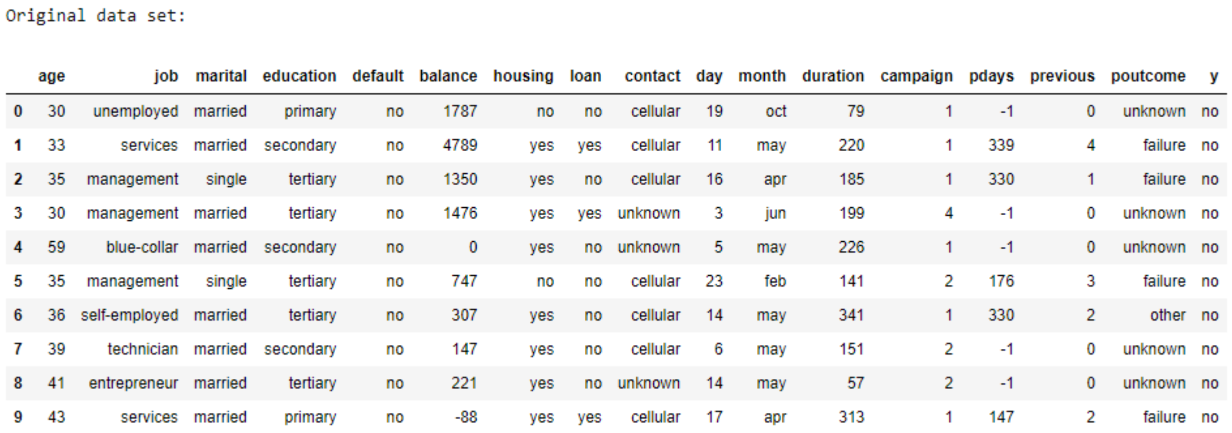
Yash Murthy

Data Mining

Assignment 3 Write Up (Questions on Bottom)

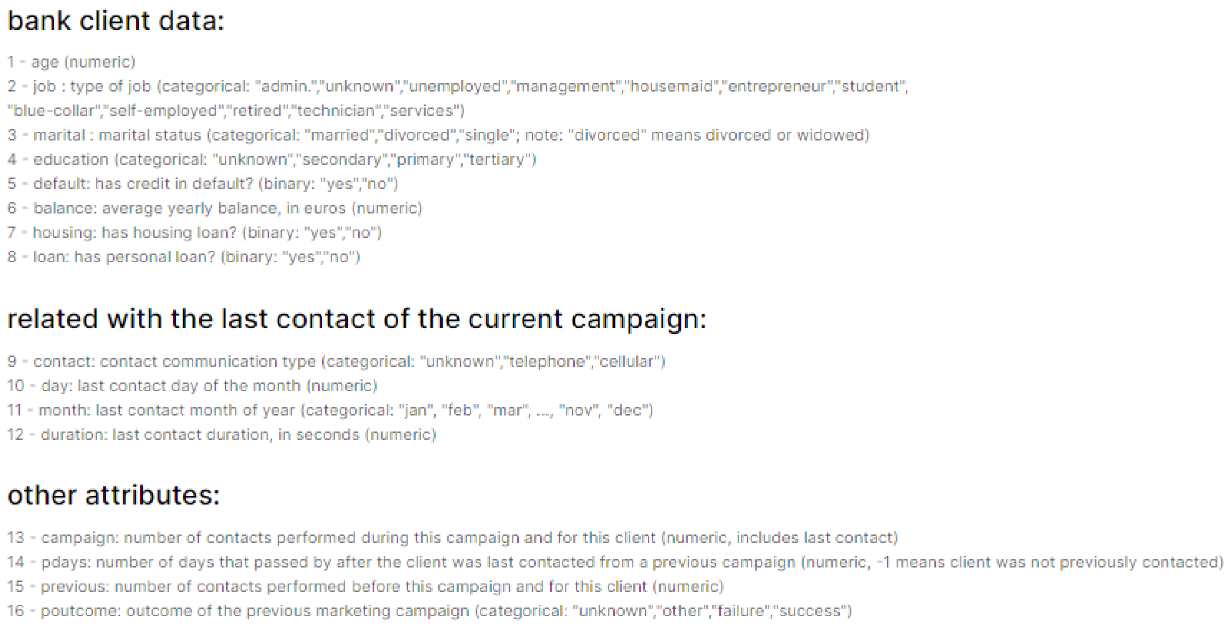
Our dataset for this assignment deals with direct marketing campaigns of a Portuguese banking institution. We are using the obtained data to determine whether or not a client will subscribe to a certain product, in this case that product is a bank term deposit. A 10 person sample from the original dataframe is as shown below. The y column (output variable) indicates whether or not the given client subscribed to the bank term deposit.



```
Original data set:
```

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
1	33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
2	35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
3	30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
4	59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no
5	35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no
6	36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	330	2	other	no
7	39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	-1	0	unknown	no
8	41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	-1	0	unknown	no
9	43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	147	2	failure	no

The following explains the meaning of each of the observed attributes. (Column headers above)



bank client data:

1 - age (numeric)
2 - job : type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
3 - marital : marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4 - education (categorical: "unknown","secondary","primary","tertiary")
5 - default: has credit in default? (binary: "yes","no")
6 - balance: average yearly balance, in euros (numeric)
7 - housing: has housing loan? (binary: "yes","no")
8 - loan: has personal loan? (binary: "yes","no")

related with the last contact of the current campaign:

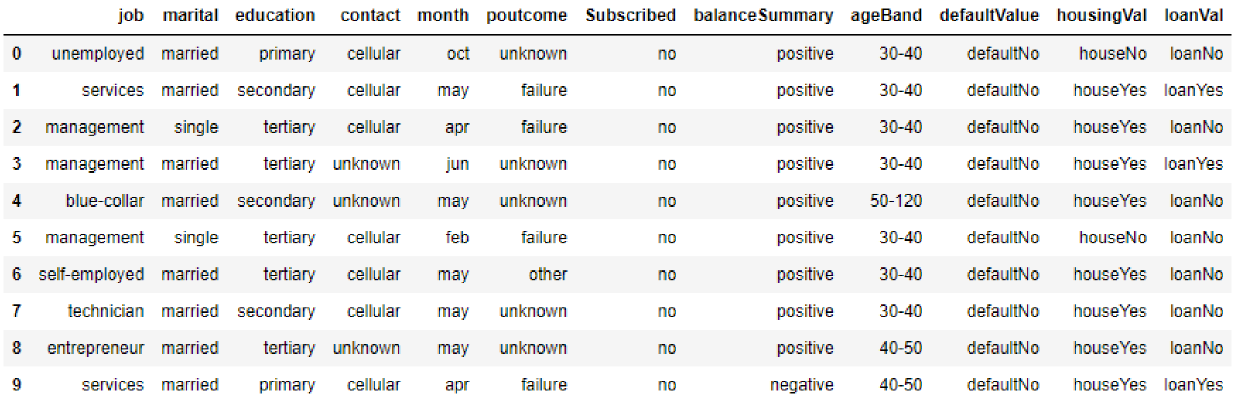
9 - contact: contact communication type (categorical: "unknown","telephone","cellular")
10 - day: last contact day of the month (numeric)
11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
12 - duration: last contact duration, in seconds (numeric)

other attributes:

13 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
15 - previous: number of contacts performed before this campaign and for this client (numeric)
16 - poutcome: outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

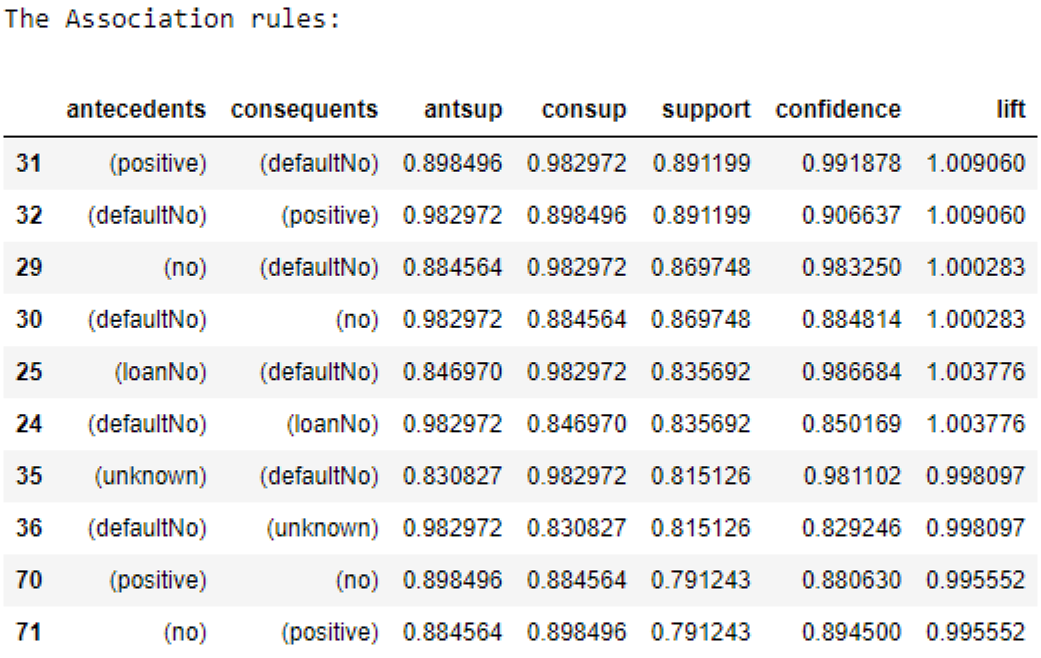
Association Rules

The next step was to find the association rules for the dataset. First, we must convert the numerical data into categorical data, as well as specify certain “yes” or “no” values from the original dataframe. We also changed the y column from the original dataframe to the column named ‘Subscribed’ for easier readability. Below the same 10 person sample as above with the categorical variable change implemented.



	job	marital	education	contact	month	poutcome	Subscribed	balanceSummary	ageBand	defaultValue	housingVal	loanVal
0	unemployed	married	primary	cellular	oct	unknown	no	positive	30-40	defaultNo	houseNo	loanNo
1	services	married	secondary	cellular	may	failure	no	positive	30-40	defaultNo	houseYes	loanYes
2	management	single	tertiary	cellular	apr	failure	no	positive	30-40	defaultNo	houseYes	loanNo
3	management	married	tertiary	unknown	jun	unknown	no	positive	30-40	defaultNo	houseYes	loanYes
4	blue-collar	married	secondary	unknown	may	unknown	no	positive	50-120	defaultNo	houseYes	loanNo
5	management	single	tertiary	cellular	feb	failure	no	positive	30-40	defaultNo	houseNo	loanNo
6	self-employed	married	tertiary	cellular	may	other	no	positive	30-40	defaultNo	houseYes	loanNo
7	technician	married	secondary	cellular	may	unknown	no	positive	30-40	defaultNo	houseYes	loanNo
8	entrepreneur	married	tertiary	unknown	may	unknown	no	positive	40-50	defaultNo	houseYes	loanNo
9	services	married	primary	cellular	apr	failure	no	negative	40-50	defaultNo	houseYes	loanYes

Upon creating this new categorized dataframe, we are now able to complete an association rules analysis. I set my minimum support value to 30%(0.3) and minimum confidence threshold to 50%(0.5). I sorted the resulting dataframe by support levels and the results are as follows.

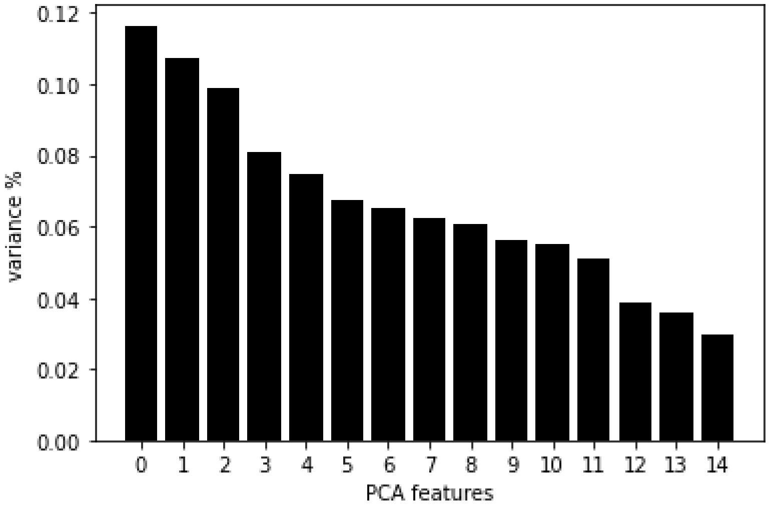


```
The Association rules:
```

	antecedents	consequents	antsup	consup	support	confidence	lift
31	(positive)	(defaultNo)	0.898496	0.982972	0.891199	0.991878	1.009060
32	(defaultNo)	(positive)	0.982972	0.898496	0.891199	0.906637	1.009060
29	(no)	(defaultNo)	0.884564	0.982972	0.869748	0.983250	1.000283
30	(defaultNo)	(no)	0.982972	0.884564	0.869748	0.884814	1.000283
25	(loanNo)	(defaultNo)	0.846970	0.982972	0.835692	0.986684	1.003776
24	(defaultNo)	(loanNo)	0.982972	0.846970	0.835692	0.850169	1.003776
35	(unknown)	(defaultNo)	0.830827	0.982972	0.815126	0.981102	0.998097
36	(defaultNo)	(unknown)	0.982972	0.830827	0.815126	0.829246	0.998097
70	(positive)	(no)	0.898496	0.884564	0.791243	0.880630	0.995552
71	(no)	(positive)	0.884564	0.898496	0.791243	0.894500	0.995552

Principal Component Analysis

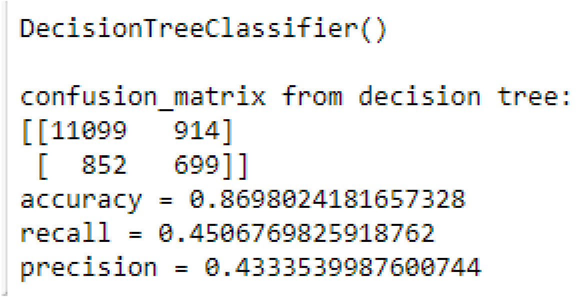
The first step in the PCA is to reduce the dimensionality of the data. In this case, I made a plot of the explained variances in a bar graph.



We can see that most of variance is explained by the first eleven features.

Decision Tree

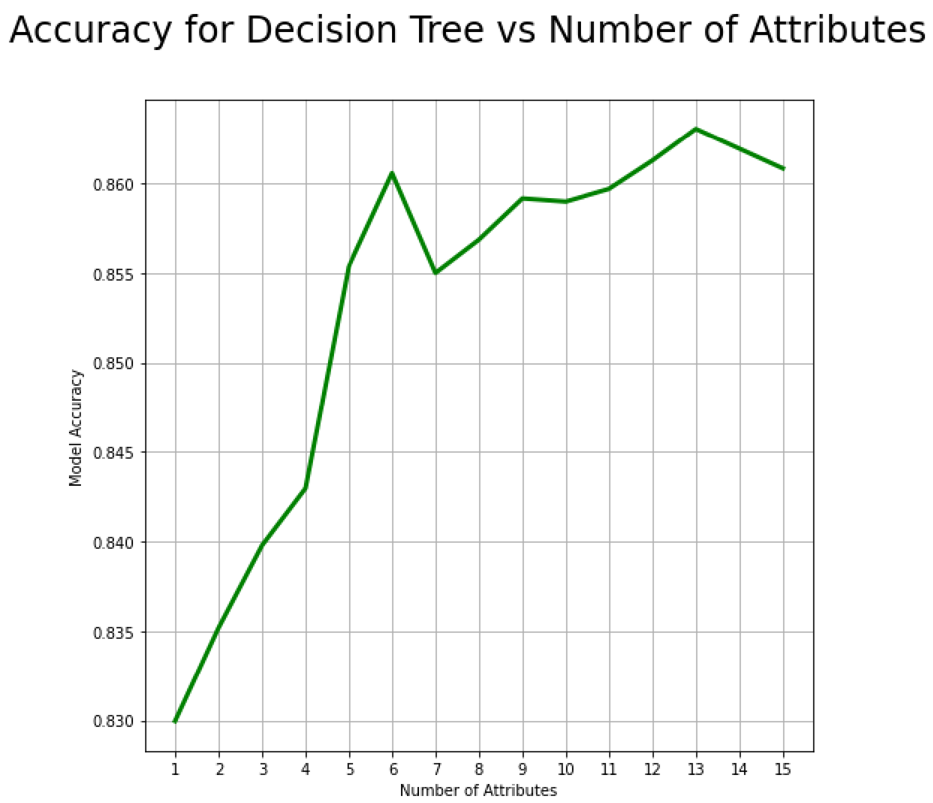
I created a decision tree matrix as well as the accuracy, recall, and precision values from the analysis. The results are as follows:



```
DecisionTreeClassifier()

confusion_matrix from decision tree:
[[11099   914]
 [  852   699]]
accuracy = 0.8698024181657328
recall = 0.4506769825918762
precision = 0.4333539987600744
```

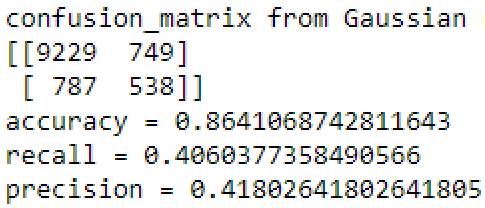
I then plotted the accuracies based on the number of components in order to see what number of components used would yield the most accurate results.


Accuracy for Decision Tree vs Number of Attributes

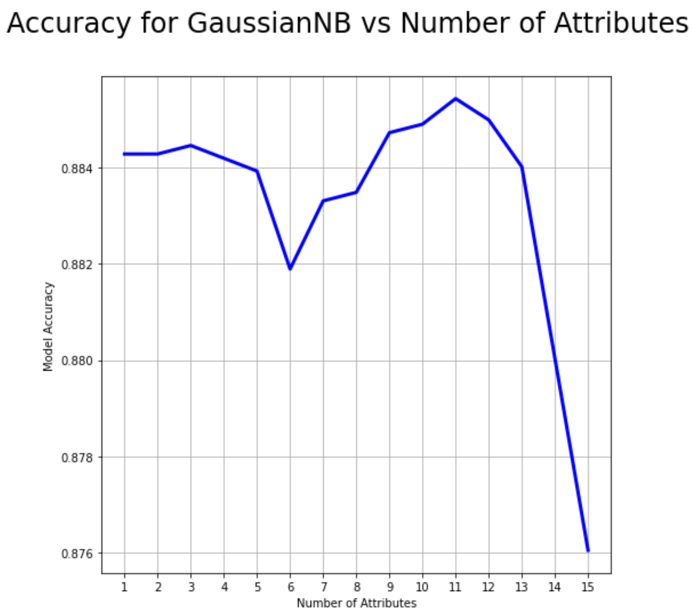
The highest peaks in the graph were for total number of components 6 and 13. I chose to show 6 in this case to minimize the number of features/processing used(code contains both results). The retained features from this included ‘job', 'balance', 'day', 'month', 'duration', 'campaign'.

Gaussian NB

Similar to the decision tree, I created a similar output for the Gaussian NB:



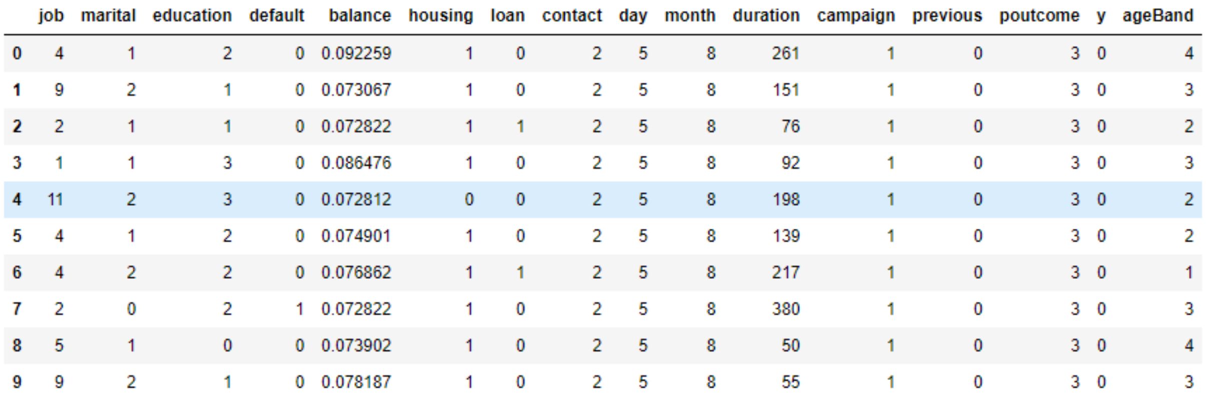
```
confusion_matrix from Gaussian naive bayes:
[[9229  749]
 [ 787  538]]
accuracy = 0.8641068742811643
recall = 0.4060377358490566
precision = 0.41802641802641805
```


Accuracy for GaussianNB vs Number of Attributes

The highest peak in the graph was clearly 11 components so I looked for those features in the original dataframe. In this case those retained features were 'job', 'marital', 'education', 'housing', 'loan', 'contact', 'day', 'month', 'duration', 'poutcome', 'ageBand'.

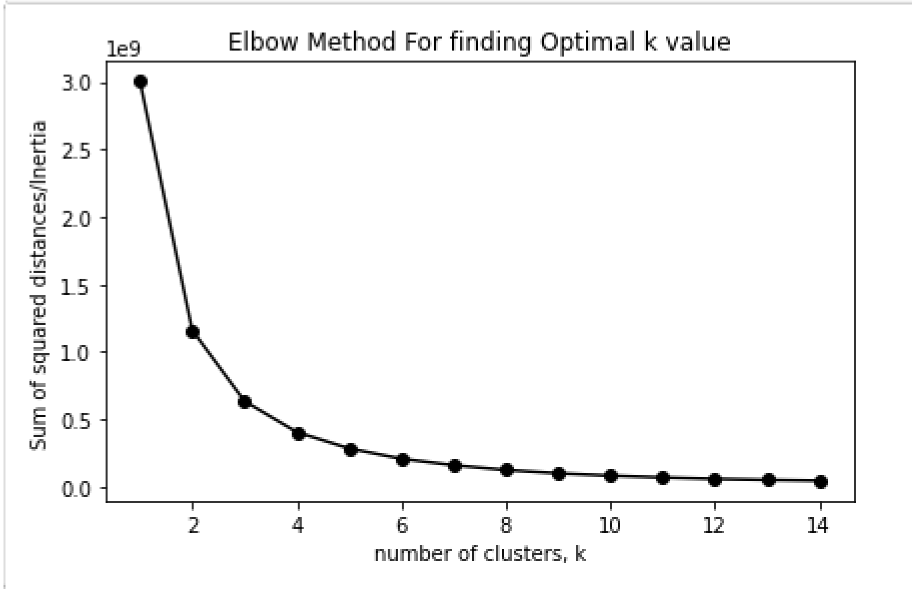
K-Means Clustering

For k means clustering, I Used the numerical data from ‘df3’.

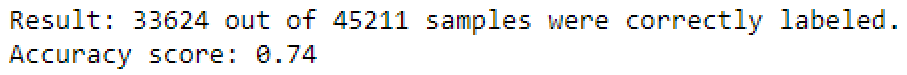


	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	previous	poutcome	y	ageBand
0	4	1	2	0	0.092259	1	0	2	5	8	261	1	0	3	0	4
1	9	2	1	0	0.073067	1	0	2	5	8	151	1	0	3	0	3
2	2	1	1	0	0.072822	1	1	2	5	8	76	1	0	3	0	2
3	1	1	3	0	0.086476	1	0	2	5	8	92	1	0	3	0	3
4	11	2	3	0	0.072812	0	0	2	5	8	198	1	0	3	0	2
5	4	1	2	0	0.074901	1	0	2	5	8	139	1	0	3	0	2
6	4	2	2	0	0.076862	1	1	2	5	8	217	1	0	3	0	1
7	2	0	2	1	0.072822	1	0	2	5	8	380	1	0	3	0	3
8	5	1	0	0	0.073902	1	0	2	5	8	50	1	0	3	0	4
9	9	2	1	0	0.078187	1	0	2	5	8	55	1	0	3	0	3

I graphed the number of clusters vs the sum of squared distances (inertia) within the range of predictive attributes (15).



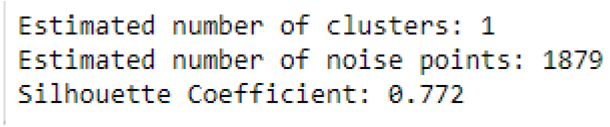
Using the elbow method, we see that the 'elbow' in the graph is at k=3. This means that the change in inertia is not significant any longer and likely neither is the variance, after that elbow point. I then ran a predictive model with the k means and it yielded an accuracy score of 0.74.



```
Result: 33624 out of 45211 samples were correctly labeled.
Accuracy score: 0.74
```

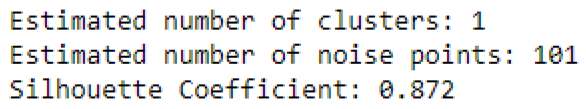
DB Scan

I also used the scalar (numerical) data for my DBScan. For this analysis I ran the DB Scan with two separate eps values (10 and 20). The eps value refers to how close points should be in order to be considered as part of a cluster. I set my min_samples (also known as minPoints) to 200 for each scan. The results were higher for the eps of 20 as shown:



```
Estimated number of clusters: 1
Estimated number of noise points: 1879
Silhouette Coefficient: 0.772
```

The silhouette coefficient is derived from the mean intracluster distance and the mean distance to the nearest cluster for a given point. The silhouette score ranged from -1 to 1. A score of 1 is the best, -1 the worst, and 0 signifies overlapping clusters. I also decided to raise the eps to 200 to see how that would alter the silhouette score. This means that there is a greater distance for a given point to be a neighbor with another point.



```
Estimated number of clusters: 1
Estimated number of noise points: 101
Silhouette Coefficient: 0.872
```

Questions

1. What can you deduce from the data set?   (in other words, what attribute values are indicative of "success")

Based on the results of the association rules, the best support came from the individuals who does not have any credit in default and has a positive bank balance. However, with there being such a large number of attributes to be analyzed, it is important to look at the remaining tests to see which ones are important. I think the best outcomes came from the classifier tests. The decision tree classifier provided an accuracy of about 86.5% when utilizing 13 attributes and ~86% when using 6 attributes. The 13 features from the decision tree were ('job', 'marital', 'education', 'default', 'balance', 'housing', 'loan', 'day', 'month', 'duration', 'campaign', 'previous', 'poutcome'). A similar, slightly lower accuracy of ~86.4% was obtained from the Gaussian Naïve Bayes classification. The resulting retained features were ('job', 'marital', 'education', 'housing', 'loan', 'contact', 'day', 'month', 'duration', 'poutcome', 'ageBand'). The overlap amongst the results from the NB and decision tree classification are job, marital, education, housing, loan, day, month, duration, and poutcome. These 9 attributes are what I would consider the biggest values to success.

2. Which mining techniques yielded the best results for what? How do you define "best results"? Please included tables and/or graphs to justify your statements about which are best.

I chose to define “best results” in this case as the results that yielded the highest accuracy. For predictive measures, the decision tree provided the highest accuracy when utilizing the 13 attributes listed in question 1. The next highest accuracy was provided by the NB classification. Based on my PCA it seemed that 11 attributes were important in explaining the variance in the dataset. K means only provided an accuracy of 74% when predicting whether a customer would get a subscription. My DBScan did not provide predictive results, but rather provided an analysis on the clusters, which ended up being quite high. All graphs and tables are posted in the write up above.

3. How useful was Principle Component Analysis?   For what number of components did you get the "best results". Please include tables and/or graphs to justify your statement.

Principal component analysis was not very helpful in this analysis. Based on the variance bar chart from above, we can see that most of the variances for the 15 attributes do not have much variation (highest is ~0.12 and lowest is ~0.03). I think with such a wide array of information or a majority of the attributes not contributing to the to whether or not a customer would subscribe to a bank loan deposit may lead to this lack of use for the PCA.