in text.split() if word not in␣
 ↪stopwords_dict])
    return text
```

[31]:
```
import re
#df['description_right'] = df['description_right'].map(lambda x: re.sub(r'\W+',␣
 ↪' ', str(x)))
#df['description_right'] = df['description_right'].replace(r'\W+', ' ',␣
 ↪regex=True)

#df['description_right']=df['description_right'].apply(lambda x:
 ↪mapping_replacer(x, contraction))

#df['description_right']=df['description_right'].apply(lambda x:clean_text(x))
```

```
df['title_right'] = df['title_right'].map(lambda x: re.sub(r'\W+', ' ', str(x)))
df['title_right'] = df['title_right'].replace(r'\W+', ' ', regex=True)

df['title_right']=df['title_right'].apply(lambda x:mapping_replacer(x,␣
 ↪contraction))

df['title_right']=df['title_right'].apply(lambda x:clean_text(x))

#df['title_right']=df['title_right'].apply(lambda x:remove_stopword(x))

#df['title_right']=df['title_right'].apply(lambda x: lexicon_normalization(x))
```

[32]:
```
#df['description_left'] = df['description_left'].map(lambda x: re.sub(r'\W+', '␣
 ↪', str(x)))
#df['description_left'] = df['description_left'].replace(r'\W+', ' ',␣
 ↪regex=True)

#df['description_left']=df['description_left'].apply(lambda x:␣
 ↪mapping_replacer(x, contraction))

#df['description_left']=df['description_left'].apply(lambda x:clean_text(x))


df['title_left'] = df['title_left'].map(lambda x: re.sub(r'\W+', ' ', str(x)))
df['title_left'] = df['title_left'].replace(r'\W+', ' ', regex=True)

df['title_left']=df['title_left'].apply(lambda x:mapping_replacer(x,␣
 ↪contraction))

df['title_left']=df['title_left'].apply(lambda x:clean_text(x))

#df['title_left']=df['title_left'].apply(lambda x: remove_stopword(x))

#df['title_left']=df['title_left'].apply(lambda x: lexicon_normalization(x))
```

[33]:
```
df['title_right']
```

[33]:
```
0         corsair vengeance lpx cmk64gx4m4a2666c16 prij…
1             dh0072balwl hp 72 gb 3g 15k 2 5 dp sas null
2           397409 b21 hp 1gb 2x512mb pc2 5300 sdram null
3         disco duro ssd kingston technology ssdnow v30…
4         corsair vengeance lpx 32gb 2 x 16gb ddr4 3000…
                              …
68456           null 512743 001 hp 72 gb 6g 15k 2 5 dp sas
68457                     ssd 750 evo 2 5 sata iii 120gb en
```

```
68458     7th generation intel core i7 7700 3 6ghz sock…
68459     m rock ships en us m rock ships new camera ba…
68460     xl2430 mon zowie 24 led panorami panorami tra…
Name: title_right, Length: 65018, dtype: object
```

[34]:
```
df['title_left']
```

[34]:
```
0         corsair vengeance lpx black 64gb 4x16gb ddr4 …
1         dh0072balwl hp 72 gb 3g 15k 2 5 dp sas null n…
2         sisk sdsdj 1024 bxp 1gb 9p sd class 2 secure …
3         disco duro interno solido hdd ssd kingston v3…
4         corsair vengeance led 32gb 2 x 16gb ddr4 dram…
                                 …
68456     dg0300farvv hp 300 gb 6g 10k 2 5 dp sas null …
68457     samsung 840 evo 250gb 2 5 solid state drive d…
68458     socket h4 1151 coffee lake core i7 8700k 6 co…
68459     wd blue wd5000azlx hard drive 500 gb sata 6gb…
68460     acer ka ka240h 24 full hd tn negro pantalla p…
Name: title_left, Length: 65018, dtype: object
```

[ ]:
```

```

[35]:
```
def matching_numbers(title_right, title_left):

    title_right = set(re.findall(r'[0-9]+', title_right))
    title_left = set(re.findall(r'[0-9]+', title_left))
    union = title_right.union(title_left)
    intersection = title_right.intersection(title_left)

    if len(title_right)==0 and len(title_left) == 0:
        return 1
    else:
        return (len(intersection)/ len(union))
```

7 Implementing Levenshtein Text similarity

[36]:
```
import jellyfish as jf
def engineer_features(df):

    df['title_right'] = df['title_right'].str.lower()
    df['title_left'] = df['title_left'].str.lower()

    df['levenshtein_distance'] = df.apply(
    lambda x: jf.levenshtein_distance(x['title_right'],
                                      x['title_left']), axis=1)
```

```
    df['matching_numbers'] = df.apply(
    lambda x: matching_numbers(x['title_right'],
                               x['title_left']), axis=1)

    df['matching_numbers_log'] = (df['matching_numbers']+1).apply(np.log)

    df.replace([np.inf, -np.inf], np.nan, inplace=True)
    df.fillna(value=0, inplace=True)

    return df
```

8 All Required Features

[37]:
```
df.head()
```

[37]:
```
     id_left              category_left  id_right             category_right  \
0   2551242  Computers_and_Accessories  16272671  Computers_and_Accessories
1  16757469  Computers_and_Accessories  16476204  Computers_and_Accessories
2    232007  Computers_and_Accessories  16442945  Computers_and_Accessories
3   2066119  Computers_and_Accessories  12411100  Computers_and_Accessories
4   6656540  Computers_and_Accessories   2639431  Computers_and_Accessories

             pair_id                                      description_left  \
0   2551242#16272671  "DDR4, 2666MHz, CL16, 1.2v, XMP 2.0, Lifetime …
1  16757469#16476204  "Description:2 x 72GB 2.5-inch Serial Attached…
2    232007#16442945   "SDSDJ-1024 BXP 1GB 9p SD Class 2 Secure Digi…
3   2066119#12411100            "DISCO DURO INTERNO SOLIDO HDD SSD"@es
4    6656540#2639431   "Corsair Vengeance LED 32GB (2 x 16GB) DDR4 D…

                                  description_right  \
0   "Corsair Vengeance LPX Black 64GB (4x16GB) DD…
1   "Description:10 x 72GB 2.5-inch Serial Attach…
2   "Description:Genuine HPE 1GB FBD PC2-5300(2x5…
3   "DISCO DURO INTERNO SOLIDO HDD SSD KINGSTON V…
4   "Corsair Vengeance LED 32GB (2 x 16GB) DDR4 D…

                                       title_left  \
0   corsair vengeance lpx black 64gb 4x16gb ddr4 …
1   dh0072balwl hp 72 gb 3g 15k 2 5 dp sas null n…
2   sisk sdsdj 1024 bxp 1gb 9p sd class 2 secure …
3   disco duro interno solido hdd ssd kingston v3…
4   corsair vengeance led 32gb 2 x 16gb ddr4 dram…

                                      title_right  category_match
0   corsair vengeance lpx cmk64gx4m4a2666c16 prij…               1
1       dh0072balwl hp 72 gb 3g 15k 2 5 dp sas null               1
```

```
2     397409 b21 hp 1gb 2x512mb pc2 5300 sdram null                1
3   disco duro ssd kingston technology ssdnow v30…               1
4   corsair vengeance lpx 32gb 2 x 16gb ddr4 3000…                1
```

[38]:
```
df = engineer_features(df)
df = df[['title_left','title_right','levenshtein_distance','matching_numbers']]
df
```

[38]:
```
                                                title_left  \
0       corsair vengeance lpx black 64gb 4x16gb ddr4 …
1       dh0072balwl hp 72 gb 3g 15k 2 5 dp sas null n…
2       sisk sdsdj 1024 bxp 1gb 9p sd class 2 secure …
3       disco duro interno solido hdd ssd kingston v3…
4       corsair vengeance led 32gb 2 x 16gb ddr4 dram…
…                                                     …
68456   dg0300farvv hp 300 gb 6g 10k 2 5 dp sas null …
68457   samsung 840 evo 250gb 2 5 solid state drive d…
68458   socket h4 1151 coffee lake core i7 8700k 6 co…
68459   wd blue wd5000azlx hard drive 500 gb sata 6gb…
68460   acer ka ka240h 24 full hd tn negro pantalla p…

                                               title_right  \
0       corsair vengeance lpx cmk64gx4m4a2666c16 prij…
1           dh0072balwl hp 72 gb 3g 15k 2 5 dp sas null
2         397409 b21 hp 1gb 2x512mb pc2 5300 sdram null
3       disco duro ssd kingston technology ssdnow v30…
4       corsair vengeance lpx 32gb 2 x 16gb ddr4 3000…
…                                                     …
68456        null 512743 001 hp 72 gb 6g 15k 2 5 dp sas
68457                  ssd 750 evo 2 5 sata iii 120gb en
68458   7th generation intel core i7 7700 3 6ghz sock…
68459   m rock ships en us m rock ships new camera ba…
68460   xl2430 mon zowie 24 led panorami panorami tra…

       levenshtein_distance  matching_numbers
0                        78          0.800000
1                        61          1.000000
2                        50          0.250000
3                        73          0.333333
4                        45          1.000000
…                       …                …
68456                    84          0.300000
68457                    51          0.285714
68458                    95          0.400000
68459                    81          0.000000
68460                    80          0.200000
```

```
[65018 rows x 4 columns]
```

[39]:
```
df=df.sort_values(by=['levenshtein_distance'], ascending=[False])
df
```

[39]:
```
                                                 title_left  \
4547    alfa awus036h 1000mw 1w deluxe bundle 802 11b…
39756   alfa awus036h 1000mw 1w deluxe bundle 802 11b…
4878    alfa awus036h 1000mw 1w deluxe bundle 802 11b…
29409   alfa awus036h 1000mw 1w deluxe bundle 802 11b…
50020   alfa awus036h 1000mw 1w deluxe bundle 802 11b…
…                                                     …
61450   hyperx fury ddr3 4 gb dimm 240 pin hyperx 240…
3369    wd blue wd5000azlx hard drive 500 gb sata 6gb…
12416   apple smart keyboard for 10 5 inch ipad pro e…
17697   tp link 5 port fast ethernet desktop switch t…
49464   samsung en lambdatek internal solid state dri…

                                               title_right  \
4547                tp link tl wn881nd pci e n 300mbps es
39756           tp link tl wn951n prijzen nl tweakers nl
4878            tp link tl wn951n prijzen nl tweakers nl
29409    tarjeta pci express red wifi n n tradineur com
50020   apple 12 9 inch ipad pro with wi fi 32 gb sil…
…                                                     …
61450   hyperx fury ddr3 4 gb dimm 240 pin hyperx 240…
3369    wd blue wd5000azlx hard drive 500 gb sata 6gb…
12416   apple smart keyboard for 10 5 inch ipad pro e…
17697   tp link 5 port fast ethernet desktop switch t…
49464   samsung en lambdatek internal solid state dri…

       levenshtein_distance  matching_numbers
4547                    260          0.000000
39756                   258          0.000000
4878                    254          0.000000
29409                   247          0.000000
50020                   246          0.111111
…                         …                …
61450                     1          1.000000
3369                      1          1.000000
12416                     0          1.000000
17697                     0          1.000000
49464                     0          1.000000

[65018 rows x 4 columns]
```

9 Examining the mean ,count and max values of columns

[40]:
```
df.describe()
```

[40]:
```
       levenshtein_distance  matching_numbers
count          65018.000000      65018.000000
mean              66.196684          0.236587
std               22.984575          0.243334
min                0.000000          0.000000
25%               54.000000          0.000000
50%               67.000000          0.181818
75%               79.000000          0.333333
max              260.000000          1.000000
```

[41]:
```
df['match'] = np.where((df['levenshtein_distance']<260) &␣
 ↪(df['matching_numbers']>0.4),1,0)
df['match'].value_counts()
```

[41]:
```
0    53618
1    11400
Name: match, dtype: int64
```

[42]:
```
df
```

[42]:
```
                                              title_left  \
4547    alfa awus036h 1000mw 1w deluxe bundle 802 11b…
39756   alfa awus036h 1000mw 1w deluxe bundle 802 11b…
4878    alfa awus036h 1000mw 1w deluxe bundle 802 11b…
29409   alfa awus036h 1000mw 1w deluxe bundle 802 11b…
50020   alfa awus036h 1000mw 1w deluxe bundle 802 11b…
…                                                     …
61450   hyperx fury ddr3 4 gb dimm 240 pin hyperx 240…
3369    wd blue wd5000azlx hard drive 500 gb sata 6gb…
12416   apple smart keyboard for 10 5 inch ipad pro e…
17697   tp link 5 port fast ethernet desktop switch t…
49464   samsung en lambdatek internal solid state dri…

                                             title_right  \
4547                tp link tl wn881nd pci e n 300mbps es
39756            tp link tl wn951n prijzen nl tweakers nl
4878             tp link tl wn951n prijzen nl tweakers nl
29409     tarjeta pci express red wifi n n tradineur com
50020   apple 12 9 inch ipad pro with wi fi 32 gb sil…
…                                                     …
61450   hyperx fury ddr3 4 gb dimm 240 pin hyperx 240…
3369    wd blue wd5000azlx hard drive 500 gb sata 6gb…
12416   apple smart keyboard for 10 5 inch ipad pro e…
```

```
17697   tp link 5 port fast ethernet desktop switch t…
49464   samsung en lambdatek internal solid state dri…

       levenshtein_distance  matching_numbers  match
4547                    260          0.000000      0
39756                   258          0.000000      0
4878                    254          0.000000      0
29409                   247          0.000000      0
50020                   246          0.111111      0
…                       …              …     …
61450                     1          1.000000      1
3369                      1          1.000000      1
12416                     0          1.000000      1
17697                     0          1.000000      1
49464                     0          1.000000      1

[65018 rows x 5 columns]
```

10 Applying Validation set For Finding Model Accuracy

[43]:
```
data1 = pd.read_json('computers_gs.json',lines = True)
data1.to_csv('test.csv',index = False)
df_test = pd.read_csv('test.csv')
df_test['label'].value_counts()
```

[43]:
```
0    800
1    300
Name: label, dtype: int64
```

[44]:
```
df_test.columns
```

[44]:
```
Index(['id_left', 'category_left', 'cluster_id_left', 'id_right',
       'category_right', 'cluster_id_right', 'label', 'pair_id', 'brand_left',
       'brand_right', 'description_left', 'description_right',
       'keyValuePairs_left', 'keyValuePairs_right', 'price_left',
       'price_right', 'specTableContent_left', 'specTableContent_right',
       'title_left', 'title_right'],
      dtype='object')
```

[45]:
```
df_test.drop(df_test.columns[[2,5,8,9,12,13,14,15,16,17]],axis=1,inplace =True)
df_test
```

[45]:
```
       id_left             category_left  id_right  \
0       581109  Computers_and_Accessories  16637861
1      3083228  Computers_and_Accessories   3424944
2      5942105  Computers_and_Accessories    770253
```

```
3      1282014  Computers_and_Accessories  16999524
4      7969280  Computers_and_Accessories   6000979
…        …                          …       …
1095  10596400  Computers_and_Accessories  15781433
1096  10186131  Computers_and_Accessories  14419422
1097   2803663                Video_Games  15959011
1098   8856662  Computers_and_Accessories  16732972
1099  11774848  Computers_and_Accessories    713690

                category_right  label            pair_id  \
0     Computers_and_Accessories      0    581109#16637861
1     Computers_and_Accessories      1   3083228#3424944
2     Computers_and_Accessories      0    5942105#770253
3             Other_Electronics      0   1282014#16999524
4     Computers_and_Accessories      0    7969280#6000979
…                           …    …               …
1095  Computers_and_Accessories      1  10596400#15781433
1096  Computers_and_Accessories      1  10186131#14419422
1097                Video_Games      1   2803663#15959011
1098  Computers_and_Accessories      0   8856662#16732972
1099  Computers_and_Accessories      0    11774848#713690

                                      description_left  \
0     "GV-RX480G1 GAMING-4GD, Core Clock: 1202MHz, B…
1     "\n                More>>>\n                \n…
2     "Apple Mac mini - DTS - 1 x Core i5 2.8 GHz - …
3     "8 port Switch for adding more ports to your r…
4     "A drive USB flash Kingston DataTraveler® 100 …
…                                                   …
1095  "\n                More>>>\n                \n…
1096  "1506MHz Core, 1683MHz Boost, 8008MHz 256-bit …
1097  "DUAL-GTX1060-O6G, Core Clock: 1607MHz, Boost …
1098  "Intuitive closure uses a responsive flex zone…
1099  "Kingston DataTraveler 100 G3 - Unidad flash U…

                                     description_right  \
0     "GV-RX550GAMING OC-2GD, Boost: 1219MHz, Memory…
1     "\n\n                                  Every det…
2                                       "Null"@es
3     "This product is ENERGY STAR qualified for its…
4                                                NaN
…                                                  …
1095   "Description:StorageWorks 20/40GB Internal DL…
1096  "ASUS GeForce GTX 1070 ASUS Turbo 8GB GDDR5 VR…
1097  "1594MHz Core, 1809MHz Boost, 8008MHz 192-bit …
1098  "The Intel Core i3 processor is the perfect en…
1099  "Kingston DataTraveler SE9 - Unidad flash USB …
```

```
                                                  title_left  \
0       "Gigabyte Radeon RX 480 G1 Gaming 4096MB GDDR…
1       "Benq ZOWIE RL2455 24" Full HD TN Grey comput…
2       "Apple Mac mini 2.8GHz Intel Core i5"@es "Mac…
3       "TP-LINK 8-Port Fast Ethernet Desktop Switch …
4       "Pen Kingston DataTraveler 100 G3 32GB USB3.0…
…                                                     …
1095    "Hewlett Packard Enterprise SP/CQ Drive DLT 2…
1096    "ASUS GeForce GTX 1070 TURBO 8GB GDDR5 Graphi…
1097    "Asus GeForce GTX 1060 Dual OC 6144MB GDDR5 P…
1098    "Speck FlapTop notebook sleeve" " Speck sleev…
1099    "Kingston Technology DataTraveler 100 Generat…

                                                 title_right
0       "Gigabyte Radeon RX 550 Gaming OC 2048MB GDDR…
1       "Zowie RL2455 E-Sports 24" Full HD LED Monito…
2       "Mac Mini Qc I5 2.6ghz/8gb/1tb/iris Graphics"…
3       "Tripp Lite 750VA 450W UPS Eco Green Battery …
4       "Kingston DataTraveler 100 G3 128GB USB3.0"@e…
…                                                     …
1095    "340769-001 HP StorageWorks Internal DLT SCSI…
1096    "ASUS NVIDIA GeForce GTX 1070 Turbo 8GB Graph…
1097    "ASUS GeForce GTX 1060 DUAL OC 6GB GDDR5 Grap…
1098    "Intel Core i3 6300 / 3.8 GHz processor" " In…
1099    "Kingston Technology DataTraveler SE9 32GB US…

[1100 rows x 10 columns]
```

[46]:
```
df_test['title_right'].isnull().sum()
```

[46]:
```
0
```

[47]:
```
df_test.description_right.fillna(df_test.title_left, inplace = True)
df_test['description_right']
```

[47]:
```
0        "GV-RX550GAMING OC-2GD, Boost: 1219MHz, Memory…
1        "\n\n                                  Every det…
2                                                 "Null"@es
3       "This product is ENERGY STAR qualified for its…
4        "Pen Kingston DataTraveler 100 G3 32GB USB3.0…
                              …
1095     "Description:StorageWorks 20/40GB Internal DL…
1096    "ASUS GeForce GTX 1070 ASUS Turbo 8GB GDDR5 VR…
1097    "1594MHz Core, 1809MHz Boost, 8008MHz 192-bit …
1098    "The Intel Core i3 processor is the perfect en…
1099    "Kingston DataTraveler SE9 - Unidad flash USB …
```

```
Name: description_right, Length: 1100, dtype: object
```

[48]:
```
df_test['description_right'].isnull().sum()
```

[48]:
```
0
```

[49]:
```
df_test['description_left'].isnull().sum()
```

[49]:
```
78
```

[50]:
```
df_test.description_left.fillna(df_test.title_left, inplace = True)
df_test['title_left']
```

[50]:
```
0        "Gigabyte Radeon RX 480 G1 Gaming 4096MB GDDR…
1        "Benq ZOWIE RL2455 24" Full HD TN Grey comput…
2        "Apple Mac mini 2.8GHz Intel Core i5"@es "Mac…
3        "TP-LINK 8-Port Fast Ethernet Desktop Switch …
4        "Pen Kingston DataTraveler 100 G3 32GB USB3.0…
                              …
1095     "Hewlett Packard Enterprise SP/CQ Drive DLT 2…
1096     "ASUS GeForce GTX 1070 TURBO 8GB GDDR5 Graphi…
1097     "Asus GeForce GTX 1060 Dual OC 6144MB GDDR5 P…
1098     "Speck FlapTop notebook sleeve" " Speck sleev…
1099     "Kingston Technology DataTraveler 100 Generat…
Name: title_left, Length: 1100, dtype: object
```

[51]:
```
df_test['title_left'].isnull().sum()
```

[51]:
```
0
```

[52]:
```
df_test['category_match'] = np.where(df_test['category_left'] ==␣
 ↪df_test['category_right'],1,0)
df_test
```

[52]:

	id_left	category_left	id_right	\
0	581109	Computers_and_Accessories	16637861	
1	3083228	Computers_and_Accessories	3424944	
2	5942105	Computers_and_Accessories	770253	
3	1282014	Computers_and_Accessories	16999524	
4	7969280	Computers_and_Accessories	6000979	
…	…	…	…	
1095	10596400	Computers_and_Accessories	15781433	
1096	10186131	Computers_and_Accessories	14419422	
1097	2803663	Video_Games	15959011	
1098	8856662	Computers_and_Accessories	16732972	
1099	11774848	Computers_and_Accessories	713690	

```
                category_right  label            pair_id  \
0     Computers_and_Accessories      0    581109#16637861
1     Computers_and_Accessories      1    3083228#3424944
2     Computers_and_Accessories      0     5942105#770253
3             Other_Electronics      0   1282014#16999524
4     Computers_and_Accessories      0    7969280#6000979
…                           …    …               …
1095  Computers_and_Accessories      1  10596400#15781433
1096  Computers_and_Accessories      1  10186131#14419422
1097                Video_Games      1   2803663#15959011
1098  Computers_and_Accessories      0   8856662#16732972
1099  Computers_and_Accessories      0    11774848#713690

                                      description_left  \
0     "GV-RX480G1 GAMING-4GD, Core Clock: 1202MHz, B…
1     "\n                More>>>\n                \n…
2     "Apple Mac mini - DTS - 1 x Core i5 2.8 GHz - …
3     "8 port Switch for adding more ports to your r…
4     "A drive USB flash Kingston DataTraveler® 100 …
…                                                   …
1095  "\n                More>>>\n                \n…
1096  "1506MHz Core, 1683MHz Boost, 8008MHz 256-bit …
1097  "DUAL-GTX1060-O6G, Core Clock: 1607MHz, Boost …
1098  "Intuitive closure uses a responsive flex zone…
1099  "Kingston DataTraveler 100 G3 - Unidad flash U…

                                     description_right  \
0     "GV-RX550GAMING OC-2GD, Boost: 1219MHz, Memory…
1     "\n\n                                  Every det…
2                                       "Null"@es
3     "This product is ENERGY STAR qualified for its…
4      "Pen Kingston DataTraveler 100 G3 32GB USB3.0…
…                                                   …
1095   "Description:StorageWorks 20/40GB Internal DL…
1096  "ASUS GeForce GTX 1070 ASUS Turbo 8GB GDDR5 VR…
1097  "1594MHz Core, 1809MHz Boost, 8008MHz 192-bit …
1098  "The Intel Core i3 processor is the perfect en…
1099  "Kingston DataTraveler SE9 - Unidad flash USB …

                                            title_left  \
0      "Gigabyte Radeon RX 480 G1 Gaming 4096MB GDDR…
1      "Benq ZOWIE RL2455 24" Full HD TN Grey comput…
2      "Apple Mac mini 2.8GHz Intel Core i5"@es "Mac…
3      "TP-LINK 8-Port Fast Ethernet Desktop Switch …
4      "Pen Kingston DataTraveler 100 G3 32GB USB3.0…
…                                                   …
1095   "Hewlett Packard Enterprise SP/CQ Drive DLT 2…
```

```
1096   "ASUS GeForce GTX 1070 TURBO 8GB GDDR5 Graphi…
1097   "Asus GeForce GTX 1060 Dual OC 6144MB GDDR5 P…
1098   "Speck FlapTop notebook sleeve" " Speck sleev…
1099   "Kingston Technology DataTraveler 100 Generat…

                                          title_right  category_match
0      "Gigabyte Radeon RX 550 Gaming OC 2048MB GDDR…               1
1      "Zowie RL2455 E-Sports 24" Full HD LED Monito…               1
2      "Mac Mini Qc I5 2.6ghz/8gb/1tb/iris Graphics"…               1
3      "Tripp Lite 750VA 450W UPS Eco Green Battery …               0
4      "Kingston DataTraveler 100 G3 128GB USB3.0"@e…               1
…                                                    …             …
1095   "340769-001 HP StorageWorks Internal DLT SCSI…               1
1096   "ASUS NVIDIA GeForce GTX 1070 Turbo 8GB Graph…               1
1097   "ASUS GeForce GTX 1060 DUAL OC 6GB GDDR5 Grap…               1
1098   "Intel Core i3 6300 / 3.8 GHz processor" " In…               1
1099   "Kingston Technology DataTraveler SE9 32GB US…               1

[1100 rows x 11 columns]
```

[53]:
```
df_test['category_match'].value_counts()
```

[53]:
```
1    1063
0      37
Name: category_match, dtype: int64
```

[54]:
```
df_test.drop(df_test.index[df_test['category_match'] == 0],inplace = True)
```

[55]:
```
df_test['category_match'].value_counts()
```

[55]:
```
1    1063
Name: category_match, dtype: int64
```

[56]:
```
df_test = engineer_features(df_test)
df_test =␣
 ↪df_test[['title_left','title_right','levenshtein_distance','matching_numbers','label']]
df_test
```

[56]:
```
                                          title_left  \
0      "gigabyte radeon rx 480 g1 gaming 4096mb gddr…
1      "benq zowie rl2455 24" full hd tn grey comput…
2      "apple mac mini 2.8ghz intel core i5"@es "mac…
4      "pen kingston datatraveler 100 g3 32gb usb3.0…
5      "fire, 7" display, wi-fi, 8 gb - includes spe…
…                                                    …
1095   "hewlett packard enterprise sp/cq drive dlt 2…
1096   "asus geforce gtx 1070 turbo 8gb gddr5 graphi…
```

```
1097   "asus geforce gtx 1060 dual oc 6144mb gddr5 p…
1098   "speck flaptop notebook sleeve" " speck sleev…
1099   "kingston technology datatraveler 100 generat…

                                          title_right  levenshtein_distance  \
0      "gigabyte radeon rx 550 gaming oc 2048mb gddr…                    11
1      "zowie rl2455 e-sports 24" full hd led monito…                    89
2      "mac mini qc i5 2.6ghz/8gb/1tb/iris graphics"…                    73
4      "kingston datatraveler 100 g3 128gb usb3.0"@e…                    41
5      "apple ipad mini 4 wi-fi - tablet 128 gb 7.9"…                   118
…                                                  …                     …
1095   "340769-001 hp storageworks internal dlt scsi…                   101
1096   "asus nvidia geforce gtx 1070 turbo 8gb graph…                    36
1097   "asus geforce gtx 1060 dual oc 6gb gddr5 grap…                    41
1098   "intel core i3 6300 / 3.8 ghz processor" " in…                    81
1099   "kingston technology datatraveler se9 32gb us…                    23

      matching_numbers  label
0             0.166667      0
1             0.666667      1
2             0.750000      0
4             0.600000      0
5             0.111111      0
…                ….    …
1095          0.500000      1
1096          0.500000      1
1097          0.500000      1
1098          0.000000      0
1099          0.166667      0

[1063 rows x 5 columns]
```

[57]:
```
X_train=df[['levenshtein_distance','matching_numbers']]
X_test=df_test[['levenshtein_distance','matching_numbers']]
y_train=df['match']
y_test=df_test['label']
y_train
```

[57]:
```
4547     0
39756    0
4878     0
29409    0
50020    0
        ..
61450    1
3369     1
12416    1
```

```
17697    1
49464    1
Name: match, Length: 65018, dtype: int32
```

11 Creating Function For Finding Confusion Matrix

[58]:
```
def get_confusion_matrix_values(y_test, y_pred):
    cm = confusion_matrix(y_test, y_pred)
    return(cm[0][0], cm[0][1], cm[1][0], cm[1][1])
```

[ ]:
```

```

12 Model Building: Decision Tree,Random Forest and Support Vector Classifier

[59]:
```
from sklearn.svm import SVC
classifiers = {
    "DecisionTreeClassifier":DecisionTreeClassifier(criterion='gini',␣
 ↪splitter='best', max_depth=1, min_samples_split=4,random_state=42),
    "Support Vector Classifier":SVC(kernel='rbf', gamma=0.1),
    "RandomForestClassifier":
 ↪RandomForestClassifier(n_estimators=1000,max_depth=4,random_state=42,n_jobs=-1),

}

df_results = pd.DataFrame(columns=['model', 'accuracy', 'precision',
                                   'true_pos','false_pos',
                                      'true_neg','false_neg'])

for key in classifiers:

    classifier = classifiers[key]
    model = classifier.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred, zero_division=0)
    recall = recall_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred, zero_division=0)
    classification = classification_report(y_test, y_pred, zero_division=0)
    tp, fp, fn, tn = get_confusion_matrix_values(y_test, y_pred)

    row = {'model': key,
           'accuracy': accuracy,
           'precision': precision,
```

```
            'Recall': recall,
            'f1': f1,
            'true_pos': tp,
            'false_pos': fp,
            'true_neg': tn,
            'false_neg': fn,
           }
    df_results = df_results.append(row, ignore_index=True)

df_results.head(10)
```

[59]:

	model	accuracy	precision	true_pos	false_pos	true_neg	false_neg	Recall	f1
0	DecisionTreeClassifier	0.793979	0.605744	612	151	232	68	0.773333	0.679356
1	Support Vector Classifier	0.796802	0.622807	634	129	213	87	0.710000	0.663551
2	RandomForestClassifier	0.793979	0.605744	612	151	232	68	0.773333	0.679356

13 Estimating Results in Binary

[60]:
```
results = pd.DataFrame(data={'predictions': y_pred, 'actual': y_test})
results['result'] = np.where(results['predictions']==results['actual'], 1, 0)
results
```

[60]:

	predictions	actual	result
0	0	0	1
1	1	1	1
2	1	0	0
4	1	0	0
5	0	0	1
…	…	…	…
1095	1	1	1
1096	1	1	1
1097	1	1	1
1098	0	0	1
1099	0	0	1

[1063 rows x 3 columns]

14 Final Results with Match and Not Matched Classification

[61]:
```
results['predictions'].replace(0, 'Not match',inplace=True)
results['predictions'].replace(1, 'Match',inplace=True)

results['actual'].replace(0, 'Not Match',inplace=True)
results['actual'].replace(1, 'Match',inplace=True)

results['result'].replace(0, 'False',inplace=True)
results['result'].replace(1, 'True',inplace=True)
results
```

[61]:

	predictions	actual	result
0	Not match	Not Match	True
1	Match	Match	True
2	Match	Not Match	False
4	Match	Not Match	False
5	Not match	Not Match	True
…	…	…	…
1095	Match	Match	True
1096	Match	Match	True
1097	Match	Match	True
1098	Not match	Not Match	True
1099	Not match	Not Match	True

[1063 rows x 3 columns]

[ ]:

[ ]: