S1 - HOPE PCA_v3

November 15, 2020

0.1 Import data CSV.

[1]:
```
# pip install pymysql
from sqlalchemy import create_engine
import pymysql
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

[2]:
```
dfPCA = pd.read_csv('hope_dataset_cleaned.csv')
```

[3]:
```
dfPCA.head(10)
```

[3]:
```
   pedido.data.attributes.age pedido.data.attributes.diagnostic_main  \
0                        75.0                     FISTULA PERITONEAL
1                        75.0                     FISTULA PERITONEAL
2                        75.0                     FISTULA PERITONEAL
3                        75.0                     FISTULA PERITONEAL
4                        75.0                     FISTULA PERITONEAL
5                        75.0                     FISTULA PERITONEAL
6                        75.0                     FISTULA PERITONEAL
7                        75.0                     FISTULA PERITONEAL
8                        75.0                     FISTULA PERITONEAL
9                        75.0                     FISTULA PERITONEAL

  pedido.data.attributes.gender  articulo  respuesta.articlesRevisedYear  \
0                          male  27395425                           2018
1                          male  28560554                           2018
2                          male  28641726                           2017
3                          male  26245344                           2016
4                          male  28942543                           2018
5                          male  24782153                           2014
6                          male  28002229                           2018
7                          male  27505109                           2017
8                          male  24850546                           2015
9                          male  29371050                           2019

   respuesta.articlesRevisedMonth  \
```

```
0                                1
1                                4
2                               12
3                               12
4                                6
5                                6
6                                9
7                                4
8                                1
9                                4
```

```
                                respuesta.pubmed_keys  utilidad
0  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       1.0
1  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
2  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
3  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
4  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
5  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
6  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
7  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
8  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
9  Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       NaN
```

Get only data with 'utilidad' value defined

[4]:
```
dfPCA = dfPCA[pd.notnull(dfPCA['utilidad'])]

print(dfPCA.shape[0])
```

```
51
```

[5]:
```
dfPCA.head(10)
```

[5]:
```
     pedido.data.attributes.age pedido.data.attributes.diagnostic_main  \
0                          75.0                     FISTULA PERITONEAL
32                         75.0                     FISTULA PERITONEAL
230                        36.0           INSUFICIENCIA RESPIRATORIA
290                        51.0                       POLITRAUMATISMO
299                        51.0                       POLITRAUMATISMO
300                        18.0                         ABDOMEN AGUDO
303                        18.0                         ABDOMEN AGUDO
304                        18.0                         ABDOMEN AGUDO
305                        18.0                         ABDOMEN AGUDO
311                        76.0                           TORACOTOMIA

    pedido.data.attributes.gender  articulo  respuesta.articlesRevisedYear  \
0                            male  27395425                           2018
```

```
32                              male  28694230                              2017
230                             male  28805236                              2011
290                             male  27537587                              2011
299                             male  28148670                              2019
300                             male  25055513                              2019
303                             male  29279563                              2017
304                             male  29279563                              2017
305                             male  28065368                              2017
311                             male  30762794                              2019

     respuesta.articlesRevisedMonth  \
0                                  1
32                                12
230                                3
290                                3
299                                3
300                                3
303                                2
304                                2
305                               11
311                                3

                                   respuesta.pubmed_keys  utilidad
0    Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       1.0
32   Abdomen,Blood Culture,Catharsis,Diuresis,Drug …       1.0
230  Abdomen,Analgesics,Antitubercular Agents,Cipro…       0.0
290  Abdomen,Analgesics,Bone,Catharsis,Electroconvu…       0.0
299  Abdomen,Analgesics,Bone,Catharsis,Electroconvu…       1.0
300  Abdomen,Anti-Bacterial Agents,Diuresis,Operati…       1.0
303  Abdomen,Anti-Bacterial Agents,Diuresis,Operati…       0.0
304  Abdomen,Anti-Bacterial Agents,Diuresis,Operati…       0.0
305  Abdomen,Anti-Bacterial Agents,Diuresis,Operati…       1.0
311  Abdomen,Amiodarone,Analgesia,Angiodysplasia,Hy…       1.0
```

Remove “articulo” and “gender” to remove attributes without value

[6]:
```
dfPCA = dfPCA.drop([
    'pedido.data.attributes.gender',
    'articulo'
], axis=1)

dfPCA.head(10)
```

[6]:
```
     pedido.data.attributes.age pedido.data.attributes.diagnostic_main  \
0                          75.0                       FISTULA PERITONEAL
32                         75.0                       FISTULA PERITONEAL
230                        36.0              INSUFICIENCIA RESPIRATORIA
```

```
290                          51.0                          POLITRAUMATISMO
299                          51.0                          POLITRAUMATISMO
300                          18.0                            ABDOMEN AGUDO
303                          18.0                            ABDOMEN AGUDO
304                          18.0                            ABDOMEN AGUDO
305                          18.0                            ABDOMEN AGUDO
311                          76.0                              TORACOTOMIA

     respuesta.articlesRevisedYear  respuesta.articlesRevisedMonth  \
0                             2018                               1
32                            2017                              12
230                           2011                               3
290                           2011                               3
299                           2019                               3
300                           2019                               3
303                           2017                               2
304                           2017                               2
305                           2017                              11
311                           2019                               3

                                 respuesta.pubmed_keys  utilidad
0    Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…       1.0
32   Abdomen,Blood Culture,Catharsis,Diuresis,Drug …       1.0
230  Abdomen,Analgesics,Antitubercular Agents,Cipro…       0.0
290  Abdomen,Analgesics,Bone,Catharsis,Electroconvu…       0.0
299  Abdomen,Analgesics,Bone,Catharsis,Electroconvu…       1.0
300  Abdomen,Anti-Bacterial Agents,Diuresis,Operati…       1.0
303  Abdomen,Anti-Bacterial Agents,Diuresis,Operati…       0.0
304  Abdomen,Anti-Bacterial Agents,Diuresis,Operati…       0.0
305  Abdomen,Anti-Bacterial Agents,Diuresis,Operati…       1.0
311  Abdomen,Amiodarone,Analgesia,Angiodysplasia,Hy…       1.0
```

Expand pubmed_keys attribute

[7]:
```
dfPCA['respuesta.pubmed_keys'] = dfPCA['respuesta.pubmed_keys'].apply(lambda x :
 ↪ x.split(','))

dfPCA = dfPCA.explode('respuesta.pubmed_keys').reset_index(drop=True)

dfPCA.head(10)
```

[7]:
```
   pedido.data.attributes.age pedido.data.attributes.diagnostic_main  \
0                        75.0                     FISTULA PERITONEAL
1                        75.0                     FISTULA PERITONEAL
2                        75.0                     FISTULA PERITONEAL
3                        75.0                     FISTULA PERITONEAL
4                        75.0                     FISTULA PERITONEAL
```

```
5                          75.0                      FISTULA PERITONEAL
6                          75.0                      FISTULA PERITONEAL
7                          75.0                      FISTULA PERITONEAL
8                          75.0                      FISTULA PERITONEAL
9                          75.0                      FISTULA PERITONEAL

   respuesta.articlesRevisedYear  respuesta.articlesRevisedMonth  \
0                           2018                               1
1                           2018                               1
2                           2018                               1
3                           2018                               1
4                           2018                               1
5                           2018                               1
6                           2018                               1
7                           2018                               1
8                           2018                               1
9                           2018                               1

  respuesta.pubmed_keys  utilidad
0               Abdomen       1.0
1        Adenocarcinoma       1.0
2           Antiemetics       1.0
3         Blood Culture       1.0
4             Catharsis       1.0
5              Diuresis       1.0
6               Fistula       1.0
7           Gastrectomy       1.0
8     Incisional Hernia       1.0
9            Intestines       1.0
```

1 PCA

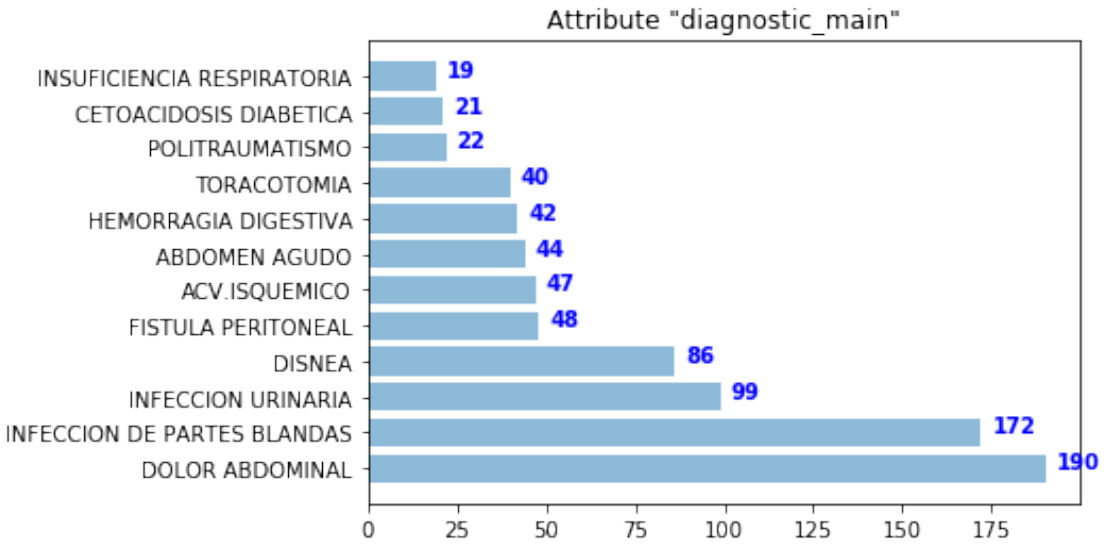
1.1 Transform (factorice) from Categories to continuous atributes

Transform 'pedido.data.attributes.diagnostic_main' atribute

[8]:
```
categoriesORGDiagnosticMain = dfPCA['pedido.data.attributes.diagnostic_main'].
↪value_counts()

print("total: " + str(categoriesORGDiagnosticMain.size))

y_values = np.arange(len(categoriesORGDiagnosticMain.index))

plt.barh(y_values, categoriesORGDiagnosticMain.values, align='center', alpha=0.
↪5)
plt.yticks(y_values, categoriesORGDiagnosticMain.index)
```

```
for i, v in enumerate(categoriesORGDiagnosticMain.values):
    plt.text(v + 3, i, str(v), color='blue', fontweight='bold')

plt.title('Attribute "diagnostic_main"')
plt.show()
```

```
total: 12
```



[9]:
```
dataDiagnosticMain, categoriesDiagnosticMain = pd.factorize(dfPCA['pedido.data.
↪attributes.diagnostic_main'])

categoriesDiagnosticMain
```

[9]:
```
Index(['FISTULA PERITONEAL', 'INSUFICIENCIA RESPIRATORIA', 'POLITRAUMATISMO',
       'ABDOMEN AGUDO', 'TORACOTOMIA', 'INFECCION DE PARTES BLANDAS',
       'DOLOR ABDOMINAL', 'INFECCION URINARIA', 'HEMORRAGIA DIGESTIVA',
       'ACV.ISQUEMICO', 'DISNEA', 'CETOACIDOSIS DIABETICA'],
      dtype='object')
```

0 => first element found => 'FISTULA PERITONEAL'

1 => second element found => 'INSUFICIENCIA RESPIRATORIA'

...

[10]:
```
dfPCA['pedido.data.attributes.diagnostic_main'] = dataDiagnosticMain

dfPCA.head(10)
```

[10]:

	pedido.data.attributes.age	pedido.data.attributes.diagnostic_main	\
0	75.0	0	
1	75.0	0	
2	75.0	0	
3	75.0	0	
4	75.0	0	
5	75.0	0	
6	75.0	0	
7	75.0	0	
8	75.0	0	
9	75.0	0	

	respuesta.articlesRevisedYear	respuesta.articlesRevisedMonth	\
0	2018	1	
1	2018	1	
2	2018	1	
3	2018	1	
4	2018	1	
5	2018	1	
6	2018	1	
7	2018	1	
8	2018	1	
9	2018	1	

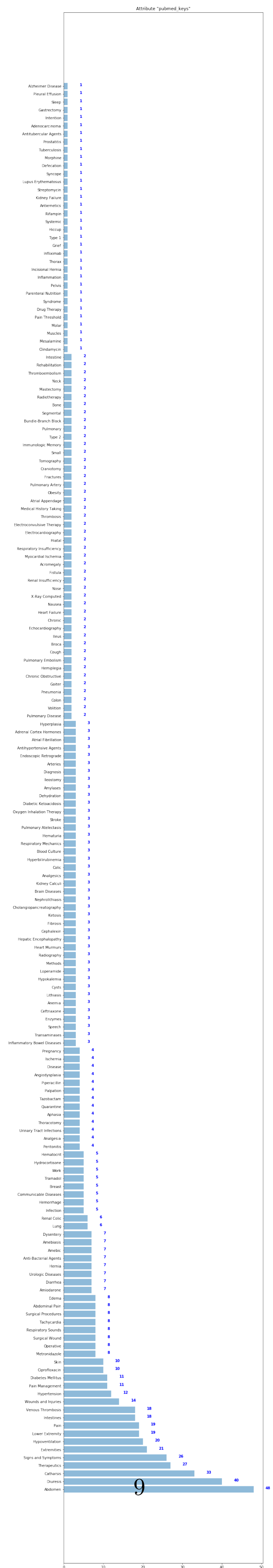
	respuesta.pubmed_keys	utilidad
0	Abdomen	1.0
1	Adenocarcinoma	1.0
2	Antiemetics	1.0
3	Blood Culture	1.0
4	Catharsis	1.0
5	Diuresis	1.0
6	Fistula	1.0
7	Gastrectomy	1.0
8	Incisional Hernia	1.0
9	Intestines	1.0

Transform 'respuesta.pubmed_keys' atribute

[11]:
```
categoriesORGPubMedKeys = dfPCA['respuesta.pubmed_keys'].value_counts()

print("total: " + str(categoriesORGPubMedKeys.size))

y_values = np.arange(len(categoriesORGPubMedKeys.index))

plt.figure(figsize=(10,80))
plt.barh(y_values, categoriesORGPubMedKeys.values, align='center', alpha=0.5)
plt.yticks(y_values, categoriesORGPubMedKeys.index)
```

```
for i, v in enumerate(categoriesORGPubMedKeys.values):
    plt.text(v + 3, i, str(v), color='blue', fontweight='bold', fontsize=10)

plt.title('Attribute "pubmed_keys"')

plt.show()
```

```
total: 177
```


Attribute "pubmed_keys"

[12]:
```
dataPubMedKeys, categoriesPubMedKeys = pd.factorize(dfPCA['respuesta.
↪pubmed_keys'])
```

[13]:
```
dfPCA['respuesta.pubmed_keys'] = dataPubMedKeys
```

[14]:
```
dfPCA.head(10)
```

[14]:

	pedido.data.attributes.age	pedido.data.attributes.diagnostic_main	respuesta.articlesRevisedYear	respuesta.articlesRevisedMonth	respuesta.pubmed_keys	utilidad
0	75.0	0	2018	1	0	1.0
1	75.0	0	2018	1	1	1.0
2	75.0	0	2018	1	2	1.0
3	75.0	0	2018	1	3	1.0
4	75.0	0	2018	1	4	1.0
5	75.0	0	2018	1	5	1.0
6	75.0	0	2018	1	6	1.0
7	75.0	0	2018	1	7	1.0
8	75.0	0	2018	1	8	1.0
9	75.0	0	2018	1	9	1.0

1.2 Standardize the Data

[15]:
```
from sklearn.preprocessing import StandardScaler

features = ['pedido.data.attributes.age',
            'pedido.data.attributes.diagnostic_main',
            'respuesta.articlesRevisedYear',
            'respuesta.articlesRevisedMonth',
            'respuesta.pubmed_keys']

# Separating out the features
x = dfPCA.loc[:, features].values# Separating out the target
dfPCA['utilidad']=pd.Categorical(dfPCA['utilidad'])
my_color=dfPCA['utilidad'].cat.codes

featuresTransformed = StandardScaler().fit_transform(x)

featuresTransformed
```

[15]:
```
array([[ 0.9570084 , -2.21979287,  0.46732162, -1.19652238, -1.28027939],
       [ 0.9570084 , -2.21979287,  0.46732162, -1.19652238, -1.25989395],
       [ 0.9570084 , -2.21979287,  0.46732162, -1.19652238, -1.2395085 ],
       …,
       [-1.66989784,  1.91795905,  0.82317202, -0.9329572 ,  2.30755826],
       [-1.66989784,  1.91795905,  0.82317202, -0.9329572 ,  2.30755826],
       [-1.66989784,  1.91795905,  0.82317202, -0.9329572 ,  0.63595208]])
```

[16]:
```
from sklearn.decomposition import PCA

pca = PCA(n_components=4)

pca.fit(featuresTransformed)

result=pd.DataFrame(pca.transform(featuresTransformed), columns=['PCA%i' % i␣
↪for i in range(4)])

result.head(10)
```

[16]:

	PCA0	PCA1	PCA2	PCA3
0	-2.181432	-1.351986	-0.924733	1.014503
1	-2.168172	-1.353110	-0.918248	1.010853
2	-2.154912	-1.354234	-0.911762	1.007204
3	-2.141652	-1.355358	-0.905276	1.003554
4	-2.128392	-1.356483	-0.898791	0.999905
5	-2.115132	-1.357607	-0.892305	0.996255
6	-2.101872	-1.358731	-0.885819	0.992606
7	-2.088612	-1.359855	-0.879334	0.988956

```
8 -2.075352 -1.360980 -0.872848  0.985307
9 -2.062092 -1.362104 -0.866362  0.981657
```

[17]:
```
print('explained variance ratio (first three components): %s' %
str(pca.explained_variance_ratio_))
print('sum of explained variance (first three components): %s' %
str(sum(pca.explained_variance_ratio_)))
```

```
explained variance ratio (first three components): [0.31205207 0.24763151
0.19472659 0.14857077]
sum of explained variance (first three components): 0.9029809474497172
```

[18]:
```
pd.DataFrame(pca.components_,columns=features,index = ['PC1','PC2','PC3','PC4'])
```

[18]:

	pedido.data.attributes.age	pedido.data.attributes.diagnostic_main	\
PC1	0.202186	0.688327	
PC2	-0.299028	-0.000864	
PC3	-0.908866	-0.006936	
PC4	-0.015928	-0.145721	

	respuesta.articlesRevisedYear	respuesta.articlesRevisedMonth	\
PC1	-0.236152	-0.080358	
PC2	-0.648047	0.698266	
PC3	0.153411	-0.221720	
PC4	-0.702451	-0.673071	

	respuesta.pubmed_keys
PC1	0.650462
PC2	-0.055149
PC3	0.318152
PC4	-0.179024

PC1 => diagnostic_main / pubmed_keys PC2 => articlesRevisedMonth PC3 => pubmed_keys PC4 => age

[ ]: