S1 - HOPE PCA

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
```

[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	FISTULA PERITONEAL	
1	75	FISTULA PERITONEAL	
2	36	INSUFICIENCIA RESPIRATORIA	
3	51	POLITRAUMATISMO	
4	51	POLITRAUMATISMO	
5	18	ABDOMEN AGUDO	
6	18	ABDOMEN AGUDO	
7	18	ABDOMEN AGUDO	
8	18	ABDOMEN AGUDO	
9	76	TORACOTOMIA	

	pedido.data.attributes.gender	\
0	male	
1	male	
2	male	
3	male	
4	male	
5	male	
6	male	
7	male	
8	male	
9	male	

	respuesta.pubmed_keys	articulo	utilidad
0	Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…	27395425	1.0

1	Abdomen,Blood Culture,Catharsis,Diuresis,Drug …	28694230	1.0
2	Abdomen,Analgesics,Antitubercular Agents,Cipro…	28805236	0.0
3	Abdomen,Analgesics,Bone,Catharsis,Electroconvu…	27537587	0.0
4	Abdomen,Analgesics,Bone,Catharsis,Electroconvu…	28148670	1.0
5	Abdomen,Anti-Bacterial Agents,Diuresis,Operati…	25055513	1.0
6	Abdomen,Anti-Bacterial Agents,Diuresis,Operati…	29279563	0.0
7	Abdomen,Anti-Bacterial Agents,Diuresis,Operati…	29279563	0.0
8	Abdomen,Anti-Bacterial Agents,Diuresis,Operati…	28065368	1.0
9	Abdomen,Amiodarone,Analgesia,Angiodysplasia,Hy…	30762794	1.0

1 PCA

1.1 Transform (factorice) from Categories to continuous atributes

Transform 'pedido.data.attributes.diagnostic_main' atribute

[4]:
```
categoriesORGDiagnosticMain = dfPCA['pedido.data.attributes.diagnostic_main'].
↪value_counts()

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

categoriesORGDiagnosticMain
```

```
total: 12
```

[4]:
```
DOLOR ABDOMINAL                13
INFECCION DE PARTES BLANDAS     9
INFECCION URINARIA              5
ABDOMEN AGUDO                   4
TORACOTOMIA                     4
CETOACIDOSIS DIABETICA          3
ACV.ISQUEMICO                   3
HEMORRAGIA DIGESTIVA            3
FISTULA PERITONEAL              2
POLITRAUMATISMO                 2
DISNEA                          2
INSUFICIENCIA RESPIRATORIA      1
Name: pedido.data.attributes.diagnostic_main, dtype: int64
```

[5]:
```
dataDiagnosticMain, categoriesDiagnosticMain = pd.factorize(dfPCA['pedido.data.
↪attributes.diagnostic_main'])

categoriesDiagnosticMain
```

[5]:
```
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'

...

[6]:
```
dfPCA['pedido.data.attributes.diagnostic_main'] = dataDiagnosticMain

dfPCA.head(10)
```

[6]:

	pedido.data.attributes.age	pedido.data.attributes.diagnostic_main	pedido.data.attributes.gender	respuesta.pubmed_keys	articulo	utilidad
0	75	0	male	Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…	27395425	1.0
1	75	0	male	Abdomen,Blood Culture,Catharsis,Diuresis,Drug …	28694230	1.0
2	36	1	male	Abdomen,Analgesics,Antitubercular Agents,Cipro…	28805236	0.0
3	51	2	male	Abdomen,Analgesics,Bone,Catharsis,Electroconvu…	27537587	0.0
4	51	2	male	Abdomen,Analgesics,Bone,Catharsis,Electroconvu…	28148670	1.0
5	18	3	male	Abdomen,Anti-Bacterial Agents,Diuresis,Operati…	25055513	1.0
6	18	3	male	Abdomen,Anti-Bacterial Agents,Diuresis,Operati…	29279563	0.0
7	18	3	male	Abdomen,Anti-Bacterial Agents,Diuresis,Operati…	29279563	0.0
8	18	3	male	Abdomen,Anti-Bacterial Agents,Diuresis,Operati…	28065368	1.0
9	76	4	male	Abdomen,Amiodarone,Analgesia,Angiodysplasia,Hy…	30762794	1.0

Transform ‘gender’ atribute

[7]:
```
categoriesORGGender = dfPCA['pedido.data.attributes.gender'].value_counts()

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

categoriesORGGender
```

```
total: 1
```

[7]:
```
male    51
Name: pedido.data.attributes.gender, dtype: int64
```

[8]:
```
dataGender, categoriesGender = pd.factorize(dfPCA['pedido.data.attributes.
↪gender'])

categoriesGender
```

[8]:
```
Index(['male'], dtype='object')
```

[9]:
```
dfPCA['pedido.data.attributes.gender'] = dataGender

dfPCA.head(10)
```

[9]:

	pedido.data.attributes.age	pedido.data.attributes.diagnostic_main	\
0	75	0	
1	75	0	
2	36	1	
3	51	2	
4	51	2	
5	18	3	
6	18	3	
7	18	3	
8	18	3	
9	76	4	

	pedido.data.attributes.gender	\
0	0	
1	0	
2	0	
3	0	
4	0	
5	0	
6	0	
7	0	
8	0	
9	0	

	respuesta.pubmed_keys	articulo	utilidad
0	Abdomen,Adenocarcinoma,Antiemetics,Blood Cultu…	27395425	1.0
1	Abdomen,Blood Culture,Catharsis,Diuresis,Drug …	28694230	1.0
2	Abdomen,Analgesics,Antitubercular Agents,Cipro…	28805236	0.0
3	Abdomen,Analgesics,Bone,Catharsis,Electroconvu…	27537587	0.0
4	Abdomen,Analgesics,Bone,Catharsis,Electroconvu…	28148670	1.0
5	Abdomen,Anti-Bacterial Agents,Diuresis,Operati…	25055513	1.0
6	Abdomen,Anti-Bacterial Agents,Diuresis,Operati…	29279563	0.0
7	Abdomen,Anti-Bacterial Agents,Diuresis,Operati…	29279563	0.0
8	Abdomen,Anti-Bacterial Agents,Diuresis,Operati…	28065368	1.0
9	Abdomen,Amiodarone,Analgesia,Angiodysplasia,Hy…	30762794	1.0

Transform 'respuesta.pubmed_keys' atribute

[10]:
```
categoriesORGPubMedKeys = dfPCA['respuesta.pubmed_keys'].value_counts()

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

categoriesORGPubMedKeys
```

```
total: 25
```

[10]:
```
Abdomen,Abdominal Pain,Amebiasis,Amebic,Catharsis,Ciprofloxacin,Diarrhea,Diuresi
s,Dysentery,Extremities,Hemorrhage,Hydrocortisone,Intestines,Lower
Extremity,Metronidazole,Pain,Pain Management,Signs and
Symptoms,Tachycardia,Therapeutics,Venous Thrombosis
5
Abdomen,Amiodarone,Analgesia,Angiodysplasia,Hypoventilation,Lung,Operative,Surgi
cal Procedures,Thoracotomy,Tramadol
4
Abdomen,Anti-Bacterial Agents,Diuresis,Operative,Peritonitis,Pregnancy,Signs and
Symptoms,Surgical Procedures,Surgical Wound,Therapeutics,Wounds and Injuries
4
Abdomen,Anti-Bacterial Agents,Antihypertensive Agents,Catharsis,Diuresis,Hernia,
Hypoventilation,Ileostomy,Intestines,Loperamide,Respiratory Mechanics,Surgical
Wound,Wounds and Injuries
3
Abdomen,Amiodarone,Anemia,Atrial
Fibrillation,Catharsis,Ceftriaxone,Cephalexin,Colic,Communicable Diseases,Diures
is,Edema,Enzymes,Hematocrit,Hematuria,Hypokalemia,Hypoventilation,Infection,Pulm
onary Atelectasis,Radiography,Renal Colic,Skin,Tachycardia,Therapeutics,Urinary
Tract Infections,Urologic Diseases,Work
3
Abdomen,Brain
Diseases,Catharsis,Dehydration,Diuresis,Extremities,Fibrosis,Hepatic
Encephalopathy,Lower Extremity,Quarantine,Signs and
Symptoms,Skin,Therapeutics,Venous Thrombosis
```

3
Abdomen,Abdominal
Pain,Amylases,Breast,Catharsis,Cholangiopancreatography,Diuresis,Endoscopic
Retrograde,Extremities,Hyperbilirubinemia,Hypertension,Hypoventilation,Lower
Extremity,Pain,Signs and Symptoms,Skin,Transaminases,Venous Thrombosis
3
Abdomen,Catharsis,Diuresis,Kidney
Calculi,Lithiasis,Methods,Nephrolithiasis,Pain,Pain Management,Renal
Colic,Urologic Diseases
3
Abdomen,Diuresis,Extremities,Hyperplasia,Hypertension,Hypoventilation,Ileus,Inte
stines,Ischemia,Lower Extremity,Pain,Pain Management,Respiratory Sounds,Signs
and Symptoms,Venous Thrombosis,Work,Wounds and Injuries
2
Abdomen,Adrenal Cortex Hormones,Amebiasis,Amebic,Catharsis,Ciprofloxacin,Colon,C
ysts,Diarrhea,Disease,Diuresis,Dysentery,Edema,Heart
Murmurs,Hematocrit,Inflammatory Bowel Diseases,Intestines,Metronidazole,Signs
and Symptoms,Speech,Therapeutics
2
Abdomen,Aphasia,Aphasia,Atrial Appendage,Broca,Catharsis,Diabetes Mellitus,Diure
sis,Hemiplegia,Hypertension,Neck,Obesity,Palpation,Rehabilitation,Respiratory
Sounds,Stroke
2
Abdomen,Analgesics,Bone,Catharsis,Electroconvulsive
Therapy,Extremities,Fractures,Immunologic Memory,Lung,Medical History
Taking,Signs and Symptoms
2
Abdomen,Acromegaly,Arteries,Breast,Bundle-Branch Block,Catharsis,Chronic
Obstructive,Cough,Craniotomy,Diabetes Mellitus,Diabetes Mellitus,Diuresis,Echoca
rdiography,Edema,Electrocardiography,Extremities,Goiter,Heart
Failure,Hypertension,Hypertension,Hypoventilation,Ischemia,Lower
Extremity,Mastectomy,Myocardial Ischemia,Nose,Oxygen Inhalation
Therapy,Piperacillin,Pulmonary,Pulmonary Artery,Pulmonary Disease,Pulmonary
Embolism,Radiotherapy,Respiratory Insufficiency,Respiratory
Sounds,Segmental,Tazobactam,Therapeutics,Thromboembolism,Thrombosis,Type
2,Venous Thrombosis,Volition 2
Abdomen,Diuresis,Extremities,Hernia,Hernia,Hiatal,Hypoventilation,Intestine,Inte
stines,Lower Extremity,Pain,Signs and Symptoms,Small,Venous Thrombosis,Wounds
and Injuries
2
Abdomen,Arteries,Catharsis,Diabetes
Mellitus,Diuresis,Extremities,Hypertension,Hypoventilation,Lower
Extremity,Oxygen Inhalation Therapy,Signs and
Symptoms,Speech,Stroke,Therapeutics,Venous Thrombosis
1
Abdomen,Clindamycin,Diuresis,Edema,Inflammation,Molar,Therapeutics,Tomography,X-
Ray Computed

```
1
Diabetes Mellitus,Diabetes Mellitus,Diabetic
Ketoacidosis,Diagnosis,Ketosis,Therapeutics,Type 1
1
Abdomen,Alzheimer Disease,Communicable
Diseases,Disease,Infection,Skin,Sleep,Urinary Tract Infections
1
Abdomen,Adrenal Cortex
Hormones,Catharsis,Ciprofloxacin,Cysts,Disease,Grief,Heart Murmurs,Inflammatory
Bowel Diseases,Infliximab,Intestines,Mesalamine,Metronidazole,Signs and
Symptoms,Syncope,Therapeutics
1
Abdomen,Blood Culture,Catharsis,Ciprofloxacin,Communicable Diseases,Diuresis,Hyp
erplasia,Infection,Pain,Pelvis,Prostatitis,Therapeutics,Tramadol
1
Abdomen,Blood Culture,Catharsis,Diuresis,Drug
Therapy,Extremities,Fistula,Hiccup,Intestines,Lower
Extremity,Morphine,Nausea,Pain,Palpation,Piperacillin,Pneumonia,Respiratory
Sounds,Tazobactam,Therapeutics,Wounds and Injuries
1
Diabetes Mellitus,Diabetic Ketoacidosis,Diagnosis,Ketosis
1
Abdomen,Adenocarcinoma,Antiemetics,Blood
Culture,Catharsis,Diuresis,Fistula,Gastrectomy,Incisional
Hernia,Intestines,Muscles,Nausea,Pain,Pain
Threshold,Palpation,Piperacillin,Pleural
Effusion,Pneumonia,Quarantine,Respiratory Sounds,Signs and Symptoms,Surgical
Wound,Tazobactam,Therapeutics,Thorax,Tomography,Wounds and Injuries,X-Ray
Computed
1
Abdomen,Analgesics,Antitubercular
Agents,Ciprofloxacin,Defecation,Diuresis,Intention,Intestines,Lupus
Erythematosus,Pain,Pain Management,Parenteral Nutrition,Rifampin,Streptomycin,Sy
ndrome,Systemic,Therapeutics,Tuberculosis,Wounds and Injuries
1
Chronic,Chronic,Diabetes Mellitus,Diabetic Ketoacidosis,Diagnosis,Ketosis,Kidney
Failure,Renal Insufficiency,Renal Insufficiency,Urologic Diseases
1
Name: respuesta.pubmed_keys, dtype: int64
```

[11]:
```
dataPubMedKeys, categoriesPubMedKeys = pd.factorize(dfPCA['respuesta.
↪pubmed_keys'])
```

[12]:
```
dfPCA['respuesta.pubmed_keys'] = dataPubMedKeys
```

[13]:
```
dfPCA.head(10)
```

[13]:

	pedido.data.attributes.age	pedido.data.attributes.diagnostic_main	pedido.data.attributes.gender	respuesta.pubmed_keys	articulo	utilidad
0	75	0	0	0	27395425	1.0
1	75	0	0	1	28694230	1.0
2	36	1	0	2	28805236	0.0
3	51	2	0	3	27537587	0.0
4	51	2	0	3	28148670	1.0
5	18	3	0	4	25055513	1.0
6	18	3	0	4	29279563	0.0
7	18	3	0	4	29279563	0.0
8	18	3	0	4	28065368	1.0
9	76	4	0	5	30762794	1.0

1.2 Standardize the Data

[14]:
```
from sklearn.preprocessing import StandardScaler

features = ['pedido.data.attributes.age',
            'pedido.data.attributes.diagnostic_main',
            'pedido.data.attributes.gender',
            'respuesta.pubmed_keys',
            'articulo']

# Separating out the features
x = dfPCA.loc[:, features].values# Separating out the target
#y = dfPCA.loc[:,['utilidad']].values# Standardizing the features
dfPCA['utilidad']=pd.Categorical(dfPCA['utilidad'])
my_color=dfPCA['utilidad'].cat.codes
#dfPCA = dfPCA.drop('utilidad', 1)

featuresTransformed = StandardScaler().fit_transform(x)

featuresTransformed
```

[14]:
```
array([[ 0.91709628, -2.24479066,  0.        , -1.81819247, -0.19317962],
       [ 0.91709628, -2.24479066,  0.        , -1.65608091,  0.31397864],
```

```
[-0.8747549 , -1.85670821,  0.        , -1.49396934,  0.35732434],
[-0.18558137, -1.46862576,  0.        , -1.33185777, -0.13766811],
[-0.18558137, -1.46862576,  0.        , -1.33185777,  0.100948  ],
[-1.70176313, -1.0805433 ,  0.        , -1.16974621, -1.10687007],
[-1.70176313, -1.0805433 ,  0.        , -1.16974621,  0.54253988],
[-1.70176313, -1.0805433 ,  0.        , -1.16974621,  0.54253988],
[-1.70176313, -1.0805433 ,  0.        , -1.16974621,  0.06842018],
[ 0.96304118, -0.69246085,  0.        , -1.00763464,  1.12171293],
[ 0.96304118, -0.69246085,  0.        , -1.00763464,  1.12171293],
[ 0.96304118, -0.69246085,  0.        , -1.00763464,  0.66295943],
[ 0.96304118, -0.69246085,  0.        , -1.00763464, -1.74384989],
[-1.19636921, -0.30437839,  0.        , -0.84552307,  0.41640915],
[-1.19636921, -0.30437839,  0.        , -0.84552307,  0.32427757],
[ 1.10087588,  0.08370406,  0.        , -0.68341151, -0.59661724],
[ 1.10087588,  0.08370406,  0.        , -0.68341151, -0.59661724],
[ 1.19276569,  0.47178651,  0.        , -0.52129994,  0.25862449],
[ 0.87115137,  0.08370406,  0.        , -0.35918837,  1.08237552],
[ 0.87115137,  0.08370406,  0.        , -0.35918837,  1.10153914],
[ 0.87115137,  0.08370406,  0.        , -0.35918837,  0.15789492],
[ 0.45764726,  0.85986897,  0.        , -0.19707681,  0.21066839],
[ 0.45764726,  0.85986897,  0.        , -0.19707681,  0.6678502 ],
[ 0.45764726,  0.85986897,  0.        , -0.19707681,  1.17931899],
[ 0.41170236,  1.24795142,  0.        , -0.03496524, -0.74218844],
[ 0.41170236,  1.24795142,  0.        , -0.03496524, -0.57050462],
[ 1.10087588,  0.08370406,  0.        ,  0.12714633, -0.77544256],
[ 1.10087588,  0.08370406,  0.        ,  0.12714633, -1.31346715],
[ 1.10087588,  0.08370406,  0.        ,  0.12714633,  0.11654809],
[-1.19636921, -0.30437839,  0.        ,  0.28925789, -2.6686866 ],
[-1.19636921, -0.30437839,  0.        ,  0.45136946,  0.89516893],
[-1.19636921, -0.30437839,  0.        ,  0.45136946,  0.51002923],
[-1.19636921, -0.30437839,  0.        ,  0.45136946, -1.57727317],
[-1.19636921, -0.30437839,  0.        ,  0.45136946,  0.41640915],
[-1.19636921, -0.30437839,  0.        ,  0.45136946, -1.10097381],
[ 0.22792275,  0.47178651,  0.        ,  0.61348103, -0.16001844],
[ 0.22792275,  0.47178651,  0.        ,  0.61348103,  0.79928405],
[ 0.22792275,  0.47178651,  0.        ,  0.61348103,  0.14857688],
[ 1.10087588,  0.08370406,  0.        ,  0.77559259,  0.98519776],
[ 1.10087588,  0.08370406,  0.        ,  0.77559259, -2.45688246],
[ 0.31981255,  0.08370406,  0.        ,  0.93770416, -1.61222666],
[ 0.31981255,  0.08370406,  0.        ,  0.93770416,  0.81440854],
[ 0.31981255,  0.08370406,  0.        ,  0.93770416,  0.98519776],
[ 0.41170236,  1.24795142,  0.        ,  1.09981573,  1.04535678],
[ 0.45764726,  0.47178651,  0.        ,  1.26192729,  0.35655197],
[ 0.45764726,  1.63603387,  0.        ,  1.42403886,  0.24450666],
[ 0.45764726,  1.63603387,  0.        ,  1.42403886, -1.68007619],
[-0.64503039, -0.30437839,  0.        ,  1.58615043, -1.48790854],
[-1.51798352,  2.02411633,  0.        ,  1.74826199,  1.4155516 ],
```

```
       [-1.51798352,  2.02411633,  0.        ,  1.91037356,  1.4155516 ],
       [-1.51798352,  2.02411633,  0.        ,  2.07248513,  0.14101718]])
```

[15]:
```
from sklearn.decomposition import PCA

pca = PCA(n_components=3)

pca.fit(featuresTransformed)

result=pd.DataFrame(pca.transform(featuresTransformed), columns=['PCA%i' % i␣
 ↪for i in range(3)])

result.head(10)
```

[15]:

	PCA0	PCA1	PCA2
0	-2.921363	0.605486	-0.397869
1	-2.780250	0.283468	-0.756023
2	-2.300288	-1.074338	0.313487
3	-1.972936	-0.198337	0.223268
4	-1.960351	-0.352165	0.043905
5	-1.560182	-0.693746	1.880072
6	-1.473186	-1.757068	0.640240
7	-1.473186	-1.757068	0.640240
8	-1.498193	-1.451418	0.996627
9	-1.188089	-0.071128	-1.503128

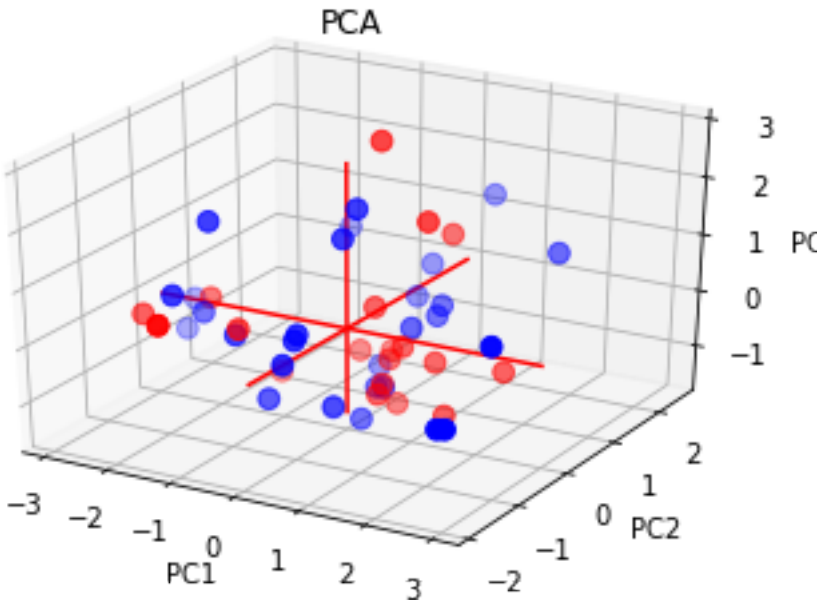
[16]:
```
print('explained variance ratio (first three components): %s' %
str(pca.explained_variance_ratio_))
print('sum of explained variance (first three components): %s' %
str(sum(pca.explained_variance_ratio_)))
```

```
explained variance ratio (first three components): [0.44726263 0.25669502
0.25009763]
sum of explained variance (first three components): 0.9540552760689917
```

[17]:
```
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D # note: remove when fix this issue =>␣
 ↪https://github.com/matplotlib/matplotlib/issues/16192

fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
ax.scatter(result['PCA0'], result['PCA1'], result['PCA2'], c=my_color.
 ↪replace([0,1],['r','b']), cmap="Set2_r", s=60)
#ax.scatter(result['PCA0'], result['PCA1'], result['PCA2'], c=my_color,␣
 ↪cmap="Set2_r", s=60)

# make simple, bare axis lines through space:
```

```
xAxisLine = ((min(result['PCA0']), max(result['PCA0'])), (0, 0), (0,0))
ax.plot(xAxisLine[0], xAxisLine[1], xAxisLine[2], 'r')
yAxisLine = ((0, 0), (min(result['PCA1']), max(result['PCA1'])), (0,0))
ax.plot(yAxisLine[0], yAxisLine[1], yAxisLine[2], 'r')
zAxisLine = ((0, 0), (0,0), (min(result['PCA2']), max(result['PCA2'])))
ax.plot(zAxisLine[0], zAxisLine[1], zAxisLine[2], 'r')

# label the axes
ax.set_xlabel("PC1")
ax.set_ylabel("PC2")
ax.set_zlabel("PC3")
ax.set_title("PCA")
```

[17]: Text(0.5, 0.92, 'PCA')



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

[18]:

	pedido.data.attributes.age	pedido.data.attributes.diagnostic_main	\
PC1	-0.050066	0.705007	
PC2	0.760475	0.071805	
PC3	-0.630542	-0.130930	

	pedido.data.attributes.gender	respuesta.pubmed_keys	articulo
PC1	-1.110223e-16	0.705462	0.052744
PC2	-1.318390e-16	0.030411	-0.644668
PC3	-1.110223e-16	0.142297	-0.751682

[ ]: