WOMEN IN STEM FIELD

Written by:

Farhana Afroze

Email: fafroze@buffalo.edu

Samuel Fergot

Email: sdfergot@buffalo.edu

Abstract: **In this project we would like to analyze how women doing on stem field, both in school and the workplace, how they have been treated in stem field or tech job field. We will analyze the data and then implement some algorithm to find out if our hypothesis about this is true.**

I.INTRODUCTION

This project we would like to find out how women doing in stem field. The dataset we chose for analyzing this topic is from the Pew Research Center 2017 Stem Survey. This dataset was made for analyzing US adult careers and their education in the STEM field. It has sample size of 4914 adults and all of these samples were taken from all 50 U.S states and The District of Columbia. Our project topic is about women is STEM field. We would like to analyze how women are doing in the stem field, both in school and the workplace. We believe this dataset can be used to answer several relevant questions. Do more women pursue STEM majors in university than man? Do women tend to pursue a STEM major or a non-STEM major? Do women like Math and Science courses more than non-STEM courses? In the job sector do they get less appreciated? Do they face harassment in this field more? Do men get more value than women even though both genders have similar qualifications and play the same role in the workplace? We would like to analyze all these issues using the dataset through EDA and also implement some machine learning algorithm to justify some of these issues. In conclusion, we would like to address issues like these. Also, we would like to analyze this dataset and implement some algorithm so we can answer all these questions.

II.HYPOTHESIS

We would like to create some hypothesis and run some machine learning algorithm to justify our hypothesis. Following are the hypothesis we would like to justify through our algorithm:

- Hypothesis-1: Discrimination against women is a major problem and women face discrimination in stem or tech field more than man.
- Hypothesis-2: Women are less interested in math, science and engineering field.
- Hypothesis-3: Females are behind in stem field
- Hypothesis-4: Education, income, stem_degree and occupation feature has more effect on male than female, means male educated more and income more than women.

If our hypothesis become true then we would like to more women come to this field and raise awareness about the inequality and harassment that women face.

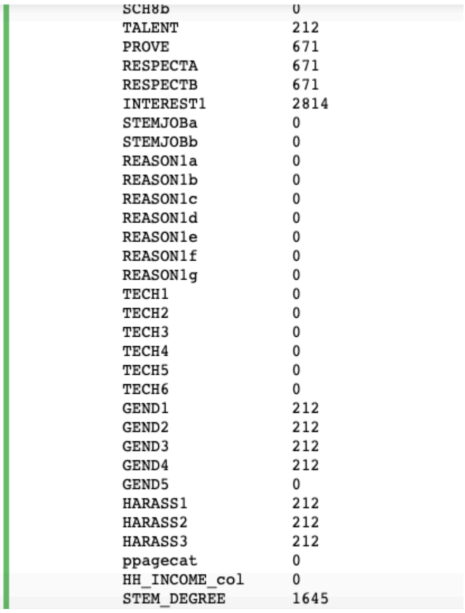
III. Data Cleaning

At first, we had to clean our dataset. We performed various cleaning method. Following is the full of list the methods we performed-

1. In our dataset we had like 220 columns. So, we would like to remove some columns that we don’t want in our experiment. The topic is about women in stem field. So, we would like to remove those columns that doesn’t have any effect of this

topic. Because this dataset is about U.S adult career and education of stem field so there are so many irrelevant columns in the dataset we needed to drop. After dropping variables, we got total 40 columns.

2.Next we tried to find missing values in the dataset.

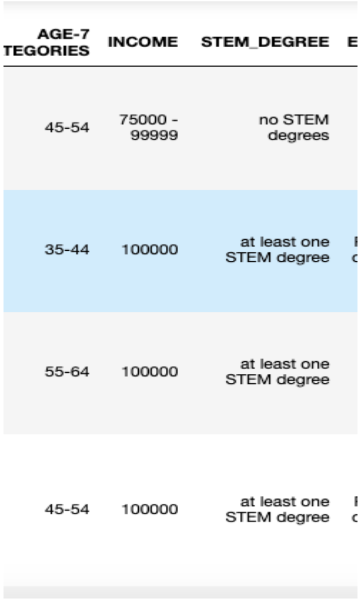


```
SCH8b              0
TALENT             212
PROVE              671
RESPECTA           671
RESPECTB           671
INTEREST1          2814
STEMJOBa           0
STEMJOBb           0
REASON1a           0
REASON1b           0
REASON1c           0
REASON1d           0
REASON1e           0
REASON1f           0
REASON1g           0
TECH1              0
TECH2              0
TECH3              0
TECH4              0
TECH5              0
TECH6              0
GEND1              212
GEND2              212
GEND3              212
GEND4              212
GEND5              0
HARASS1            212
HARASS2            212
HARASS3            212
ppagecat           0
HH_INCOME_col      0
STEM_DEGREE        1645
```

As we see those are all the missing values in each column. Next, we removed the missing values. We now got '0' missing values in our dataset.

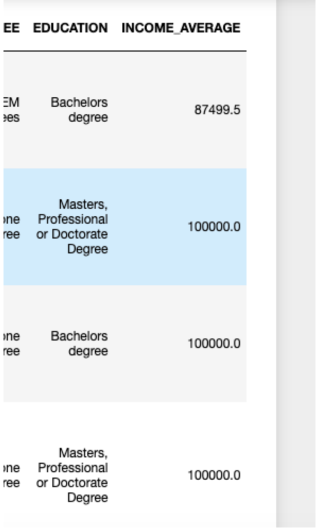
3.Next we checked if there are any duplicate rows in our dataset. We found there are no duplicate rows in our dataset.

4. Next we would like to clean the 'income' column feature. Because in the income column there are string with numbers. So, In the 'Income' column we would like to remove string that is with numbers and the special characters and replace '-' in the place of 'to' and also rename other columns and making uppercase to all columns. Here income column now looks like-



AGE-7 TEGORIES	INCOME	STEM_DEGREE	E
45-54	75000 - 99999	no STEM degrees	
35-44	100000	at least one STEM degree	I c
55-64	100000	at least one STEM degree	
45-54	100000	at least one STEM degree	I c

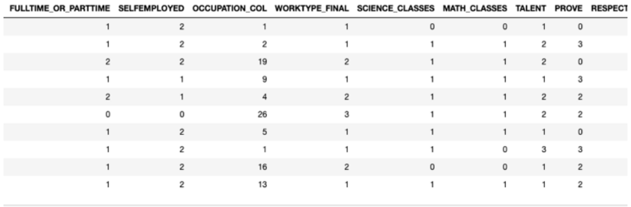
5. Next we would like to average income column. Here we would like to add new column called income average and we would like to get this doing averaging of income column range. We would like to get income average from income 30,000 to 99,999 and any value less than or greater than given as it was. In our dataset it was like 30,000 or less so we are keeping 30,000 as average and 100,000 or more than 100,000 we are keeping 100,000 as average.



EE	EDUCATION	INCOME_AVERAGE
EM ees	Bachelors degree	87499.5
one ree	Masters, Professional or Doctorate Degree	100000.0
one ree	Bachelors degree	100000.0
one ree	Masters, Professional or Doctorate Degree	100000.0

6. In the ‘age’ column every rows, if it has age ‘75+’, I would like to delete ‘75+’ and then keep the age category 18-74 and all the data points of that age range.

7. Next we did transformation all the categorical values into numerical.



FULLTIME_OR_PARTTIME	SELFEMPLOYED	OCCUPATION_COL	WORKTYPE_FINAL	SCIENCE_CLASSES	MATH_CLASSES	TALENT	PROVE	RESPECT
1	2	1	1	0	0	1	0	
1	2	2	1	1	1	2	3	
2	2	19	2	1	1	2	0	
1	1	9	1	1	1	1	3	
2	1	4	2	1	1	2	2	
0	0	26	3	1	1	2	2	
1	2	5	1	1	1	1	0	
1	2	1	1	1	0	3	3	
1	2	16	2	0	0	1	2	
1	2	13	1	1	1	1	2	

8. Next we checked outlier from histogram and removed those outliers.

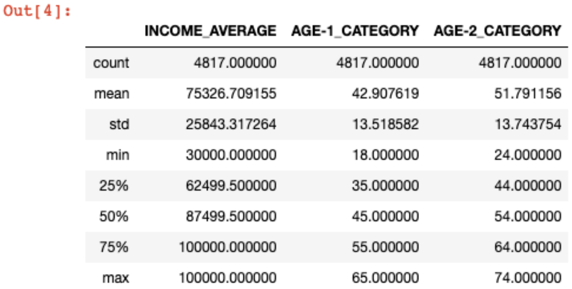
9. Next, we added two employment column as a table where '1' would be for employment and '-1' would for unemployment from our old 'EMPLOYMENT_STATUS' column.

10. We also added a new column name ‘Behind stem field’ which has 0 and 1 value. Where ‘0’ means women are not behind in stem field and ‘1’ means women are behind in stem field.

IV. EXPLORATORY DATA ANALYSIS

Here we would like to do some exploratory data analysis so we can pull out information of the dataset. We can analyze some of issues we trying to find through this EDA technique.

EDA-1: Here we would like to find out what is the min, max and average for income column, for two categories of age column so we can get idea about age and income of people who are participating in survey.

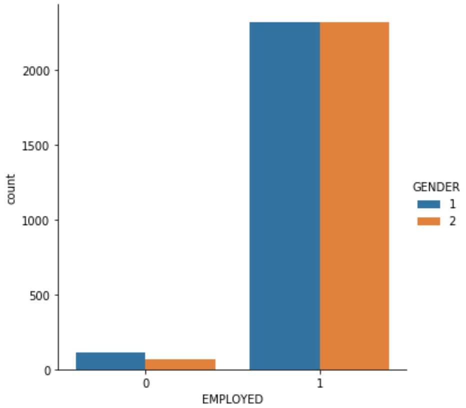


Out[4]:

	INCOME_AVERAGE	AGE-1_CATEGORY	AGE-2_CATEGORY
count	4817.000000	4817.000000	4817.000000
mean	75326.709155	42.907619	51.791156
std	25843.317264	13.518582	13.743754
min	30000.000000	18.000000	24.000000
25%	62499.500000	35.000000	44.000000
50%	87499.500000	45.000000	54.000000
75%	100000.000000	55.000000	64.000000
max	100000.000000	65.000000	74.000000

As we see here for income average column we have average for 4817 people is- 75,326 and min income is- 30,000 and max income is- 100,000. We see in another two columns ‘age-1-category’ and ‘age-2-category’. In ‘age-1-category’ column we see minimum age is- 18 and maximum age is 65 and another column ‘age-2-category’ we see minimum age is 24 and maximum age is 74.

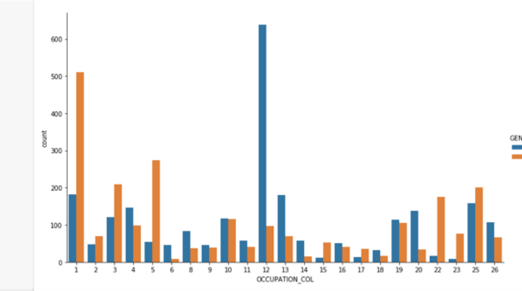
EDA-2: Next we would like to find out how many people are unemployed and employed through visualization and would like to see who is more unemployed? Male or Female?



Here we see ‘EMPLOYED’ has two categorical values, so employed = 1 means working and employed = 0 means not working. Here we see another variable called ‘Gender’ where Gender = 1 is Female and Gender = 2 is Male. So, we can see both male and female working has same numbers while for non-working individual female number is little bit more than male.

So, we can say by doing EDA women are little more unemployed.

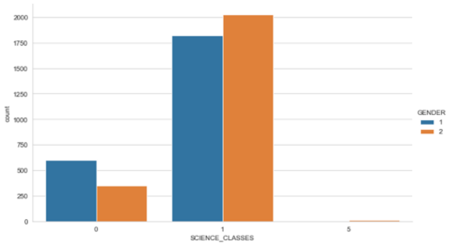
EDA-3: Here we would like to see in stem field how many male and female. There are so many occupation column, so would like to see which stem and other occupation field mostly women goes and are they behind in stem field or math, science field or technology field than man?

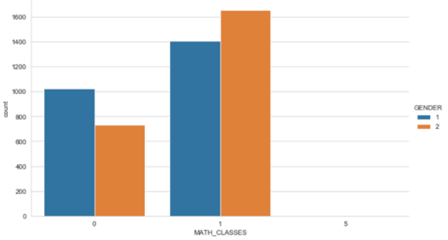


We found that for computer and math sector we see 182 female and 510 male which is pretty low. In architecture and engineering sector we see 54 is female and 274 is male. In medical doctor field we see 58 female and 40 male and this is where we see more female than male. In health care support field we see 57 female and 15 male. In other health care practitioner, we see 638 females and 97 males. As we can see in medical related field there are more female than male and we see in computer, math, engineering field there are more male than female. So, basically we can tell some of stem field there are more women but for math, engineering and tech field there are more man.

EDA-4: Here I would like to see who like or dislike science class more? And also who like or dislike math class more? Male or Female? If girls dislike math and science classes more than boys, we would like to more girls to like these classes so that in future they get interested in math and tech job fields.

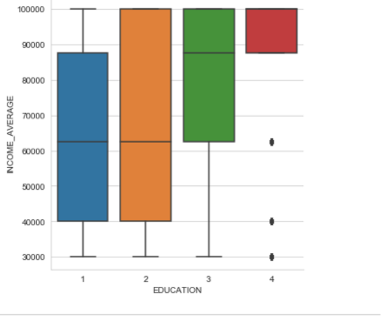
```
Out[16]: <seaborn.axisgrid.FacetGrid at 0x11cf89d90>
```

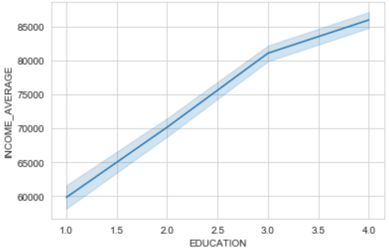




Here 1 is female and 2 is male. And also 0 is dislike label and 1 is like label. we see female less like math and science classes and they dislike math and science classes more than male.

EDA-5: Next we would like to see if higher education has more higher income.





Here we see Education which has Four categories-

High school graduate or less - 1 Some college, including Associate degree - 2 Bachelor’s degree - 3 Masters, professional or doctorate degree - 4 Here from the plot we can see education-1 which is high school graduate or less has median value little above 60,000 and lower quartile is about 40,000 and upper quartile is about 88,000. Next, we see for education-2 which is some college, including associate degree has median value above 60,000 and lower quartile is about 40,000 and upper quartile is about 100,000. Next, we see for education-3 which is Bachelor’s degree has median value is about 88,000 and lower quartile is about 62,000 and upper quartile is about 100,000. Next, we see for education-4 which is master’s degree, professional or doctorate degree has median value 100,000 and lower- quartile is about 88,000 and also we see some outliers here.

we see from correlation that education and income is showing low correlation. It has correlation 0.35 between education and income.

And also see in linear plot as the education increase, income is increasing too.

EDA-6: Here we would like to find out REASON1A, REASON1B, REASON1C, REASON1D and REASON1F survey columns responses and we would like to find the reasons why there are not more women working in science, technology, engineering and math jobs. There are some features in the dataset called-

REASON1A - From an early age, girls are not encouraged to pursue these subjects in school

REASON1B - Women are less likely than men to believe that they can succeed in these fields

REASON1C - Women do not pursue these jobs because there are so few female role models in these fields

REASON1D- Women face discrimination in the recruitment process, hiring and promotions

REASON1F - Women are just less interested in science, technology, engineering and math than men

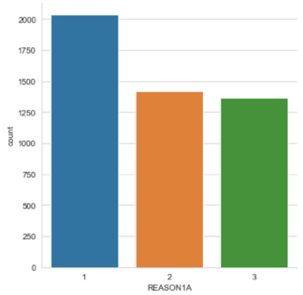
All those has cateegorical variables-

"A major reason"-1,

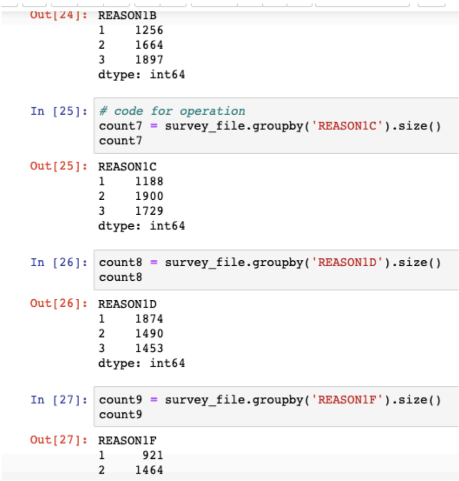
"A minor reason"-2,

"Not a reason"-3,

So, we would like to know from people what they think about this-



From plot we see almost 2000 individual thinks girls are not encouraged to pursue these subjects in school is the major reason why not more women working in science, technology, engineering and math jobs.



```
Out[24]: REASON1B
         1    1256
         2    1664
         3    1897
         dtype: int64

In [25]: # code for operation
         count7 = survey_file.groupby('REASON1C').size()
         count7

Out[25]: REASON1C
         1    1188
         2    1900
         3    1729
         dtype: int64

In [26]: count8 = survey_file.groupby('REASON1D').size()
         count8

Out[26]: REASON1D
         1    1874
         2    1490
         3    1453
         dtype: int64

In [27]: count9 = survey_file.groupby('REASON1F').size()
         count9

Out[27]: REASON1F
         1     921
         2    1464
```

REASON1B - Women are less likely than men to believe that they can succeed in these fields.

Here we see 1256 individual thinks REASON1B is the major reason more women not working these science, technology,engineering and math jobs.

REASON1C - Women do not pursue these jobs because there are so few female role models in these fields

Here we see 1188 individual thinks REASON1C is the major reason that women don't work in these fields

REASON1D- Women face discrimination in the recruitment process, hiring and promotions

Here we see 1874 individual thinks REASON1D is the major reason that women don't work in these fields

REASON1F - Women are just less interested in science, technology, engineering and math than men Here we see 921 individuals thinks REASON1F is the major reason that women don't work in these fields

EDA-7:

Here we like to see GEND1 and GEND2 columns survey response

Reason for doing this: Would like to know about 'GEND1', 'GEND2' features so that we could know if there is balance of men and women in workplace and what workplace treats women regrading in recruitment and hiring process

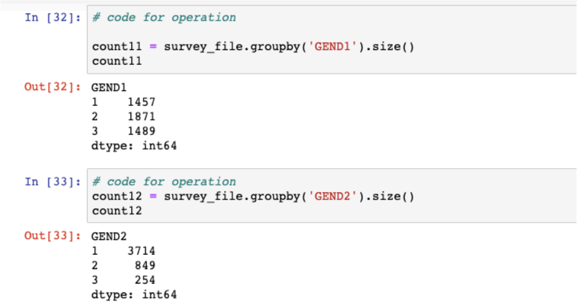
GEND1- What is the balance of men and women in your workplace?

GEND2 - How would you say your workplace treats women when it comes to the recruitment and hiring process?

"GEND1":{"There are more men":1, "There are more women":2, "There is an even mix of men and women":3, "Refused":5}

"GEND2":{"Usually treated fairly in the recruitment and hiring process":1, "Sometimes treated fairly and sometimes treated unfairly":2, "Usually treated unfairly in the recruitment and hiring process":3, "Refused":5}

Outcome:



```
In [32]: # code for operation

         count11 = survey_file.groupby('GEND1').size()
         count11

Out[32]: GEND1
         1    1457
         2    1871
         3    1489
         dtype: int64

In [33]: # code for operation
         count12 = survey_file.groupby('GEND2').size()
         count12

Out[33]: GEND2
         1    3714
         2     849
         3     254
         dtype: int64
```

In 'GEND1' 1871 people are saying there are more women

In 'GEND2' 3714 people saying- "Usually treated fairly in the recruitment and hiring process"

EDA-8:

operation name: Here ploting 'HARASS1' and 'HARASS3' columns survey response because we would like to see sexual harassment in work place**.**

HARASS1- As far as you know, would you say sexual harassment is a big problem, a small problem or not a problem in your workplace?

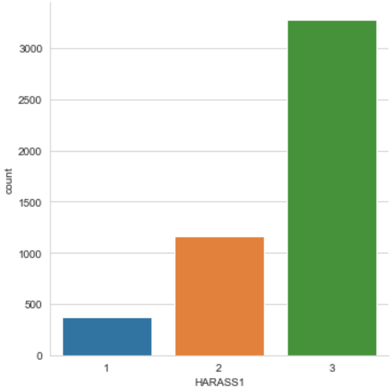
HARASS2- Overall, would you say sexual harassment is a big problem, a small problem or not a problem in jobs in the industry where you work?

All the values:

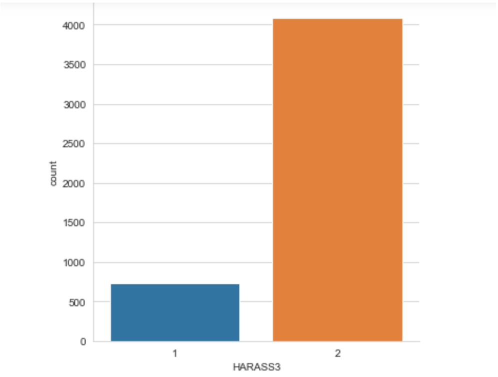
"HARASS1":{"Big problem":1, "Small problem":2, "Not a problem":3, "Refused":5},

"HARASS3":{"Yes, I have experienced sexual harassment at work":1, "No, I have not experienced sexual harassment at work":2, "Refused":5},

Outcome:



In 'HARASS1'- As we can see most individual said it's not a problem.



In 'HARASS3' - Here we see majority of individual said that they have not experienced sexual harassment at work

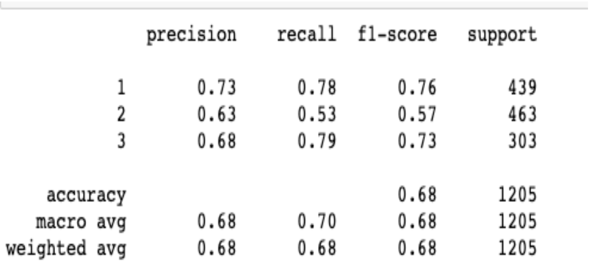
V: IMPLEMENTING ALGORITHM

First Algorithm:

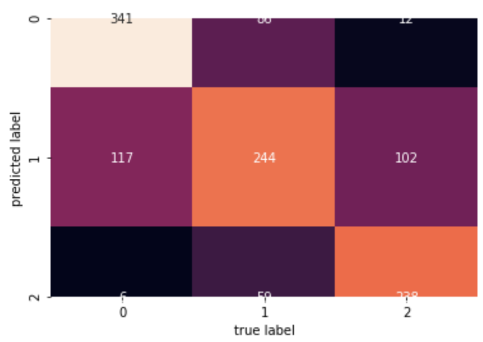
Here we would like to find out if discrimination against women is a major or minor problem or not problem. We would like to find out if women face discrimination in stem field or in a tech job is a major or minor or not problem. We are using Naïve bayes algorithm for this.

So, first we selected some of features. For example, we selected one feature called 'GEND2' which is giving us information about how their workplace treating women when it comes to the recruitment and hiring process and based on that we would like to find out the label called 'tech3' which has 3 classes class 1- A major problem, class 2- a minor problem and class 3- not a problem. So tech3 label finding out the discrimination against women based on some features on the survey dataset. First, we split the dataset into training and testing sets and for training set, we kept 75% of the data and rest we kept for testing set. Next, we ran our algorithm. we got the accuracy on the testing set is 68% which is not bad as we can tell on the unknown and unseen data the accuracy of naive bayes model giving the accuracy rate is about 68%. For precision we got 73% accuracy for class-1, 63% accuracy for class-2 and 68% accuracy for class-3. F1 score is what percent positive prediction were correct. We e got 76% for class-1, 57% for class-2 and 73% for class-3 for F1 score. And recall is how much positive case we did catch. we found for class-1 we got 78%, class-2 we got 53% and class-3 we got 79%. So, we see class-1 (major problem) we see it has most positive accuracy rate which is 76%. We see for class-3 the positive rate we did catch is 79% which is higher than class-2 but f1-score which is accuracy of positive rate for class03 is 73%, so we 6% was incorrect. 6% of the class-3 label given positive case where it supposed to be in other class label. In confusion matrix we see the most data points which we predicted and true label is class-1. we have 341 data points classified correctly. So, we can see class-1 (a major problem) has more effect on our model. we can tell that based on our naive bayes algorithm is that discrimination against women is a major problem. As class-1 has more data points than other classes.

Here we would like to print classification report and confusion matrix:



```
              precision    recall  f1-score   support

           1       0.73      0.78      0.76       439
           2       0.63      0.53      0.57       463
           3       0.68      0.79      0.73       303

    accuracy                           0.68      1205
   macro avg       0.68      0.70      0.68      1205
weighted avg       0.68      0.68      0.68      1205
```

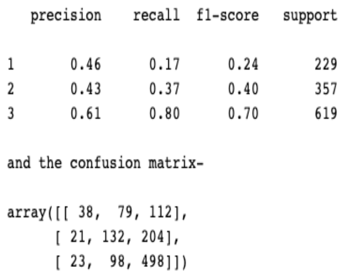


Second algorithm:
Here, we would like to see if women are less interested in science, technology, engineering and math than man and would like to find if it's a major, minor or not a problem.

As this is a classification problem so I used decision tree classifier to find out 3 classes which are major, minor and not a problem in finding women being interested in some of stem field.

First, we selected some of feature and then split the dataset into training and testing algorithm. After implementing algorithm and running the algorithm with different kind of max_depth we got the accuracy 55% which is pretty low. We found high precision score we got for class-3, high recall score we got for class-3 and high f1-score we got for class-3 too. So, the class-3 label is telling us women are not less interested in science, technology, engineering and math than man. we found 70% of the true positive accuracy rate for class-3. we found in the confusion matrix 498 data points in the class-3 label. Even though our algorithm mismatching so many points it's still telling us that class-3 has the most higher data points which representing that women are not less interested in science, technology and math.

Next, we would like to print classification report and confusion matrix:

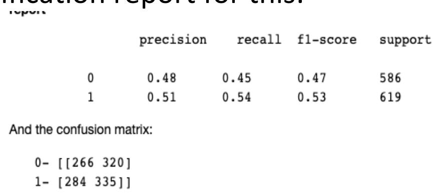


```
        precision    recall  f1-score   support

1         0.46      0.17      0.24       229
2         0.43      0.37      0.40       357
3         0.61      0.80      0.70       619

and the confusion matrix-

array([[ 38,  79, 112],
       [ 21, 132, 204],
       [ 23,  98, 498]])
```

Third algorithm:
Here we would like to find out if women behind in stem field. We chose the k-means clustering to do this.

At first, we created a new column which is called behind stem field which has 2 labels and we would like to use k-mean clustering to see how this true label and k-mean label differs. we scaled our data set and then we divided our dataset into training and testing and then train the algorithm. We got 50% accuracy. We see k-means algorithm captured 50% of the data accurately. We found in the confusion matrix for class-0 we got 266 correct points and for class-1 we got 335 correct points. This k-means algorithm showing that 335 data points telling us female behind in stem field. Where 266 data points showing us that female is not behind in stem field. Next, we got most correct positive prediction for label 0 is 47% and for label-1 is 53%. So, k-means trying to tell that women are behind in stem field based on some features.

Here is the confusion matrix and classification report for this:

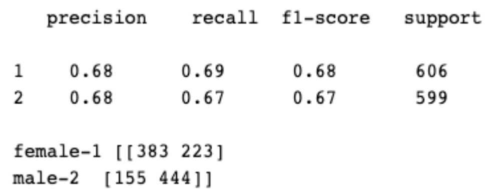


```
report

              precision    recall  f1-score   support

         0       0.48      0.45      0.47       586
         1       0.51      0.54      0.53       619

And the confusion matrix:

   0- [[266 320]
   1- [284 335]]
```

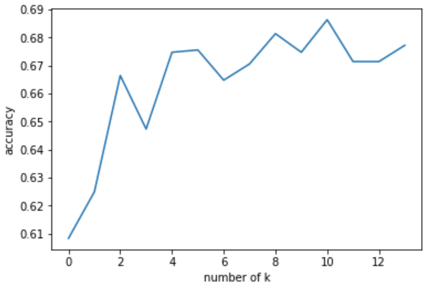
Fourth algorithm:
Here we would like to see if education, income, stem_degree and occupation has more effect on female or male. For this I chose k-nearest neighbor classifier.

First, we split the dataset into training and testing set. Then we used the k-neighbors classifier to see how algorithm performing on the unknown dataset. We iterate through many neighbors k = 1 to 15 to see which one is doing best. We found when k=11 the accuracy we got 69%. We found from the confusion matrix that total 383 class label- female got correctly predicted where 444 data points got correctly predicted. Here for f1-score total correctly predicted points is about 67% for male class and for 68% for female class correctly predicted. We can see f1-score showing those income, stem_degree and occupation has slightly more effect on male than female. But we can say that male and female both having almost equally affected by those features. We can say man and women both are educated and both are earning.

Next, we would like to print out classification report and k-accuracy plot.



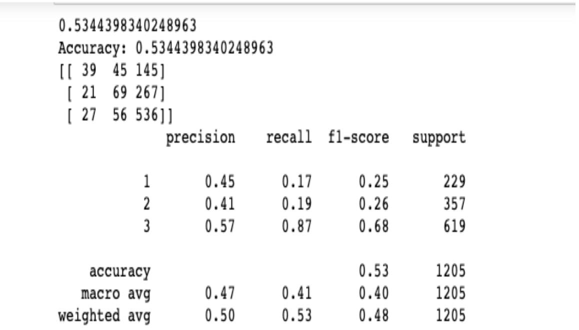
```
      precision    recall  f1-score   support

1       0.68       0.69      0.68       606
2       0.68       0.67      0.67       599

female-1 [[383 223]
male-2   [155 444]]
```



Fifth algorithm:
Here we would like to find out if discrimination against women is a major or minor problem or not problem. We would like to find out if women face discrimination in stem field or tech job is a major or minor or not problem. Here We chose logistic regression to analyze this finding.

First, we split the dataset into training and testing set. After implementing the algorithm, we found the accuracy is about 53% which much lower than the naive bayes algorithm. This logistic regression has more effect and accuracy on the label-3 which showing discrimination against women is not problem. As the logistic regression giving less accuracy than naive bayes we would like to choose naive bayes instead of logistic regression. And logistic regression is good for binary outcome and in our case the labels are not binary outcome. So, we would like to discard this algorithm and stick on the naive bayes which predicted that discrimination against women is a major problem.

Next, we would like to see classification report-



```
0.5344398340248963
Accuracy: 0.5344398340248963
[[ 39  45 145]
 [ 21  69 267]
 [ 27  56 536]]
              precision    recall  f1-score   support

           1       0.45      0.17      0.25       229
           2       0.41      0.19      0.26       357
           3       0.57      0.87      0.68       619

    accuracy                           0.53      1205
   macro avg       0.47      0.41      0.40      1205
weighted avg       0.50      0.53      0.48      1205
```

VI. CONCLUSION AND RECOMMENDATIONS

Hypothesis we made:

- hypothesis 1: Discrimination against women is a major problem and women face discrimination in stem or tech field
- hypothesis 2: Women are less interested in math, science and engineering field is a major problem
- hypothesis 3: Females are behind in stem field
- hypothesis 4: education, income, stem_degree and occupation has more effect on male than female.

After implementing several algorithm we found-

For hypothesis-1: Discrimination against women is a major problem in stem or tech field based on the naïve bayes algorithm. So, our hypothesis is correct.

For hypothesis- 2: Women are not less interested in science, technology or math based on decision tree classifier. So, we reject the hypothesis-2.

For hypothesis-3: Here we found that women are behind in stem field based on the k-means clustering. So, we accept the hypothesis-3.

For hypothesis-4: Here we found that education, income and occupation has same effect on both male and female. So, we reject the hypthosis-4.

So, based on the hypothesis we can say that women are behind in stem field and they face discrimination in tech or stem field. And we also find out that women are interested in stem subjects and women also earning same as man. So, as women face discrimination in stem or tech field, we need to raise awareness to stop discriminate against women. Because of discrimination and harassment women are behind in stem field even though they like math, science and technology. Women are working outside almost same rate as man but in the field of tech and math we see less women because of the discrimination they face. So, we need to inspire more women to come stem field and stop the discrimination against women in school and job.