AMAT 502 FINAL PROJECT REPORT BACHELOR'S DEGREE MAJORS

Mahesh Utlapalli, Guru Datta Abhishek Yejju, Krishna Chaitanya Yejju, Tejas Patel

December 15 2021

1 Abstract

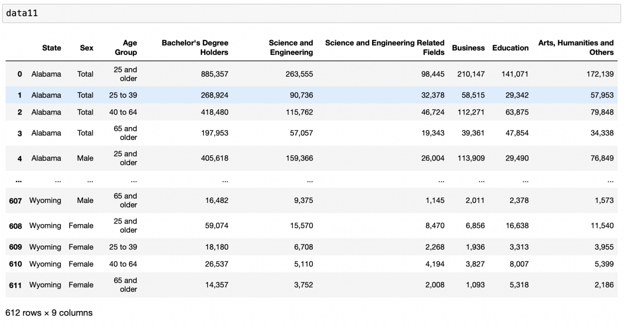
In Bachelor's Degree Majors, main aim of this project is to predict/classify Gender based on the different fields included in the dataset, For example: State, Sex, Age Group, Bachelor's Degree Holders, Science and Engineering, Science and Engineering-Related Fields, Business, Education, Arts, Humanities and Others. To classify the gender in this project, we used different algorithms like decision tree, naive Bayes classification, K-Means clustering, and logistic regression.

2 Introduction

A bachelor's degree or baccalaureate is an undergraduate academic degree awarded by colleges and universities upon completion of a course of study lasting three to six years (depending on the institution and academic discipline). In this dataset there are five most common bachelor's degrees are the Science and Engineering, Science and Engineering-Related Fields, Business, Education, Arts, Humanities and Others. In some institutions and educational systems, certain bachelor's degrees can only be taken as graduate or postgraduate educations after a first degree has been completed, although more commonly the successful completion of a bachelor's degree is a prerequisite for further courses such as a master's or a doctorate. In countries with qualifications frameworks, bachelor's degrees are normally one of the major levels in the framework (sometimes two levels where non-honors and honors bachelor's degrees are considered separately).

3 Dataset Overview

The Bachelor's Degree Majors dataset contains 9 features/columns and 612 observations.

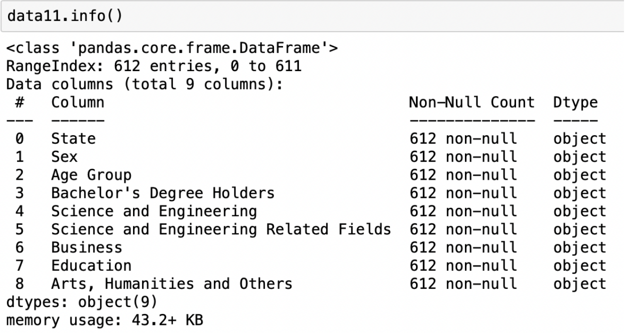


```
data11
```

	State	Sex	Age Group	Bachelor's Degree Holders	Science and Engineering	Science and Engineering Related Fields	Business	Education	Arts, Humanities and Others
0	Alabama	Total	25 and older	885,357	263,555	98,445	210,147	141,071	172,139
1	Alabama	Total	25 to 39	268,924	90,736	32,378	58,515	29,342	57,953
2	Alabama	Total	40 to 64	418,480	115,762	46,724	112,271	63,875	79,848
3	Alabama	Total	65 and older	197,953	57,057	19,343	39,361	47,854	34,338
4	Alabama	Male	25 and older	405,618	159,366	26,004	113,909	29,490	76,849
...	...	...	...	...	...	...	...	...	...
607	Wyoming	Male	65 and older	16,482	9,375	1,145	2,011	2,378	1,573
608	Wyoming	Female	25 and older	59,074	15,570	8,470	6,856	16,638	11,540
609	Wyoming	Female	25 to 39	18,180	6,708	2,268	1,936	3,313	3,955
610	Wyoming	Female	40 to 64	26,537	5,110	4,194	3,827	8,007	5,399
611	Wyoming	Female	65 and older	14,357	3,752	2,008	1,093	5,318	2,186

612 rows × 9 columns

Figure-1: Dataset



```
data11.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 612 entries, 0 to 611
Data columns (total 9 columns):
 #   Column                              Non-Null Count  Dtype 
---  ------                              --------------  ----- 
 0   State                               612 non-null    object
 1   Sex                                 612 non-null    object
 2   Age Group                           612 non-null    object
 3   Bachelor's Degree Holders           612 non-null    object
 4   Science and Engineering             612 non-null    object
 5   Science and Engineering Related Fields  612 non-null    object
 6   Business                            612 non-null    object
 7   Education                           612 non-null    object
 8   Arts, Humanities and Others         612 non-null    object
dtypes: object(9)
memory usage: 43.2+ KB
```

Figure-2: Dataset Info which includes the index dtype and column dtypes, non-null values, and memory usage.

4 Data Visualization

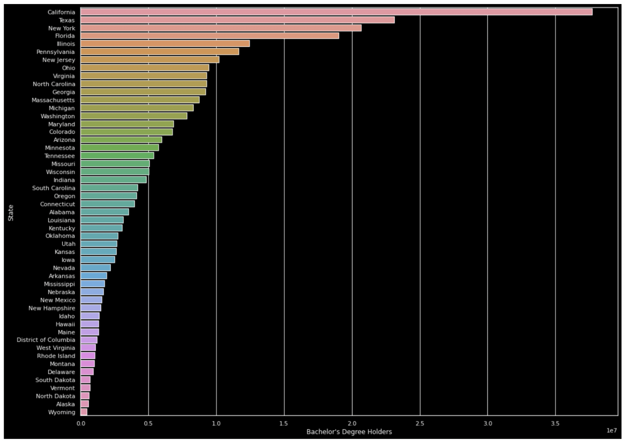


Figure-3: Number of bachelor's degree holders by states.

Figure-3 Insights are:

1. Highest number of bachelor's degree holders are in California, Texas, New York, and Florida.
2. Lowest number of bachelor's degree holders are in Vermont, North Dakota, Alaska, and Wyoming.

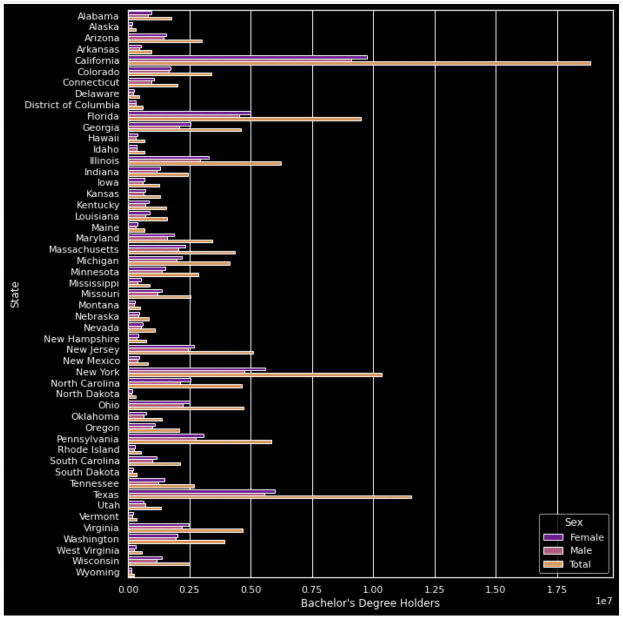


Figure-4: Number of bachelor's degree holders by states based on sex.

Figure-4 Insights are:

1. The state of 'Utah' is the only state where the number of males bachelor's degree holders is bigger than females.

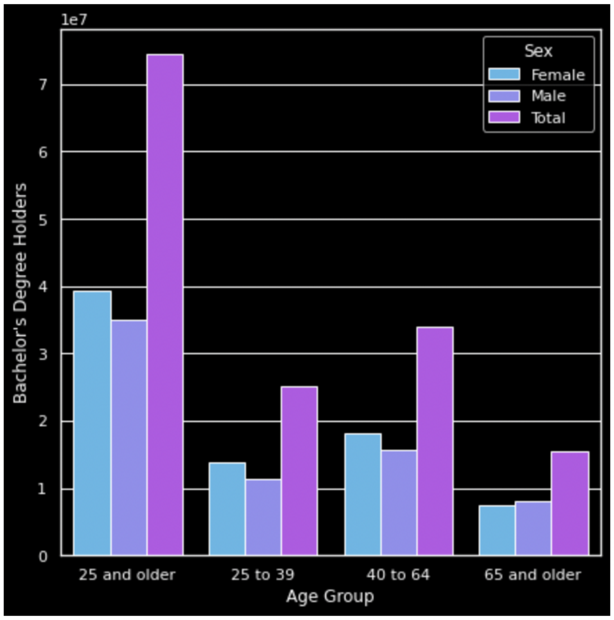


Figure-5: Number of females, males, and total based on age group.

Figure-5 Insights are:

1. The number of males is higher than females only in the '65 and older' age group and not considering total.

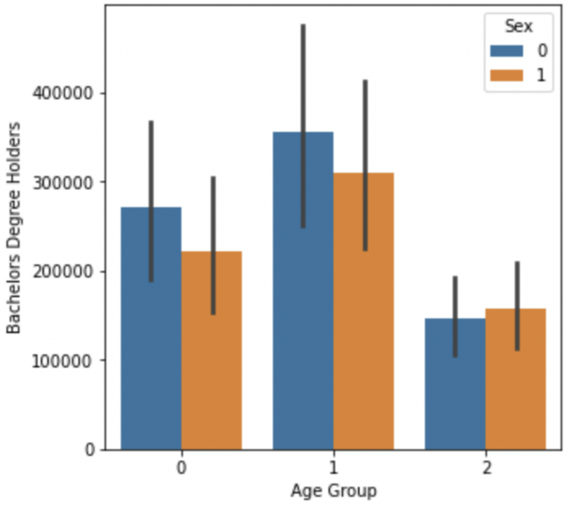


Figure-6: Graph shows which gender (Female, Male) have more Degree Holders.

Figure-6 Insights are:

1. The Age Group of 1 (40 to 64) has the more Bachelors Degree holders when compared with other age groups.

5 Methods

5.1 Naive Bayes Classification:

This algorithm is a supervised machine learning algorithm, which is working based on the Bayes theorem and is used for solving classification problems.

•Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building fast machine learning models that can make quick predictions on any dataset.

•It is a probabilistic classifier, which means it predicts based on the probability of an object.

•Some popular examples of the Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

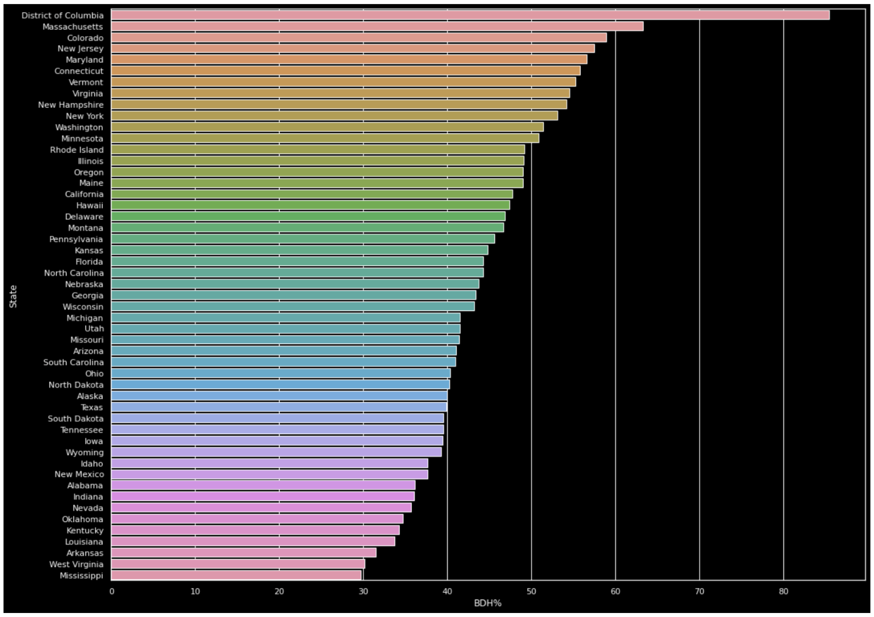


Figure-7: Comparing the percentage of bachelor's degree holders per population by states.

Figure-7 Insights are:

1. Highest percentage of bachelor's degree holders per population are in: District of Columbia, Massachusetts, and Colorado.
2. Lowest percentage of bachelor's degree holders per population are in Arkansas, West Virginia, and Mississippi.
3. The one interesting fact regarding our dataset is When we compared based on the number of bachelor's degree holders, california was in the first place, but now we can find it in the 17th position.

5.2 Decision Tree Classification

Classification algorithms use the input training data to predict the likelihood that subsequent data will fall into one category. We have different types of classifiers but compared to other Decision tree classifiers is easy to explain and does not require much normalization and scaling of data for our bachelor's degree Majors data set it is supervised learning and is perfect for classification problems. A decision tree classifier works like a flow chart, separating data points into two similar categories at a time from the tree trunk to branches then to leaves. A decision tree consists of two types of nodes decision nodes and end nodes or Leaf nodes. In the Decision tree classifier algorithm, CART uses the Gini method to create split points.
CART is a binary tree built by splitting nodes into two child nodes repeatedly based on a threshold value of an attribute with the help of the Gini Index criterion. Sklearn supports Gini criteria in decision tree classifier for Gini Index and by default, it takes Gini value, and we have implemented this because it favors larger partitions.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Figure-8: Gini Index.

Gini is calculated by subtracting the sum of the squared probabilities of each class from one

- More Gini Index = More Impure Class (1 Impure Class)
- Less Gini Index = More Pure Class
- Gini Index = 0(Pure class) i.e. It's a Leaf node and further Can't be Split.

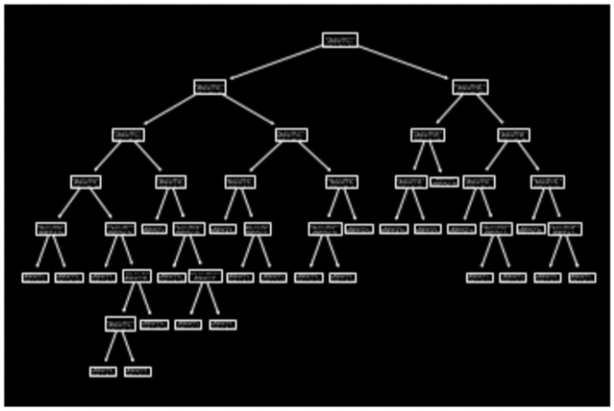


Figure-9: Decision Tree.

5.3 K-Means Clustering

Clustering refers to the process of automatically grouping data points with similar characteristics and assigning them to "clusters." Some use cases for clustering include: Recommender systems (grouping users with similar viewing patterns on Netflix, to recommend similar content).
K-Means: It is a method that aims to partition 'n' observations into 'k' clusters in which each observation belongs to the cluster with the nearest mean.

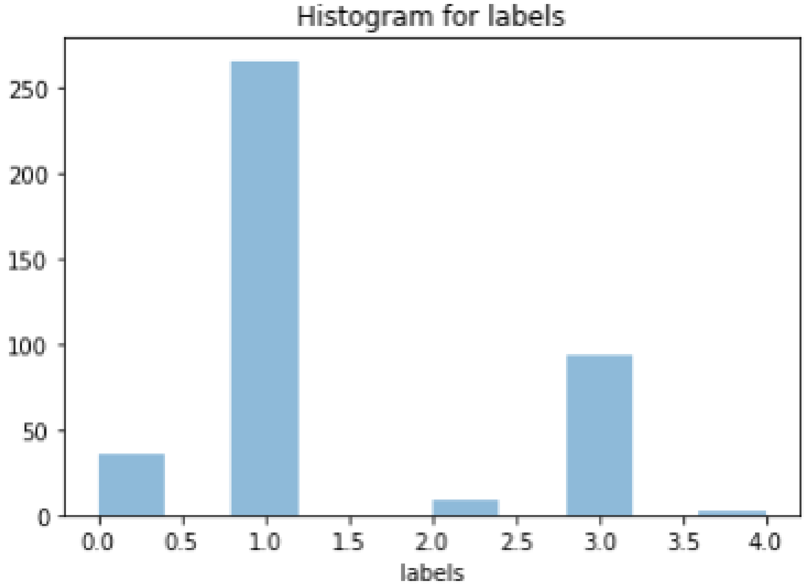


Figure-10: To summarize discrete or continuous data we used Histogram.

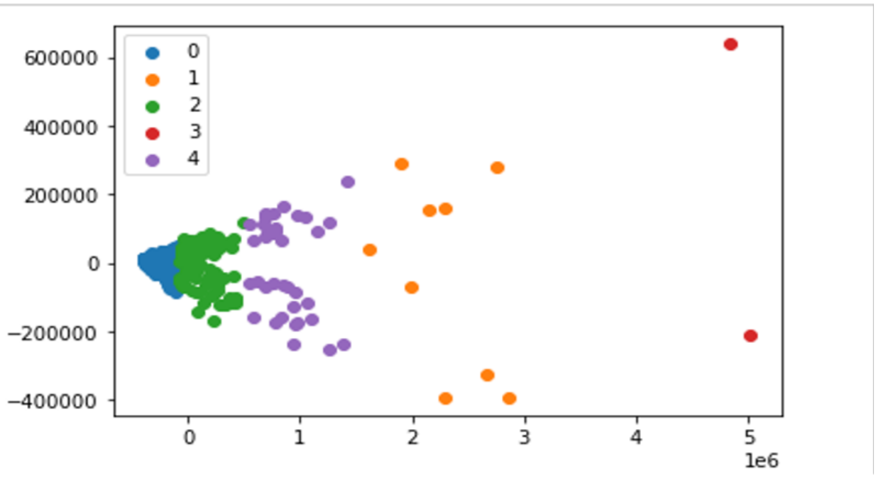
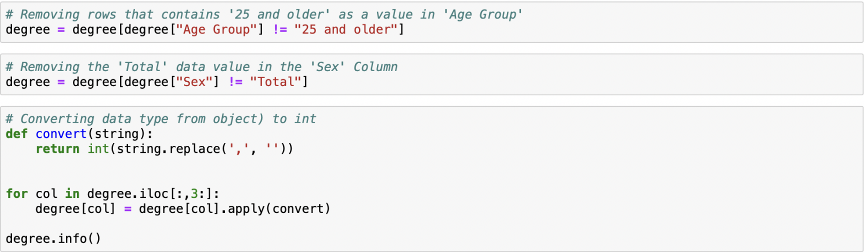


Figure-11: 2D scatter plot for all clusters.

5.4 Logistic Regression

Logistic Regression is used when the dependent variable(target) is categorical. For Example, To predict whether an email is a spam (1) or (0). Whether the tumor is malignant (1) or not (0). This method is speedy, easy to understand and it works well on non-linear data. We used logistic regression to predict whether the bachelor's degree holder is a male or a female. Preprocessing: Initially, the raw dataset contains 612 rows and 9 columns. After removing unnecessary information we left with 306 rows and 8 columns which represent only useful information. Preprocessing is important to increase the accuracy of a predicted variable.



```
# Removing rows that contains '25 and older' as a value in 'Age Group'
degree = degree[degree["Age Group"] != "25 and older"]
```

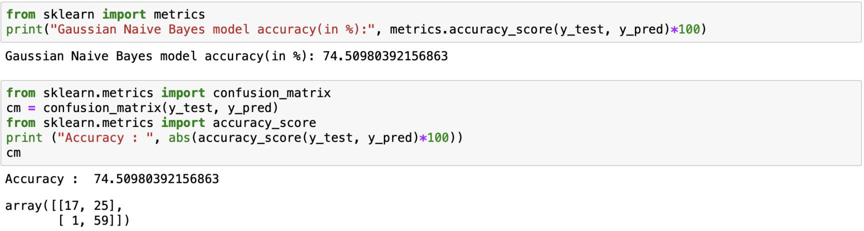
```
# Removing the 'Total' data value in the 'Sex' Column
degree = degree[degree["Sex"] != "Total"]
```

```
# Converting data type from object) to int
def convert(string):
    return int(string.replace(',', ''))


for col in degree.iloc[:,3:]:
    degree[col] = degree[col].apply(convert)

degree.info()
```

Figure-12: Applying preprocessing on data.

6 Results and Analysis

6.1 Naive Bayes



```
from sklearn import metrics
print("Gaussian Naive Bayes model accuracy(in %):", metrics.accuracy_score(y_test, y_pred)*100)
```

```
Gaussian Naive Bayes model accuracy(in %): 74.50980392156863
```

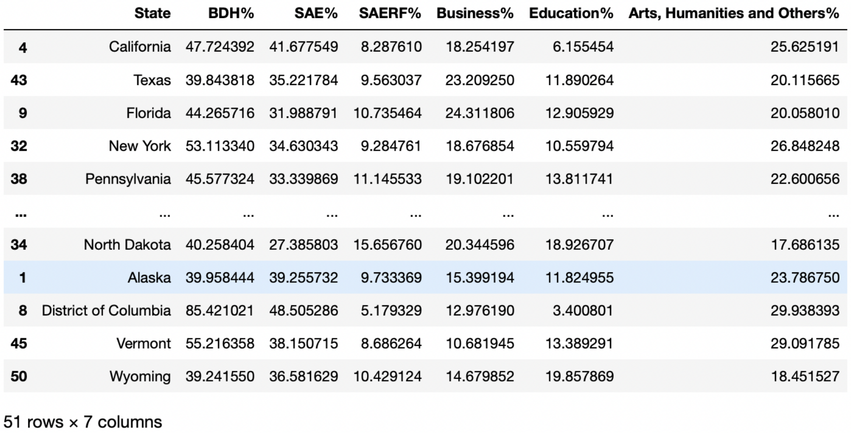
```
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
from sklearn.metrics import accuracy_score
print ("Accuracy : ", abs(accuracy_score(y_test, y_pred)*100))
cm
```

```
Accuracy :  74.50980392156863

array([[17, 25],
       [ 1, 59]])
```

Figure-13: Accuracy and Confusion Matrix.

From the above figure, accuracy of the Gaussian Naive Bayes classifier is 74% in this I predicted the gender column based on the different degree types those are: Bachelors degree holders, science and engineering, science and engineering-related fields, business, education, and arts humanities, and other columns.

The confusion matrix shows the ways in which your classification model is confused when it makes predictions.



	State	BDH%	SAE%	SAERF%	Business%	Education%	Arts, Humanities and Others%
4	California	47.724392	41.677549	8.287610	18.254197	6.155454	25.625191
43	Texas	39.843818	35.221784	9.563037	23.209250	11.890264	20.115665
9	Florida	44.265716	31.988791	10.735464	24.311806	12.905929	20.058010
32	New York	53.113340	34.630343	9.284761	18.676854	10.559794	26.848248
38	Pennsylvania	45.577324	33.339869	11.145533	19.102201	13.811741	22.600656
...	...	...	...	...	...	...	...
34	North Dakota	40.258404	27.385803	15.656760	20.344596	18.926707	17.686135
1	Alaska	39.958444	39.255732	9.733369	15.399194	11.824955	23.786750
8	District of Columbia	85.421021	48.505286	5.179329	12.976190	3.400801	29.938393
45	Vermont	55.216358	38.150715	8.686264	10.681945	13.389291	29.091785
50	Wyoming	39.241550	36.581629	10.429124	14.679852	19.857869	18.451527

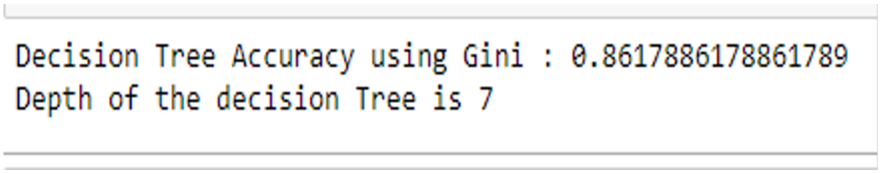
51 rows × 7 columns

Figure-14: Percentage of all bachelor degree holders per population by states.

Figure-14 Insights are:

1. Based on the above percentages, we can see that science and engineering course is crushing it in every single state.
2. The lowest percentage was in Nebraska (27.1%) and the highest in the district of Columbia (48.5%).

6.2 Decision Tree Classification



```
Decision Tree Accuracy using Gini : 0.8617886178861789
Depth of the decision Tree is 7
```

Figure-15: Decision Tree Accuracy.

Using decision tree classifier on the Bachelor Degree Majors data set gives an accuracy of 86.1% and Depth of decision tree is 7.

6.3 K-Means Clustering

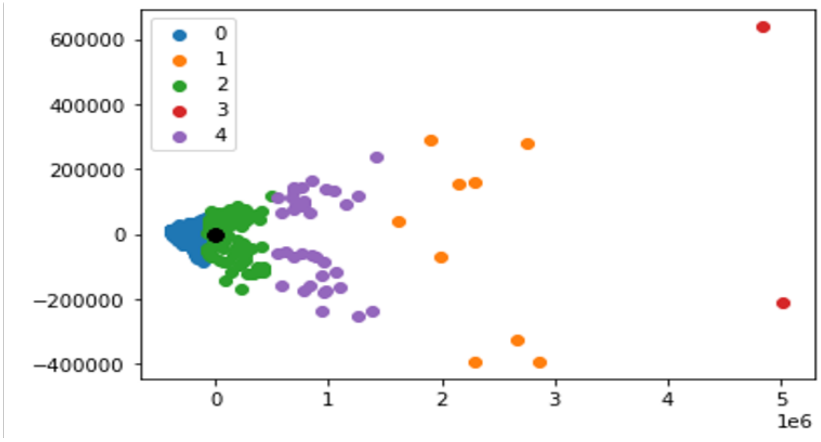
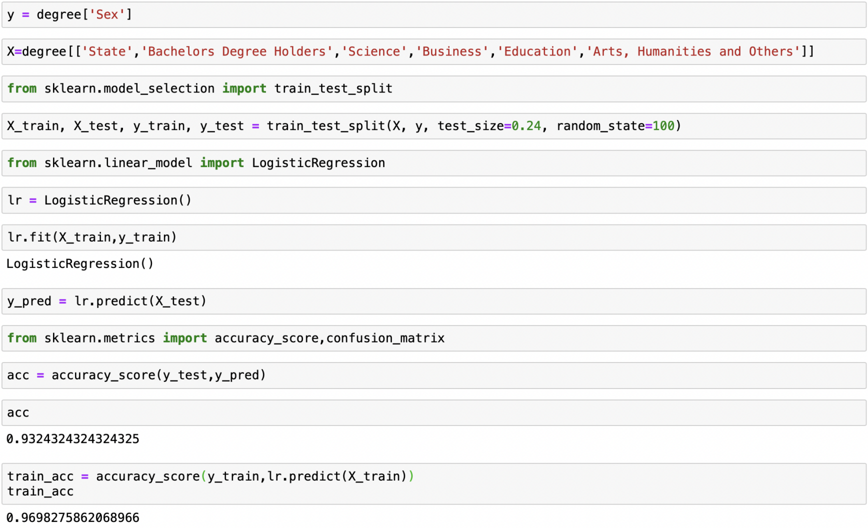


Figure-16: Centroid for all clusters.

6.4 Logistic Regression



```
y = degree['Sex']

X=degree[['State','Bachelors Degree Holders','Science','Business','Education','Arts, Humanities and Others']]

from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.24, random_state=100)

from sklearn.linear_model import LogisticRegression

lr = LogisticRegression()

lr.fit(X_train,y_train)
LogisticRegression()

y_pred = lr.predict(X_test)

from sklearn.metrics import accuracy_score,confusion_matrix

acc = accuracy_score(y_test,y_pred)

acc
0.9324324324324325

train_acc = accuracy_score(y_train,lr.predict(X_train))
train_acc
0.9698275862068966
```

Figure-17: Accuracy of the Logistic Regression.

From the above figure, accuracy of the Logistic Regression is 93% in this I predicted the gender column based on the different degree types.

7 Conclusion

We have used different algorithms like decision tree, naïve Bayes classification, K-Means clustering and lastly logistic regression in those algorithms the Accuracy of the Logistic Regression algorithm is comparatively more than other algorithms. From this, we can conclude that the prediction of Gender (Male, Female) is easy when comparing with age group and state because when we predict the gender after pre-processing of our dataset there are only binary values either 0 or 1. so this makes the more accuracy occurred in the logistic regression algorithm.

8 Future Scope

We used different algorithms like a decision tree and naive bayes classification, K-Means clustering, and logistic regression, but we will be using some other algorithms like ID3, MARS, C4.5, CHAID, Gaussian Mixture Model, linear regression, and multivariate regression where we could have compared the accuracy.

9 Work Distribution

9.1 Mahesh Utlapalli

– Visualization
– Naive Bayes Classification
– Paper Work
– Latex Documentation

9.2 Krishna Chaitanya Yejju

– Visualization
– Decision Tree Classifier
– Conclusion

9.3 Guru Datta Abhishek Yejju

– Visualization
– K-Means Clustering
– Latex Documentation

9.4 Tejas Patel

– Visualization
– Preprocessing
– Logistic Regression
– Future Scope

10 Refereces

1. https://www.kaggle.com/tjkyner/bachelor-degree-majors-by-age-sex-and-state
2. https://en.wikipedia.org/wiki/Bachelor%27s_degree
3. https://towardsdatascience.com/how-to-encode-categorical-columns-using-python-9af10b36f049
4. https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c