# Machine Learning
# Classification

Ai Paula Couto

# ABSTRACT

This project uses supervised machine learning classification models to assign labels to product titles and automate categorization. Two models are implemented: Multinomial Naive Bayes (MNB) Method and Random Forest (RF) Classifier. These algorithms are chosen because of their efficiency in working with high-dimensional data

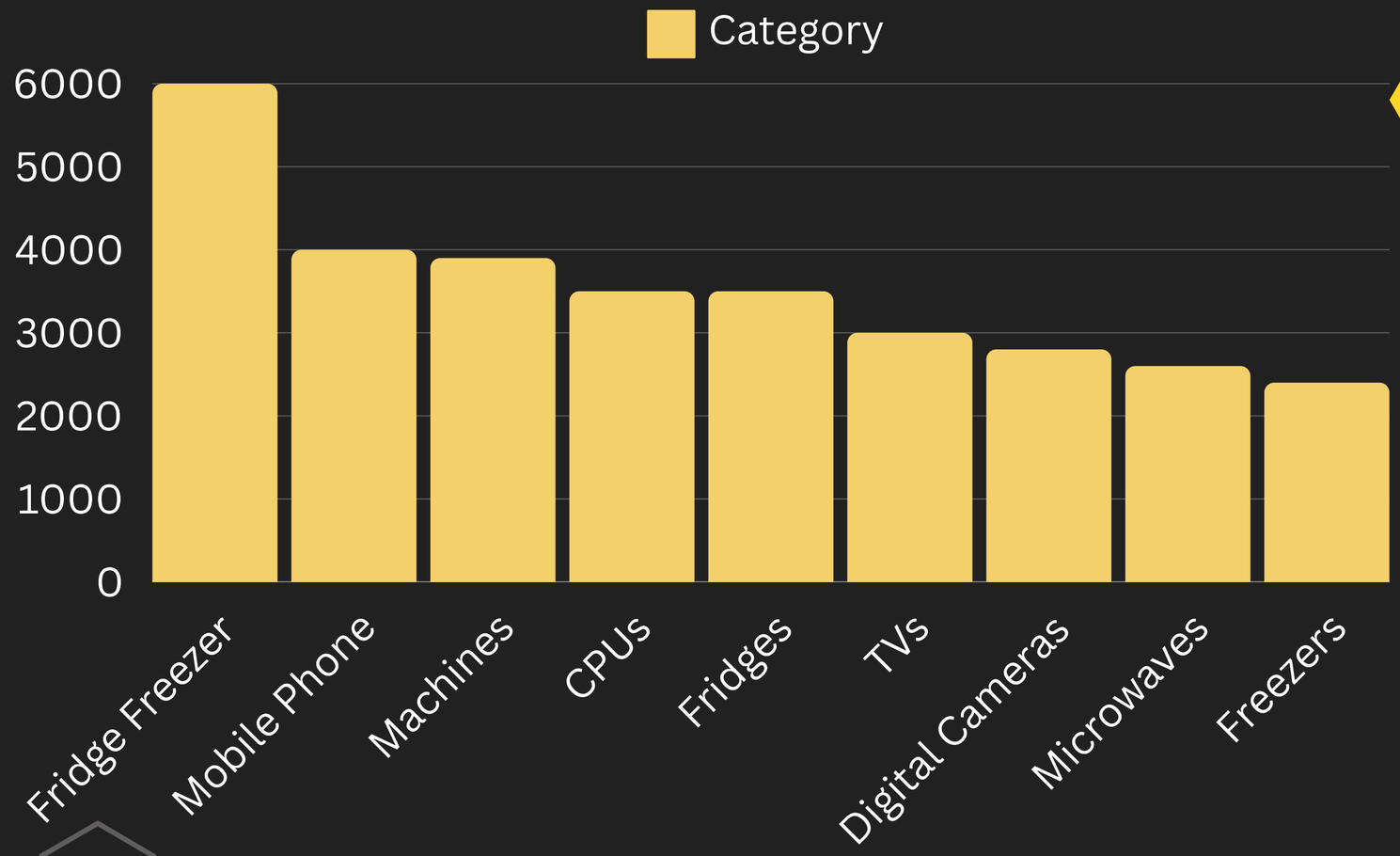
# BACKGROUND

Categorization for products can help e-commerce website provide a more concise list of items to customers by streamlining products under groups based on their name, descriptions, or other identifying information.

Classification algorithms in machine learning allow for automation by using different calculations to predict the data entered. In doing so, customers are presented with not only the items they're looking for as well as similar things but filtering it to ones that are related or similar.

# EXPLORATORY DATA ANALYSIS

Data source: Kaggle – Product Classification and Clustering



# DATA PREPARATION



- Replace all letters with lower strings
- Removal of white space and other special characters
- Vectorization
- Label Encoding

# METHODOLOGY

## Multinomial Naive Bayes

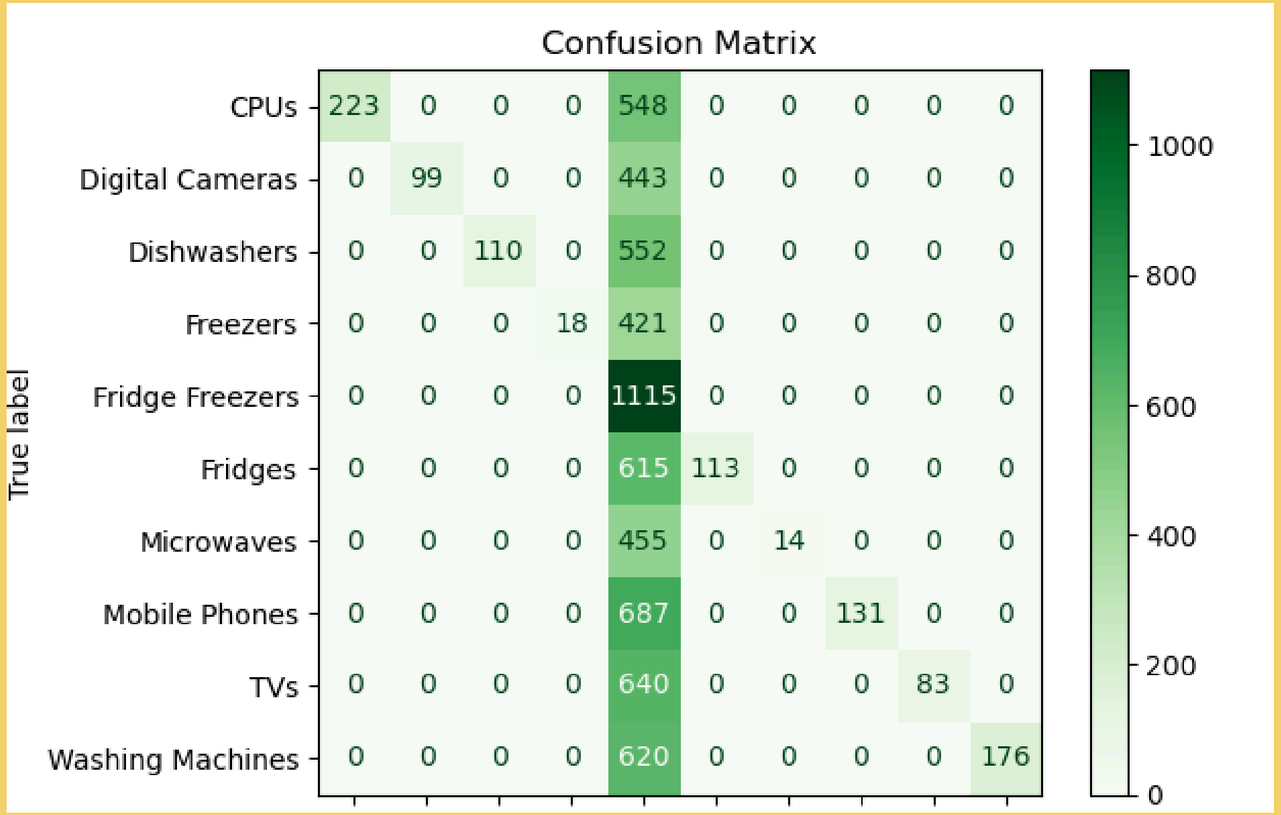
- Probabilistic Model
- Feature Independence
- Adept at handling features that represent counts, like word frequencies in documents.

## Random Forest Classification

- Ensemble or collection of decision trees
- Aggregates the results of each trees by voting
- Random Feature Selection



# ANALYSIS




## Multinomial Naive Bayes Classifier
## Accuracy Score: 29%

```
Classification Report
=====================================================

              precision    recall  f1-score   support

           0       1.00      0.29      0.45       771
           1       1.00      0.18      0.31       542
           2       1.00      0.17      0.28       662
           3       1.00      0.04      0.08       439
           4       0.18      1.00      0.31      1115
           5       1.00      0.16      0.27       728
           6       1.00      0.03      0.06       469
           7       1.00      0.16      0.28       818
           8       1.00      0.11      0.21       723
           9       1.00      0.22      0.36       796

    accuracy                           0.29      7063
   macro avg       0.92      0.24      0.26      7063
weighted avg       0.87      0.29      0.28      7063
```

## Random Forest Classifier
## Accuracy Score: 31 %

```
Classification Report
=====================================================

              precision    recall  f1-score   support

           0       1.00      0.29      0.45       771
           1       1.00      0.18      0.31       542
           2       1.00      0.17      0.28       662
           3       1.00      0.22      0.36       439
           4       0.19      1.00      0.32      1115
           5       1.00      0.16      0.27       728
           6       1.00      0.17      0.29       469
           7       1.00      0.16      0.28       818
           8       1.00      0.11      0.21       723
           9       1.00      0.22      0.36       796

    accuracy                           0.32      7063
   macro avg       0.92      0.27      0.31      7063
weighted avg       0.87      0.32      0.31      7063
```

# CONCLUSION

The accuracy scores between the Multinomial Naive Bayes Classifier and Random Forest Classification models tells us that the latter, ensemble method is a better algorithm for our dataset. Even though accuracy tests are not the be all, end all measure for a model's performance- it is an important indicator of whether the algorithm we used is the best fit for the problem we are trying to solve.

# RECOMMENDATIONS

After reviewing the result of our Multinomial Naive Bayes Classifier, we addressed the imbalance in our dataset with resampling methods. While this helped the performance when we changed the algorithm for RandomForestClassification, the classification report stayed mainly the same, showing high precision with a low recall. For the next iterations of this project, feature selection techniques could be helpful in removing bias in the dataset. Neural Networks classification models should also be considered especially when dealing with imbalanced dataset

# ETHICAL ASSESSMENT

When performing any type of research using data, one of the main ethical considerations is privacy concerns. This is especially important since the information comes from e-commerce websites where people use personal data including banking information that needs to be protected. In my initial inspection of the dataset, I did not see any information that could identify an individual. Another ethical concern that needs to be addressed is bias. Whether intentional or not, the possibility of the occurrence of bias in the dataset could lead to skewed outcomes of the study. In machine learning models, when bias is not removed, then we cannot trust the results of our calculations.

# THANK YOU