Customer Segmentation Classification

Violet Huang: Data Preprocess, Feature Encoding, Data Exploration, Data Analysis
April Wang: Feature Encoding, Logistic Regression Model, Random Forest Model
Zenthia Song: Decision Tree Classifier Model, K Nearest Neighbor Model, Model Comparison

Overview

- TASK
  - We aim to train a multi-class classification model to divide new customers into existing segments (A, B, C, D) for a automobile company to predict the potential groups of new customers.
- DATASET
  - https://www.kaggle.com/datasets/kaushiksuresh147/customer-segmentation?datasetId=841888&sortBy=voteCount
- METHOD
  - Logistic Regression, Random Forest, Decision Tree, and K Nearest Neighbours

Summary Data

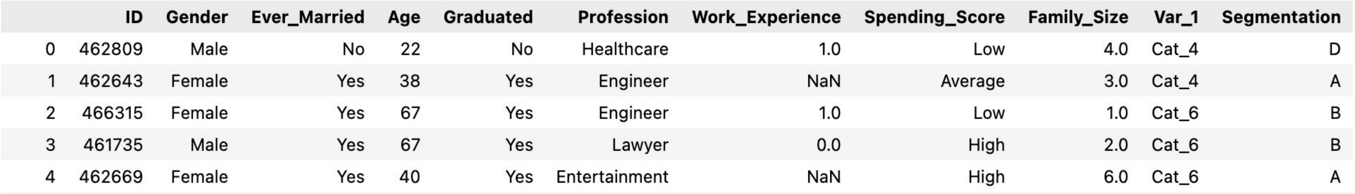
8068 training data set

2627 testing data set

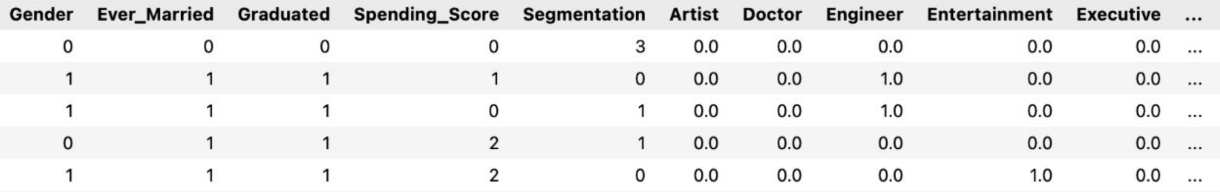
Preprocessing

7308 training data set

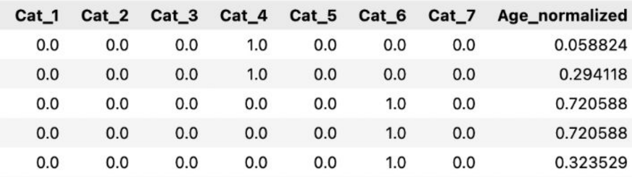
2355 testing data set



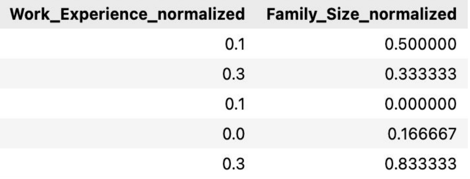
	ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1	Segmentation
0	462809	Male	No	22	No	Healthcare	1.0	Low	4.0	Cat_4	D
1	462643	Female	Yes	38	Yes	Engineer	NaN	Average	3.0	Cat_4	A
2	466315	Female	Yes	67	Yes	Engineer	1.0	Low	1.0	Cat_6	B
3	461735	Male	Yes	67	Yes	Lawyer	0.0	High	2.0	Cat_6	B
4	462669	Female	Yes	40	Yes	Entertainment	NaN	High	6.0	Cat_6	A



Gender	Ever_Married	Graduated	Spending_Score	Segmentation	Artist	Doctor	Engineer	Entertainment	Executive	...
0	0	0	0	3	0.0	0.0	0.0	0.0	0.0	...
1	1	1	1	0	0.0	0.0	1.0	0.0	0.0	...
1	1	1	0	1	0.0	0.0	1.0	0.0	0.0	...
0	1	1	2	1	0.0	0.0	0.0	0.0	0.0	...
1	1	1	2	0	0.0	0.0	0.0	1.0	0.0	...

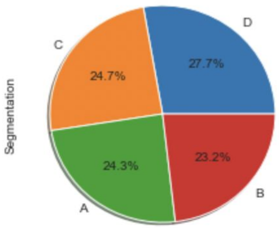


Cat_1	Cat_2	Cat_3	Cat_4	Cat_5	Cat_6	Cat_7	Age_normalized
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.058824
0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.294118
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.720588
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.720588
0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.323529

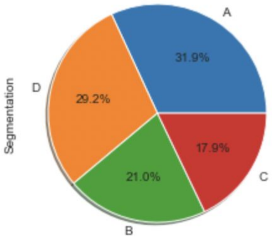


Work_Experience_normalized	Family_Size_normalized
0.1	0.500000
0.3	0.333333
0.1	0.000000
0.0	0.166667
0.3	0.833333

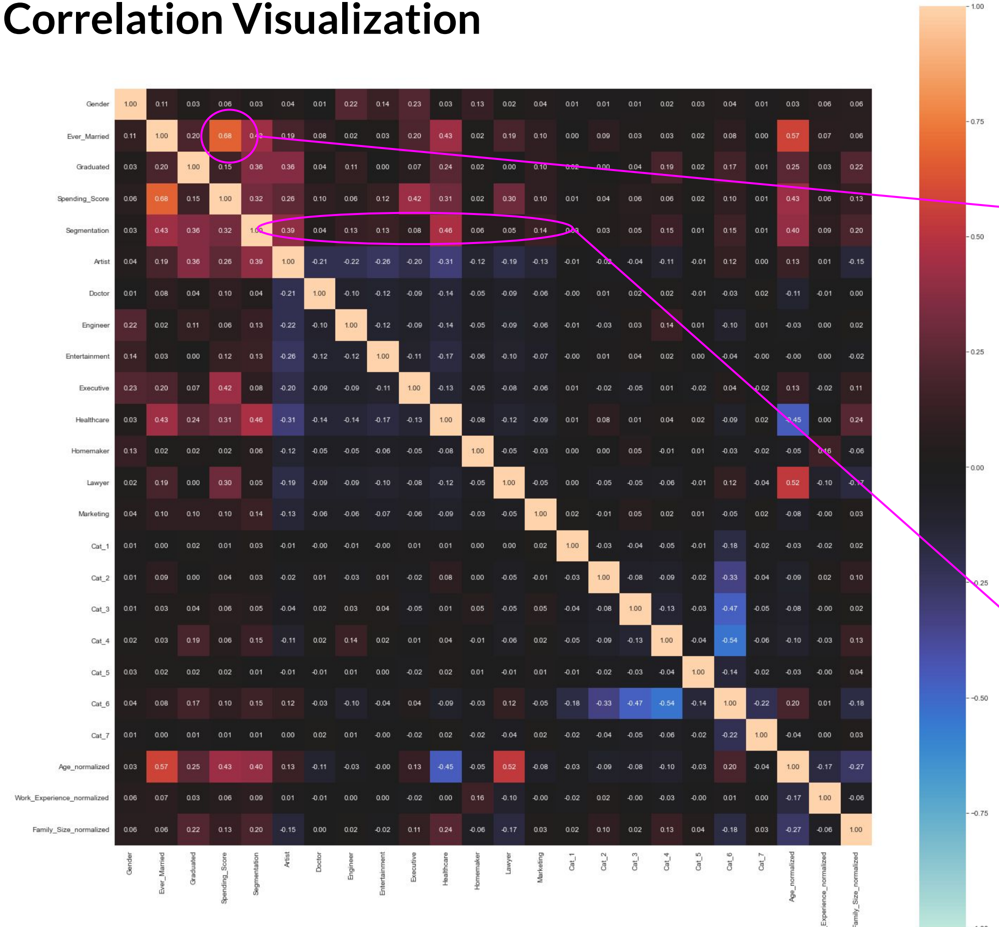
```
D     2127
C     1896
A     1867
B     1779
Name: Segmentation, dtype: int64

<AxesSubplot:ylabel='Segmentation'>
```



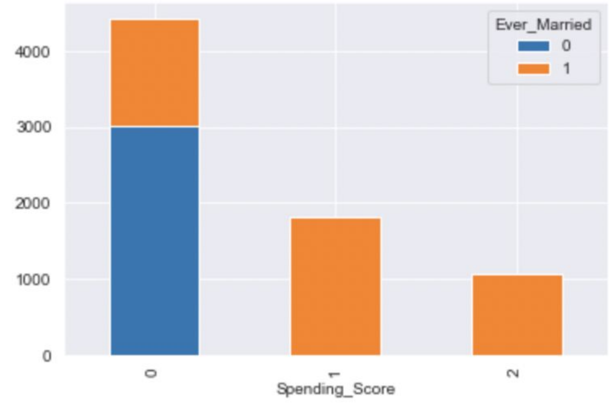
```
A     794
D     726
B     523
C     445
Name: Segmentation, dtype: int64

<AxesSubplot:ylabel='Segmentation'>
```

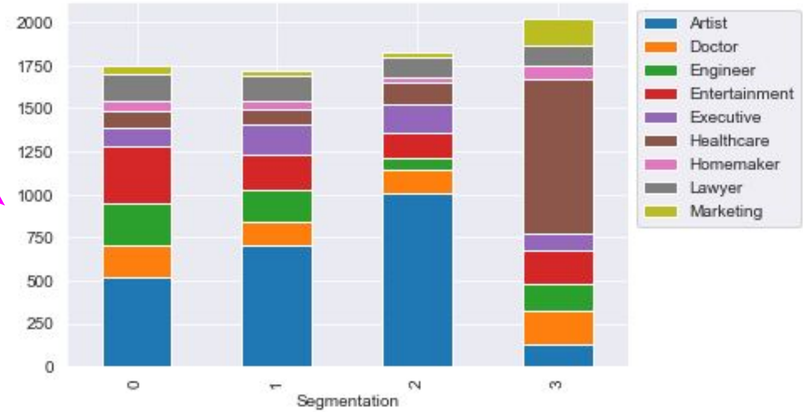


Correlation Visualization



	Gender	Ever_Married	Graduated	Spending_Score	Segmentation	Artist	Doctor	Engineer	Entertainment	Executive	Healthcare	Homemaker	Lawyer	Marketing	Cat_1	Cat_2	Cat_3	Cat_4	Cat_5	Cat_6	Cat_7	Age_normalized	_Experience_normalized	amily_Size_normalized
Gender	1.00	0.11	0.03	0.06	0.03	0.04	0.01	0.22	0.14	0.23	0.03	0.13	0.02	0.04	0.01	0.01	0.01	0.02	0.03	0.04	0.01	0.03	0.06	0.06
Ever_Married	0.11	1.00	0.20	0.68	0.43	0.19	0.08	0.02	0.03	0.20	0.43	0.02	0.19	0.10	0.00	0.09	0.03	0.03	0.02	0.08	0.00	0.57	0.07	0.06
Graduated	0.03	0.20	1.00	0.15	0.36	0.36	0.04	0.11	0.00	0.07	0.24	0.02	0.00	0.10	0.02	0.00	0.04	0.19	0.02	0.17	0.01	0.25	0.03	0.22
Spending_Score	0.06	0.68	0.15	1.00	0.32	0.26	0.10	0.06	0.12	0.42	0.31	0.02	0.30	0.10	0.01	0.04	0.06	0.06	0.02	0.10	0.01	0.43	0.06	0.13
Segmentation	0.03	0.43	0.36	0.32	1.00	0.39	0.04	0.13	0.13	0.08	0.46	0.06	0.05	0.14	0.03	0.03	0.05	0.15	0.01	0.15	0.01	0.40	0.09	0.20
Artist	0.04	0.19	0.36	0.26	0.39	1.00	-0.21	-0.22	-0.26	-0.20	-0.31	-0.12	-0.19	-0.13	-0.01	-0.02	-0.04	-0.11	-0.01	0.12	0.00	0.13	0.01	-0.15
Doctor	0.01	0.08	0.04	0.10	0.04	-0.21	1.00	-0.10	-0.12	-0.09	-0.14	-0.05	-0.09	-0.06	-0.00	0.01	0.02	0.02	-0.01	-0.03	0.02	-0.11	-0.01	0.00
Engineer	0.22	0.02	0.11	0.06	0.13	-0.22	-0.10	1.00	-0.12	-0.09	-0.14	-0.05	-0.09	-0.06	-0.01	-0.03	0.03	0.14	0.01	-0.10	0.01	-0.03	0.00	0.02
Entertainment	0.14	0.03	0.00	0.12	0.13	-0.26	-0.12	-0.12	1.00	-0.11	-0.17	-0.06	-0.10	-0.07	-0.00	0.01	0.04	0.02	0.00	-0.04	-0.00	-0.00	0.00	-0.02
Executive	0.23	0.20	0.07	0.42	0.08	-0.20	-0.09	-0.09	-0.11	1.00	-0.13	-0.05	-0.08	-0.06	0.01	-0.02	-0.05	0.01	-0.02	0.04	-0.02	0.13	-0.02	0.11
Healthcare	0.03	0.43	0.24	0.31	0.46	-0.31	-0.14	-0.14	-0.17	-0.13	1.00	-0.08	-0.12	-0.09	0.01	0.08	0.01	0.04	0.02	-0.09	0.02	-0.45	0.00	0.24
Homemaker	0.13	0.02	0.02	0.02	0.06	-0.12	-0.05	-0.05	-0.06	-0.05	-0.08	1.00	-0.05	-0.03	0.00	0.00	0.05	-0.01	0.01	-0.03	-0.02	-0.05	0.16	-0.06
Lawyer	0.02	0.19	0.00	0.30	0.05	-0.19	-0.09	-0.09	-0.10	-0.08	-0.12	-0.05	1.00	-0.05	0.00	-0.05	-0.05	-0.06	-0.01	0.12	-0.04	0.52	-0.10	-0.17
Marketing	0.04	0.10	0.10	0.10	0.14	-0.13	-0.06	-0.06	-0.07	-0.06	-0.09	-0.03	-0.05	1.00	0.02	-0.01	0.05	0.02	0.01	-0.05	0.02	-0.08	-0.00	0.03
Cat_1	0.01	0.00	0.02	0.01	0.03	-0.01	-0.00	-0.01	-0.00	0.01	0.01	0.00	0.00	0.02	1.00	-0.03	-0.04	-0.05	-0.01	-0.18	-0.02	-0.03	-0.02	0.02
Cat_2	0.01	0.09	0.00	0.04	0.03	-0.02	0.01	-0.03	0.01	-0.02	0.08	0.00	-0.05	-0.01	-0.03	1.00	-0.08	-0.09	-0.02	-0.33	-0.04	-0.09	0.02	0.10
Cat_3	0.01	0.03	0.04	0.06	0.05	-0.04	0.02	0.03	0.04	-0.05	0.01	0.05	-0.05	0.05	-0.04	-0.08	1.00	-0.13	-0.03	-0.47	-0.05	-0.08	-0.00	0.02
Cat_4	0.02	0.03	0.19	0.06	0.15	-0.11	0.02	0.14	0.02	0.01	0.04	-0.01	-0.06	0.02	-0.05	-0.09	-0.13	1.00	-0.04	-0.54	-0.06	-0.10	-0.03	0.13
Cat_5	0.03	0.02	0.02	0.02	0.01	-0.01	-0.01	0.01	0.00	-0.02	0.02	0.01	-0.01	0.01	-0.01	-0.02	-0.03	-0.04	1.00	-0.14	-0.02	-0.03	-0.00	0.04
Cat_6	0.04	0.08	0.17	0.10	0.15	0.12	-0.03	-0.10	-0.04	0.04	-0.09	-0.03	0.12	-0.05	-0.18	-0.33	-0.47	-0.54	-0.14	1.00	-0.22	0.20	0.01	-0.18
Cat_7	0.01	0.00	0.01	0.01	0.01	0.00	0.02	0.01	-0.00	-0.02	0.02	-0.02	-0.04	0.02	-0.02	-0.04	-0.05	-0.06	-0.02	-0.22	1.00	-0.04	0.00	0.03
Age_normalized	0.03	0.57	0.25	0.43	0.40	0.13	-0.11	-0.03	-0.00	0.13	-0.45	-0.05	0.52	-0.08	-0.03	-0.09	-0.08	-0.10	-0.03	0.20	-0.04	1.00	-0.17	-0.27
Work_Experience_normalized	0.06	0.07	0.03	0.06	0.09	0.01	-0.01	0.00	0.00	-0.02	0.00	0.16	-0.10	-0.00	-0.02	0.02	-0.00	-0.03	-0.00	0.01	0.00	-0.17	1.00	-0.06
Family_Size_normalized	0.06	0.06	0.22	0.13	0.20	-0.15	0.00	0.02	-0.02	0.11	0.24	-0.06	-0.17	0.03	0.02	0.10	0.02	0.13	0.04	-0.18	0.03	-0.27	-0.06	1.00





Model: Logistic Regression

Type of Model	Logistic Regression

Parameter	solver='newton-cg'
	random_state=42

Accuracy	32.4%

Model: Random Forest

Type of Model	Random Forest

Parameter	n_estimators=200
	criterion="entropy"
	max_depth=5
	random_state=42

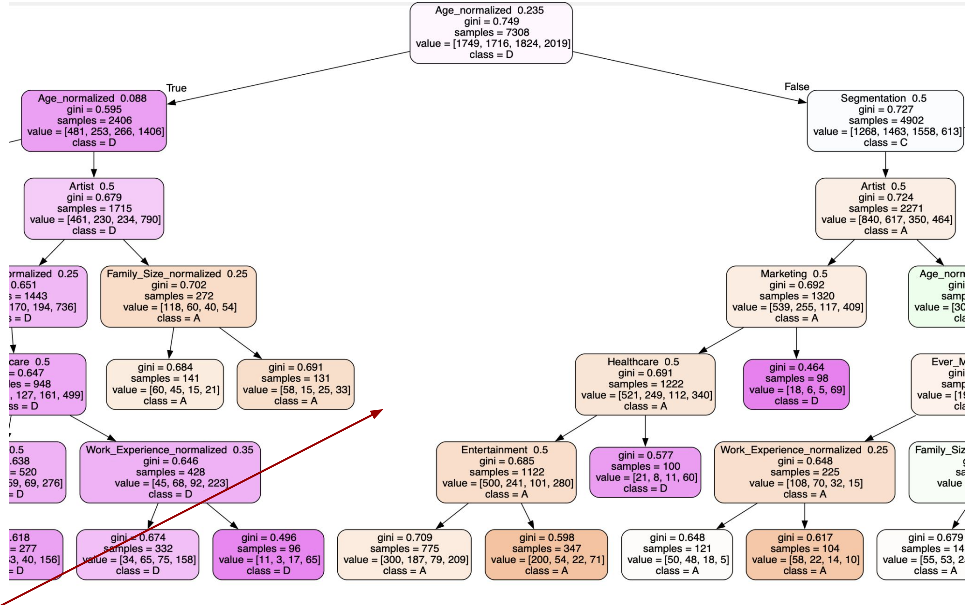
Accuracy	33.9%

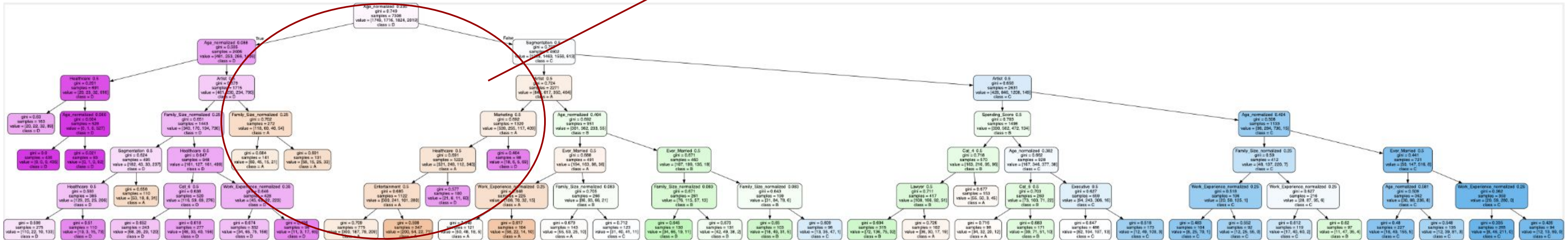
Model: Decision Tree

Type of Model	Decision Tree

Parameter	criterion='gini'
	random_state=42
	min_samples_leaf=90
	max_step=6

Accuracy	34.5%

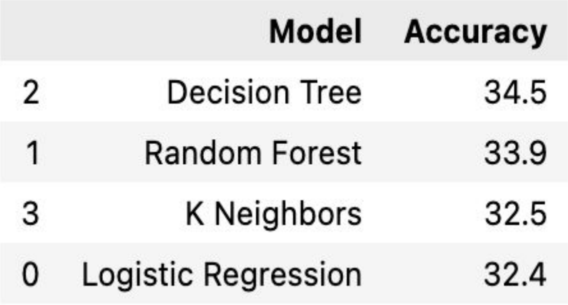




Model: K Neighbors

Type of Model	K Neighbors
Parameter	radius=1.44
Accuracy	32.5%

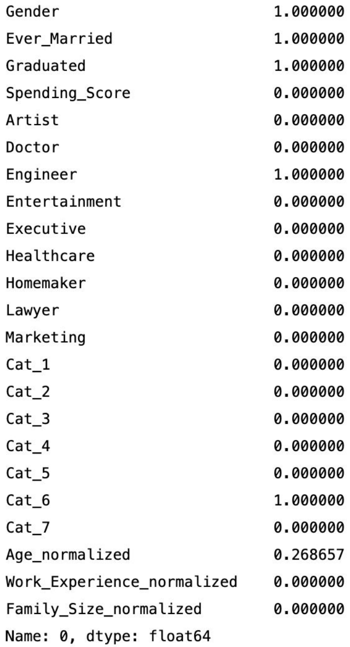
Comparing Model Performance



	Model	Accuracy
2	Decision Tree	34.5
1	Random Forest	33.9
3	K Neighbors	32.5
0	Logistic Regression	32.4

- The model with the highest accuracy is Decision Tree because there are a large number of categories in the feature set.
- Based on these four accuracy scores, we found that 33% is an average score, which is not ideally high.
- Our RF's accuracy score is lower than our DT's, which potentially mean DT outperforms RF. This is possible because aggregated/ensemble models are not universally better than their "single" counterparts, they are better if and only if the single models suffer from instability (a Machine Learning System that produces large differences in generalization patterns when small changes are made to its initial conditions).

Error Analysis



```
Gender                        1.000000
Ever_Married                  1.000000
Graduated                     1.000000
Spending_Score                0.000000
Artist                        0.000000
Doctor                        0.000000
Engineer                      1.000000
Entertainment                 0.000000
Executive                     0.000000
Healthcare                    0.000000
Homemaker                     0.000000
Lawyer                        0.000000
Marketing                     0.000000
Cat_1                         0.000000
Cat_2                         0.000000
Cat_3                         0.000000
Cat_4                         0.000000
Cat_5                         0.000000
Cat_6                         1.000000
Cat_7                         0.000000
Age_normalized                0.268657
Work_Experience_normalized    0.000000
Family_Size_normalized        0.000000
Name: 0, dtype: float64
```

Predicted Segmentation: B

Actual Segmentation: A

Error Analysis

- On the whole, the accuracy of the four models are not high.
- Potential reasons:
  - Lack of data: the amount of data may not capture the complexity of the problem
    - Solution: acquiring more data of the characteristics of old customers
  - Non-representative data: the newly acquired customers may behave inherently different from the old customers
    - Solution: using labeled new customer data to predict unlabeled new customer data
  - Irrelevant Features: We may selected irrelevant features. For Decision Tree model, most of our features have importances smaller than 0.1
    - Solution: Filter out unimportant features. Add other dimension to the data set if possible
  - Algorithm Selection Issue: The choice of four algorithms may also impact the accuracy of the model. Specifically, some more complex algorithms may be better suited this multi-class classification problem.
    - Solution: Deep learning, Naive Bayes