In [6]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.preprocessing import OrdinalEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

from sklearn.model_selection import train_test_split
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score


#Reads CSV to data frame, sets case order to index
med_dirty= pd.read_csv('/Users/herlihpj/Desktop/Data Analytics/D212 Data Mining II/Task 2/medical_clean.csv',
                       index_col=0)

#Display all columns in the dataset
print(med_dirty.columns)

# Dropping columns not relavant to the analysis
med_mine = med_dirty.drop(columns= ["Customer_id", "Interaction","UID", "City", 'State', "County", "Zip", "Lat","Lng",
                                    'Area', 'TimeZone','Job', 'Marital','Gender','Initial_admin', 'Complication_risk',
                                   'Services', 'Item1','Item2','Item3','Item4','Item5','Item6','Item7','Item8'])
#'Population',

#Ordinal Encoding to convert to numeric 0:No, 1:Yes; other variable alphabetically starting with 0
oe_dict={}
#List of Categorical Yes/No
categorical=['ReAdmis','Soft_drink','HighBlood','Stroke','Overweight','Arthritis','Diabetes','Hyperlipidemia','BackPain',
             'Anxiety', 'Allergic_rhinitis', 'Reflux_esophagitis', 'Asthma']
for col_name in categorical:
    #print(col_name+' pre: '+str(med_mine[col_name].unique()))
    #Creates column ordinal encoder
    oe_dict[col_name]=OrdinalEncoder()
    col=med_mine[col_name]
    #select non-null values of col
    col_not_null=col[col.notnull()]
    reshaped_vals=col_not_null.values.reshape(-1,1)
    encoded_vals=oe_dict[col_name].fit_transform(reshaped_vals)
    med_mine.loc[col.notnull(), col_name]=np.squeeze(encoded_vals)
    #print(col_name+' post: '+str(med_mine[col_name].unique()))
```

```
#Create an unlinked copy of the dataset to be used in a final model pipeline
med_pipe=med_mine.copy(deep=True)

#List of Continuous numerical features
numeric_cols=['Population','Children','Age','Income','VitD_levels','Doc_visits', 'Full_meals_eaten','vitD_supp',
              'Initial_days','TotalCharge','Additional_charges']
#Standardize the Numerical data
scaler = StandardScaler()
scaler.fit(med_mine[numeric_cols])
med_mine[numeric_cols] = scaler.transform(med_mine[numeric_cols])

#Export the cleaned dataset to csv
med_mine.to_csv('/Users/herlihpj/Desktop/Data Analytics/D212 Data Mining II/Task 2/medical_prepared.csv')
print('Prepared Data has been exported to CSV')

print('====================== \n Data has been prepared     \n====================== ')
SEED=7
#Store the number of columns in DF
num_cols=len(med_mine.columns)
#Create list of all PC's
pc_list=[]
for pc in range(1,num_cols+1):
    component='PC'+str(pc)
    pc_list.append(component)

pca_full=PCA(n_components=num_cols, random_state=SEED)
pca_full.fit(med_mine)
pca=pca_full.transform(med_mine)

#Show each rows contribution to each PC
pc_df=pd.DataFrame(pca, columns=pc_list)
print(pc_df)

#Creates Tables with PC's as columns and replaces index #'s with the feature labels :) great visualization
Load=pd.DataFrame(pca_full.components_.T, columns=[pc_list],index=med_mine.columns)
print(Load)
Load.to_csv('/Users/herlihpj/Desktop/Data Analytics/D212 Data Mining II/Task 2/PCA_matrix.csv')


print('Variance explained by all 24 components: ', sum(pca_full.explained_variance_ratio_*100))

#Display variance captured with each PC
var_full=pca_full.explained_variance_ratio_*100
#Calculate cumulative sum at each component
```

```
cumsum_full=np.cumsum(pca_full.explained_variance_ratio_*100)
print(cumsum_full)

#elbow Plot
var=pca_full.explained_variance_ratio_
plt.plot(var)
plt.title('Elbow PLot')
plt.xlabel('Principal component index')
plt.ylabel('Explained variance ratio')
plt.show()

#Create dataframe with each components variance ratio and cumulative sum of variance
var_full_df=pd.DataFrame(var_full.round(2), columns=['Captured Variance per PC'], index=pc_list)
var_full_df['Cumulative Sum']=cumsum_full
print(var_full_df)
#you can see first 18 PC's capture 95.4 of the variance in the dataset

#Creating a Skree PLot
plt.plot(np.cumsum(pca_full.explained_variance_ratio_*100))
plt.xlabel('Number of Components')
plt.ylabel('Explained Variance')
plt.show()

print('====================== \n Final Pipeline     \n====================== ')

#Optimal compnents is 10, from elbow plot
y = med_pipe['ReAdmis']
X = med_pipe.drop('ReAdmis', axis=1)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)


#PCA in a model pipeline
pipe = Pipeline([('scaler', StandardScaler()),
                 ('reducer', PCA(n_components=10,random_state=SEED)),
                 ('classifier', LogisticRegression())])
pipe.fit(X_train, y_train)

#print('Remaining components at 95% variance: ',len(pipe['reducer'].components_))
print(pipe['reducer'].explained_variance_ratio_)
print('Total model explained variance: ',pipe['reducer'].explained_variance_ratio_.sum().round(4))

print('Model Accuracy: ',pipe.score(X_test, y_test).round(4)*100,'%')
```

```
Index(['Customer_id', 'Interaction', 'UID', 'City', 'State', 'County', 'Zip',
       'Lat', 'Lng', 'Population', 'Area', 'TimeZone', 'Job', 'Children',
       'Age', 'Income', 'Marital', 'Gender', 'ReAdmis', 'VitD_levels',
       'Doc_visits', 'Full_meals_eaten', 'vitD_supp', 'Soft_drink',
       'Initial_admin', 'HighBlood', 'Stroke', 'Complication_risk',
       'Overweight', 'Arthritis', 'Diabetes', 'Hyperlipidemia', 'BackPain',
       'Anxiety', 'Allergic_rhinitis', 'Reflux_esophagitis', 'Asthma',
       'Services', 'Initial_days', 'TotalCharge', 'Additional_charges',
       'Item1', 'Item2', 'Item3', 'Item4', 'Item5', 'Item6', 'Item7', 'Item8'],
      dtype='object')
Prepared Data has been exported to CSV
======================
 Data has been prepared
======================
           PC1       PC2       PC3       PC4       PC5       PC6       PC7  \
0    -1.193877  0.696329  0.737259  0.785957  0.813725 -1.242777 -0.338144
1    -0.892455  0.625842 -0.464297 -0.012741 -0.022356  0.744206  0.967847
2    -1.688348  0.704117 -0.550308 -0.349948 -0.194175  1.263739 -0.791718
3    -1.911924  0.870126 -0.291960 -1.470806  0.160836 -0.336030 -0.679320
4    -2.052326 -1.910740  1.178325 -0.984226 -1.961989  0.597876  1.829450
...        ...       ...       ...       ...       ...       ...       ...
9995  0.696718 -1.338772 -0.372986 -0.877046  0.144141  0.103242  1.025211
9996  2.079594  2.738303 -0.291452 -0.171707  0.215194  0.738029 -0.347151
9997  1.970479 -0.099370 -0.989311 -0.551975  1.434169 -0.180250  0.031636
9998  1.559000 -1.014453 -1.053510  0.435158 -0.513583  0.113891  1.446193
9999  2.096816  0.144923  1.752908  1.759092  1.027891  2.482973  0.390884

           PC8       PC9      PC10  ...      PC15      PC16      PC17  \
0     0.403689 -1.398279 -0.737347  ...  0.103167  0.543193  0.155717
1     1.060478  0.678821 -0.783888  ... -0.564041 -0.323660 -0.093109
2    -0.142993  0.164692 -0.697448  ... -0.329721 -0.344662  0.065216
3     0.328929 -0.124542  0.977946  ... -0.352595 -0.233208  0.144569
4    -0.571147 -0.182156 -0.033597  ...  0.580975  0.242894 -0.533297
...        ...       ...       ...  ...       ...       ...       ...
9995  0.706845  0.927508 -0.947160  ... -0.670490  0.795544 -0.644859
9996 -1.129981 -1.052621 -0.487507  ... -0.053439 -0.486279  0.858675
9997  0.667297  0.446101 -0.810005  ... -0.431201  0.602577 -0.116944
9998  0.242195  0.113374  0.304409  ... -0.324061 -0.585368 -0.233326
9999  0.043812  0.460182  0.890653  ...  0.832262 -0.414146 -0.075967

          PC18      PC19      PC20      PC21      PC22      PC23      PC24
0     1.066278 -0.017856 -0.522134 -0.191030 -0.034688  0.008611  0.044709
1    -0.425538  0.034471 -0.099902 -0.260909  0.122537 -0.022276  0.165202
2     0.087252  0.731481 -0.402396 -0.282212 -0.126931  0.095530 -0.096447
3     0.724575 -0.785377 -0.069656  0.804406 -0.205594  0.181388 -0.050701
```

```
4     0.397863 -0.219799  0.844764 -0.171619 -0.197274 -0.120183 -0.160600
...        ...       ...       ...       ...       ...       ...       ...
9995  0.066193 -0.216153 -0.224561 -0.213926  0.616108  0.230047  0.068878
9996  0.504102  0.155623 -0.369938 -0.243338 -0.121589 -0.153858 -0.180973
9997 -0.524393  0.261296  0.789954 -0.204421 -0.055886  0.099471  0.019882
9998 -0.338000 -0.006555 -0.181474 -0.232214 -0.139114 -0.106991  0.010843
9999 -0.501934 -0.239424 -0.055198 -0.121785 -0.085260  0.126258 -0.102757

[10000 rows x 24 columns]
                        PC1       PC2       PC3       PC4       PC5  \
Population          0.022907 -0.024479  0.422490  0.372442 -0.227873
Children            0.030695  0.017855 -0.099209  0.343142  0.524892
Age                 0.058544  0.678095  0.020128 -0.019828  0.007180
Income             -0.017428 -0.018799  0.323030  0.195566  0.656085
ReAdmis             0.285376 -0.024541 -0.007675  0.002876  0.001549
VitD_levels        -0.001623  0.019598 -0.359890  0.570242 -0.328691
Doc_visits         -0.005824  0.015153  0.210468  0.598067 -0.107612
Full_meals_eaten   -0.018767  0.030988 -0.603945  0.128280  0.028863
vitD_supp           0.021713  0.015259  0.408048 -0.071463 -0.347634
Soft_drink          0.001251 -0.000747 -0.008579  0.009052  0.005526
HighBlood           0.012710  0.153080  0.007780  0.006273 -0.000762
Stroke             -0.000169  0.008616 -0.000800  0.001740  0.000608
Overweight         -0.003912  0.002074 -0.000279  0.002963 -0.013262
Arthritis           0.009054  0.001899 -0.009130  0.001979  0.003412
Diabetes            0.001379  0.001157 -0.005364 -0.001562  0.011304
Hyperlipidemia      0.002378 -0.000397  0.002097 -0.015003  0.004896
BackPain            0.010149  0.006674  0.012437 -0.000602  0.002365
Anxiety             0.007268  0.003540 -0.009054  0.001004  0.003432
Allergic_rhinitis   0.004019  0.006304  0.001206 -0.002648 -0.008835
Reflux_esophagitis  0.006438 -0.008004  0.007218 -0.000197  0.011391
Asthma             -0.004682  0.005423 -0.010664 -0.001915  0.007764
Initial_days        0.673644 -0.063665 -0.015212 -0.006497  0.005186
TotalCharge         0.674170 -0.051919 -0.016289 -0.003331  0.003659
Additional_charges  0.060806  0.711372  0.024992 -0.006303  0.008094

                        PC6       PC7       PC8       PC9      PC10  ...  \
Population          0.347549 -0.484309  0.224645  0.472293  0.024919  ...
Children            0.626540  0.276527 -0.350002  0.060118  0.000810  ...
Age                -0.003181 -0.019683  0.005198 -0.020399  0.622188  ...
Income             -0.326957  0.087129  0.554687 -0.061121  0.004504  ...
ReAdmis            -0.012014 -0.002901  0.003048 -0.001563  0.002801  ...
VitD_levels         0.140884  0.095441  0.412546 -0.487076 -0.001918  ...
Doc_visits         -0.541076  0.081382 -0.534415  0.024754  0.003854  ...
Full_meals_eaten   -0.227283  0.170414  0.189800  0.706511  0.016334  ...
vitD_supp           0.116090  0.796786  0.162713  0.177583  0.000232  ...
```

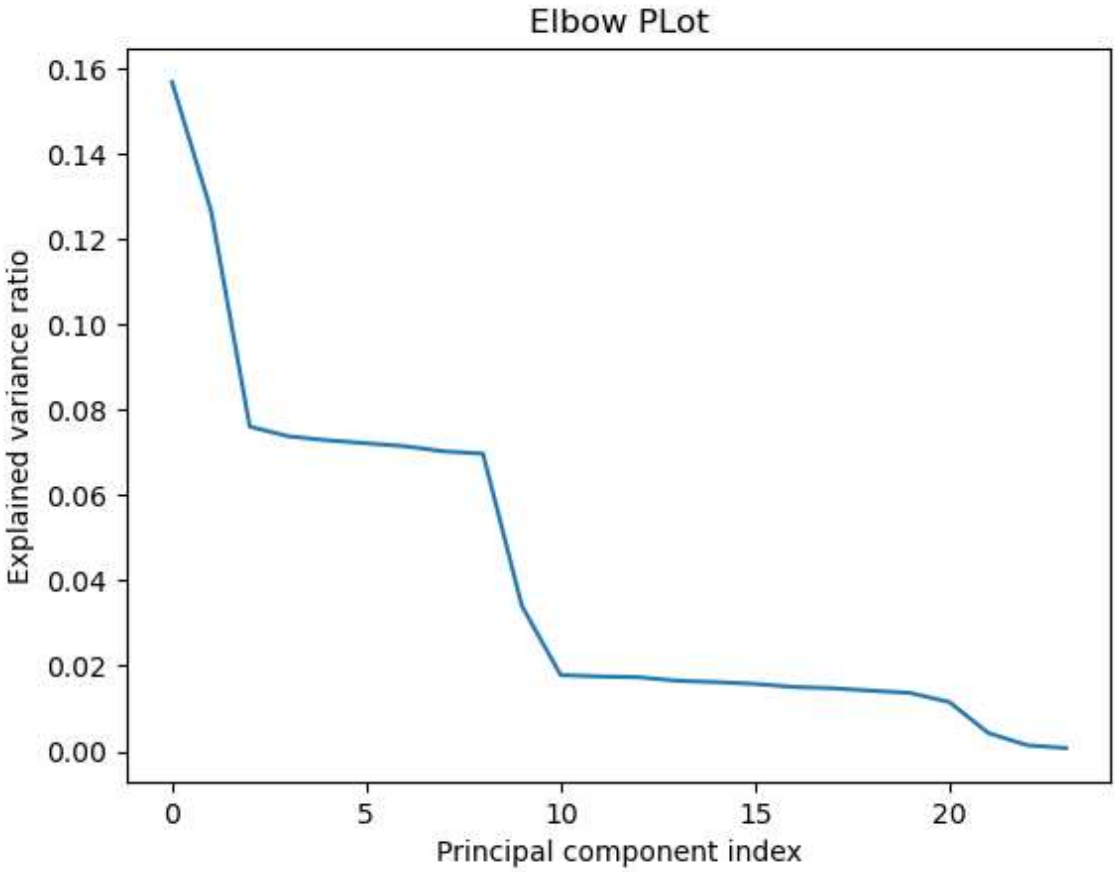
Soft_drink	-0.004183	-0.002807	-0.001685	0.006443	0.004261	...
HighBlood	-0.001074	-0.010626	0.007699	0.012580	-0.633259	...
Stroke	0.001959	0.002797	0.002342	-0.000545	-0.013898	...
Overweight	-0.002361	-0.006532	-0.005622	-0.005176	-0.032828	...
Arthritis	0.000960	-0.004534	-0.003184	0.004323	-0.000959	...
Diabetes	-0.000471	-0.001310	-0.019482	0.008264	0.004584	...
Hyperlipidemia	0.004201	0.006136	0.010076	0.002635	0.011666	...
BackPain	-0.006584	-0.006418	-0.000598	-0.003547	0.010488	...
Anxiety	-0.000617	0.006215	-0.000149	-0.002695	-0.008387	...
Allergic_rhinitis	-0.009168	-0.000308	0.005862	0.011612	-0.007253	...
Reflux_esophagitis	0.001574	-0.006332	0.004351	0.008083	-0.015009	...
Asthma	0.002764	-0.008641	0.007349	0.000729	-0.004550	...
Initial_days	-0.029412	-0.004085	0.007969	-0.004494	0.019461	...
TotalCharge	-0.028368	-0.003914	0.007935	-0.001563	-0.022017	...
Additional_charges	0.000917	-0.029857	0.014199	0.007635	-0.456226	...

	PC15	PC16	PC17	PC18	PC19	\
Population	0.002796	0.008728	0.000415	0.001312	0.002760	
Children	0.000556	-0.006804	-0.001419	-0.011767	-0.008245	
Age	-0.009964	-0.002746	0.008677	-0.018579	0.002978	
Income	-0.002558	-0.005027	0.004046	-0.005972	0.008664	
ReAdmis	-0.010667	-0.017480	-0.002390	-0.003659	-0.004080	
VitD_levels	0.008565	-0.000570	-0.000818	0.010009	0.010021	
Doc_visits	0.010700	0.006611	0.003641	0.001718	-0.016156	
Full_meals_eaten	-0.003360	-0.008528	-0.005172	-0.009040	-0.001622	
vitD_supp	-0.003499	-0.001038	0.013757	0.008607	0.004299	
Soft_drink	0.087958	0.145929	0.094489	0.140292	0.240811	
HighBlood	0.010654	-0.003684	-0.018425	0.014302	-0.007089	
Stroke	-0.052409	-0.034553	0.006019	0.038911	0.030101	
Overweight	-0.005799	-0.245936	0.546658	-0.735758	0.295449	
Arthritis	0.373971	-0.234526	-0.011209	0.063163	-0.068504	
Diabetes	0.116545	0.062825	0.258506	0.470604	0.768652	
Hyperlipidemia	0.882883	0.214405	0.060356	-0.118154	-0.127029	
BackPain	0.145360	-0.149621	-0.110273	0.061068	0.031969	
Anxiety	-0.099147	0.875694	-0.020637	-0.245726	0.061102	
Allergic_rhinitis	0.055526	0.027660	-0.023310	-0.010630	-0.001269	
Reflux_esophagitis	-0.119264	0.093998	0.073976	-0.047528	0.049926	
Asthma	-0.076884	0.108468	0.776015	0.361083	-0.483754	
Initial_days	-0.032397	-0.021285	0.004428	-0.002883	-0.012898	
TotalCharge	0.027736	0.021587	0.005720	0.003302	0.010772	
Additional_charges	0.003665	0.002048	-0.009207	0.012846	0.000300	

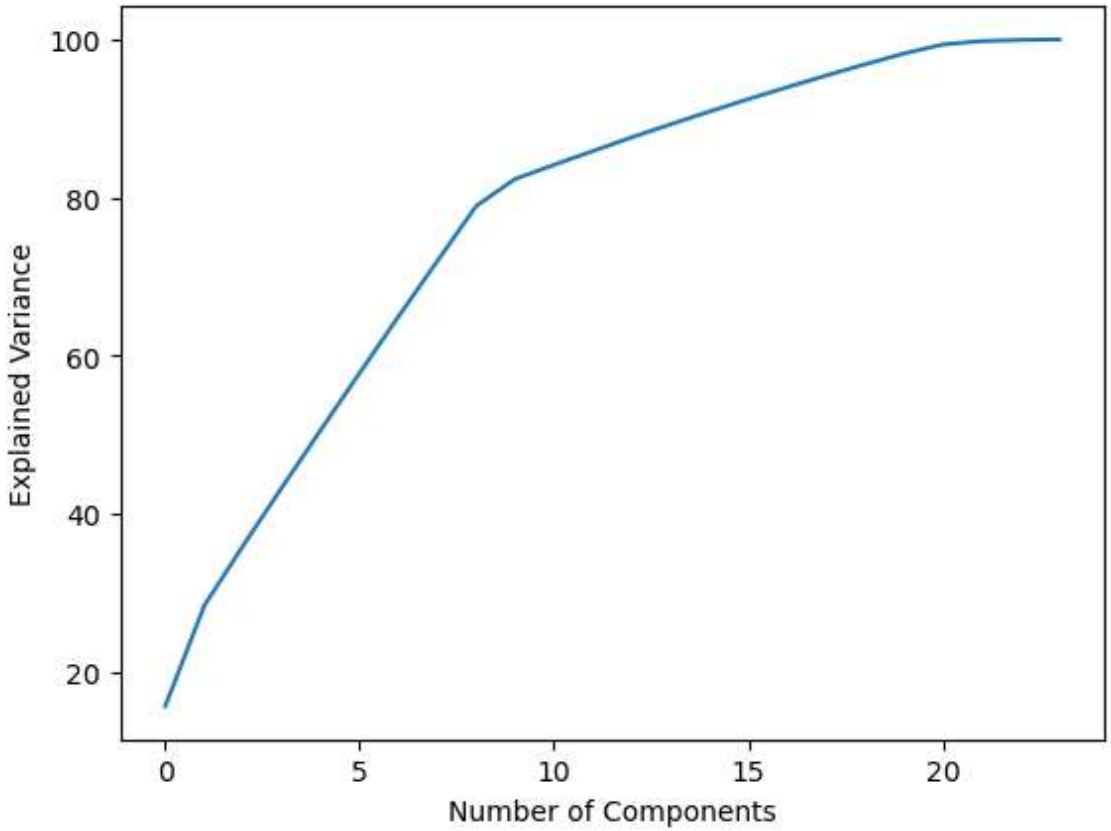
	PC20	PC21	PC22	PC23	PC24
Population	-0.004172	0.001132	0.001898	0.001490	-0.000844
Children	0.000477	-0.002054	0.001487	0.002750	-0.000284

```
Age                 0.000039 -0.008953  0.002449  0.368019  0.108028
Income             -0.002806 -0.001279  0.000312 -0.000461  0.001533
ReAdmis             0.010716  0.003152 -0.956180 -0.008993  0.036631
VitD_levels        -0.006421 -0.003352  0.002612 -0.001173 -0.002623
Doc_visits         -0.004744  0.002309  0.002524 -0.001278 -0.001729
Full_meals_eaten   -0.009700 -0.001336  0.001038  0.001407 -0.001048
vitD_supp           0.004795 -0.002622 -0.001885 -0.000199 -0.000689
Soft_drink          0.933744 -0.012059  0.006534  0.004772 -0.002870
HighBlood          -0.000554 -0.033942 -0.002384  0.737833  0.169054
Stroke              0.006315  0.995121  0.005272  0.029825  0.010374
Overweight          0.013174  0.007978  0.002757 -0.000182  0.003710
Arthritis           0.044874  0.050001 -0.020029 -0.000787 -0.021438
Diabetes           -0.310192 -0.031800 -0.011683  0.010450 -0.021133
Hyperlipidemia     -0.076033  0.048567 -0.004258  0.010994 -0.032952
BackPain           -0.075152 -0.006910 -0.013876  0.003224 -0.027519
Anxiety            -0.110673  0.046247 -0.023409  0.009787 -0.027906
Allergic_rhinitis   0.070173  0.012629 -0.018756  0.005007 -0.021385
Reflux_esophagitis  0.004257 -0.001924 -0.014959  0.007618 -0.019095
Asthma             -0.025074 -0.004141 -0.006295  0.002763  0.003567
Initial_days        0.004160 -0.000616  0.177263  0.177050 -0.689143
TotalCharge        -0.009019 -0.000489  0.228622 -0.173675  0.675638
Additional_charges  0.002615  0.003660 -0.002055 -0.507168 -0.149987

[24 rows x 24 columns]
Variance explained by all 24 components:  100.00000000000006
[ 15.67112336  28.34248605  35.94497159  43.3178299   50.59856589
  57.80982783  64.9504468   71.97173146  78.93945398  82.34166015
  84.12421232  85.87416533  87.60790675  89.25752533  90.87681586
  92.4519179   93.95529873  95.42970554  96.8464942   98.20987052
  99.35971457  99.7867892   99.92674901 100.        ]
```


Elbow PLot
Explained variance ratio
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
5
10
15
20
Principal component index

	Captured Variance per PC	Cumulative Sum
PC1	15.67	15.671123
PC2	12.67	28.342486
PC3	7.60	35.944972
PC4	7.37	43.317830
PC5	7.28	50.598566
PC6	7.21	57.809828
PC7	7.14	64.950447
PC8	7.02	71.971731
PC9	6.97	78.939454
PC10	3.40	82.341660
PC11	1.78	84.124212
PC12	1.75	85.874165
PC13	1.73	87.607907
PC14	1.65	89.257525
PC15	1.62	90.876816
PC16	1.58	92.451918
PC17	1.50	93.955299
PC18	1.47	95.429706
PC19	1.42	96.846494
PC20	1.36	98.209871
PC21	1.15	99.359715
PC22	0.43	99.786789
PC23	0.14	99.926749
PC24	0.07	100.000000




```
=======================
 Final Pipeline
=======================
[0.08749649 0.08533617 0.04739781 0.04653484 0.0462588  0.04534298
 0.04515122 0.04481873 0.04470852 0.04407286]
Total model explained variance:  0.5371
Model Accuracy:  96.27 %
```

In [ ]:

In [ ]: