Text Classification of FDA Medical Device Recalls

By: Ruth Iang

DATA 606

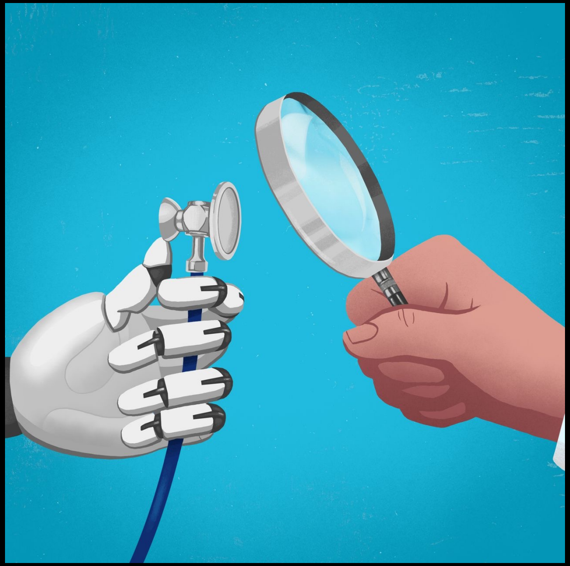
Fall 2023

Introduction

- Medical device are recalled by FDA to protect the public from harm
- Classified into 3 classes
  - Class I-Most severe,
  - Class II-Moderate,
  - Class III- negligible harm
- From year 2012-October 2023

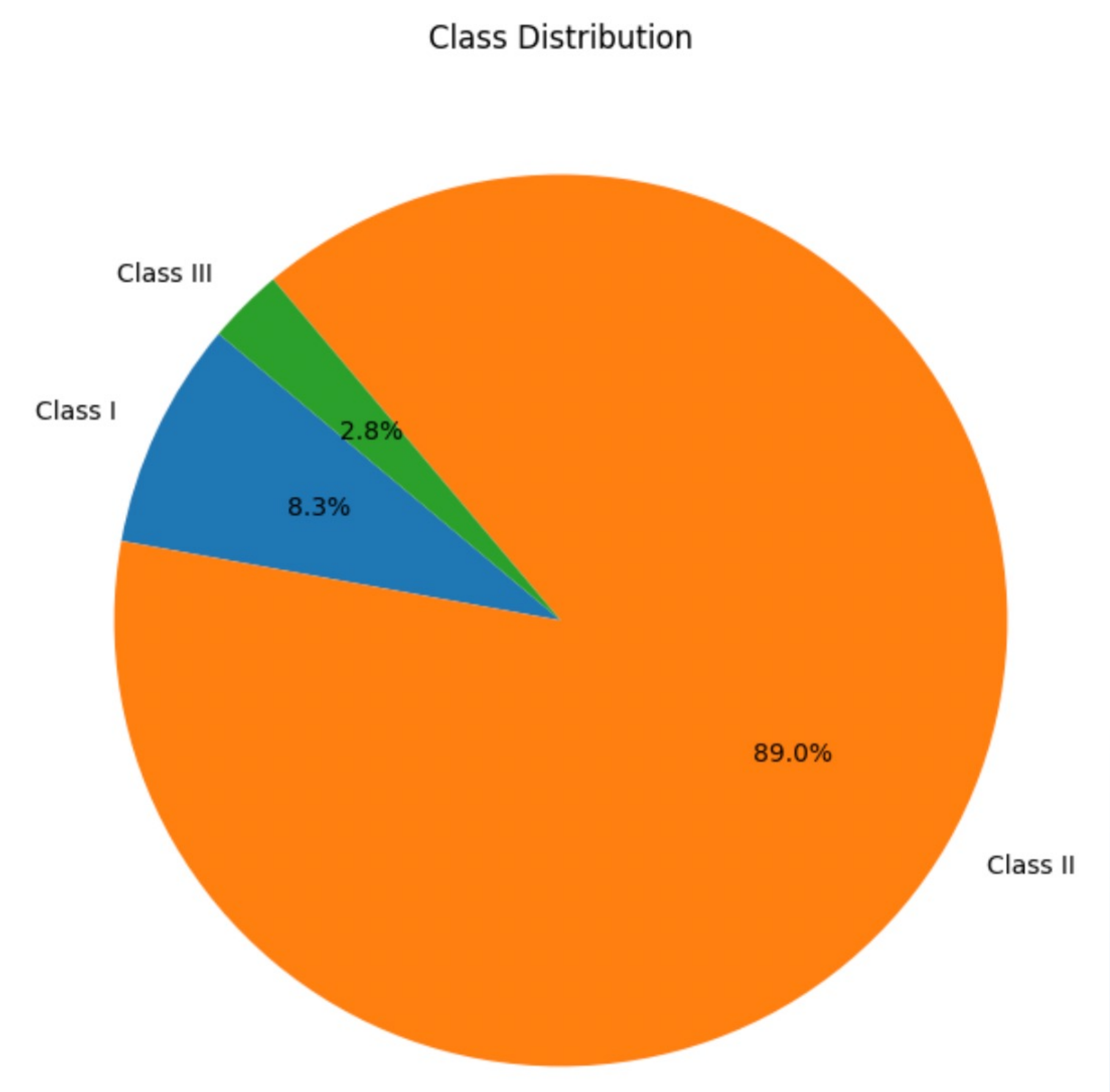
Data Cleaning/ Normalization

- Kept only relevant columns, Event Classification and Reason for Recall
- Dropped nulls and duplicates
- Made all letters/words lowercase
- Tokenized feature variable
- Stop words removal
- Print the stem of the text in the feature variable



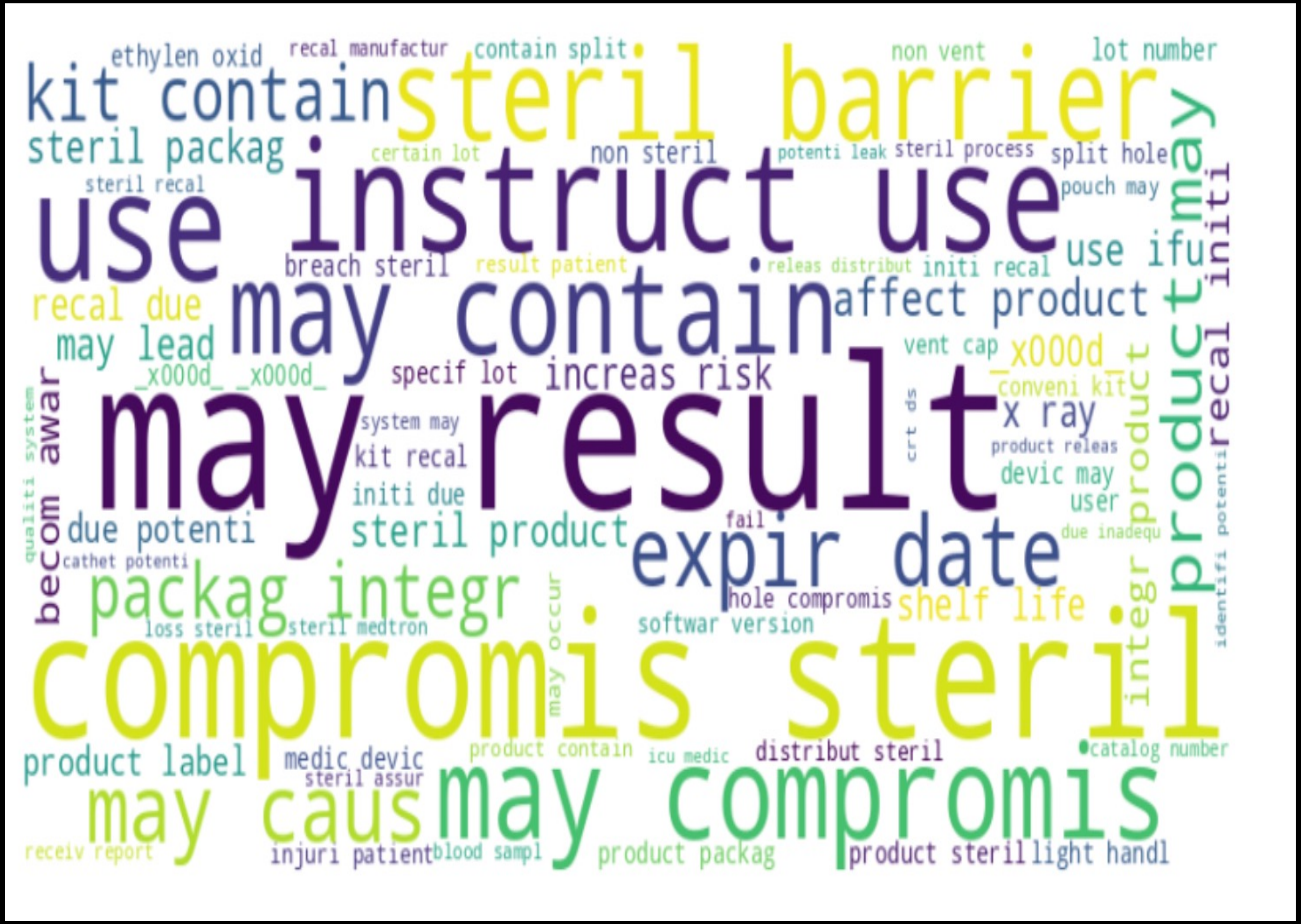
Distribution of classes

- Heavily Imbalanced dataset
- Class II makes up majority of data
- Class III and I are minority classes



Word Cloud

- Bilinear- visualize two words and their importance with each other in a single layout
- Font size shows frequency of words in dataset
- Shows main idea of dataset

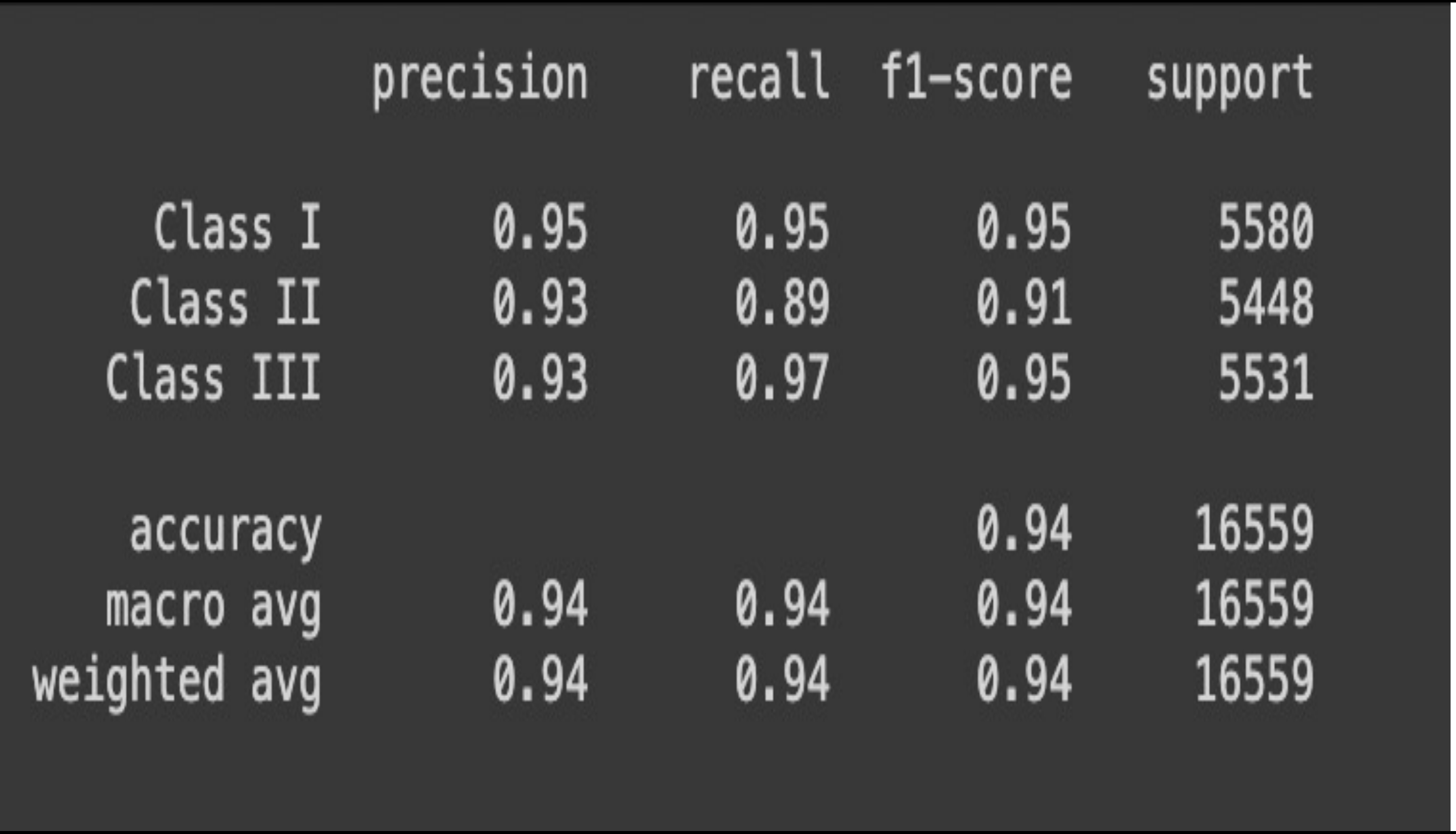


Data Preprocessing

- Used TF-IDF Vectorizer to transform feature variable into numerical column
- Used SMOTE (Synthetic Minority Over-sampling Technique)
  - to treat imbalanced distribution
  - to increase minority class
- Split 80% of data to training, and 20% to test

Multinomial Naïve bayes

Classification Report

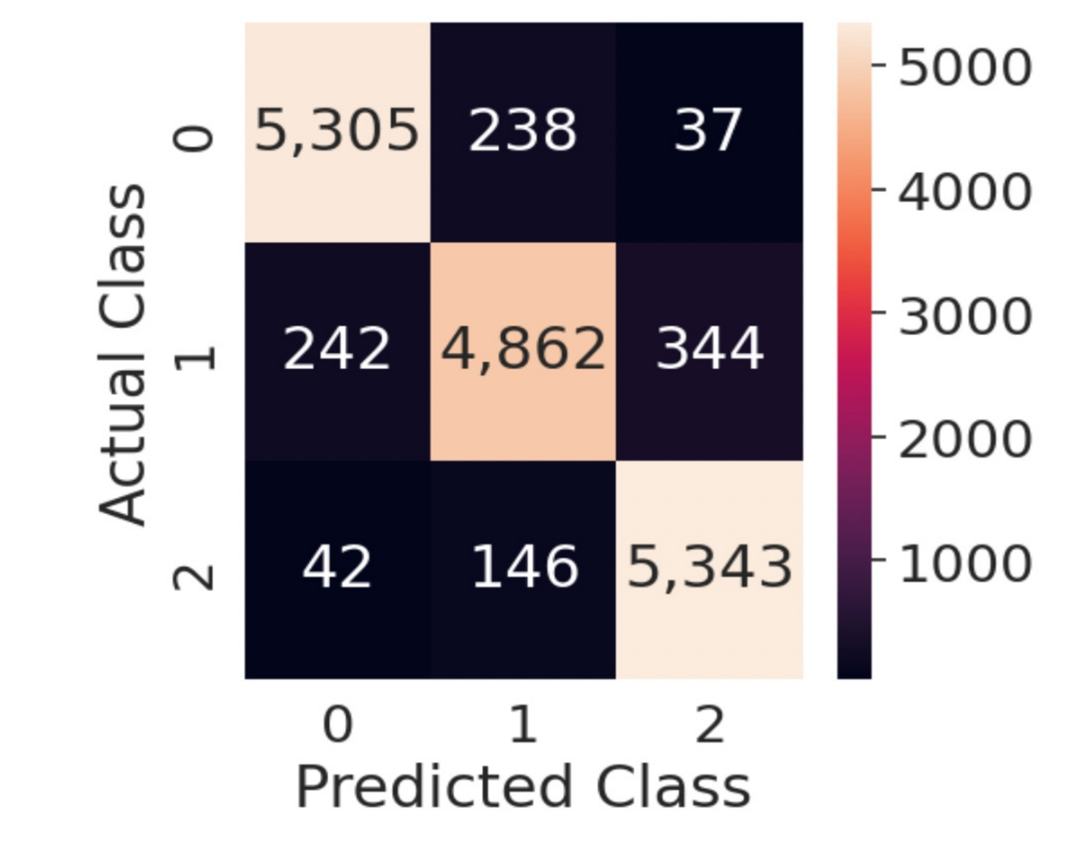


```
             precision    recall  f1-score   support

     Class I       0.95      0.95      0.95      5580
    Class II       0.93      0.89      0.91      5448
   Class III       0.93      0.97      0.95      5531

    accuracy                           0.94     16559
   macro avg       0.94      0.94      0.94     16559
weighted avg       0.94      0.94      0.94     16559
```

Confusion Matrix Heatmap



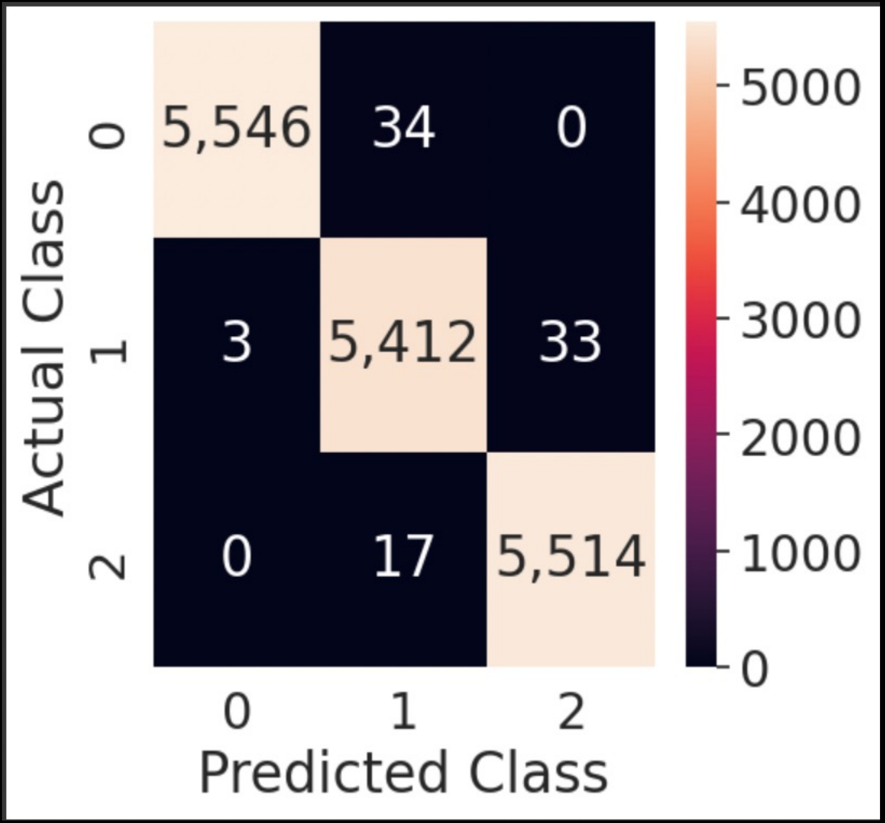
Multinomial Naïve bayes

- Accuracy: .9366
- Precision: 0.937
- F1 Score: 0.936
- Recall: 0.937

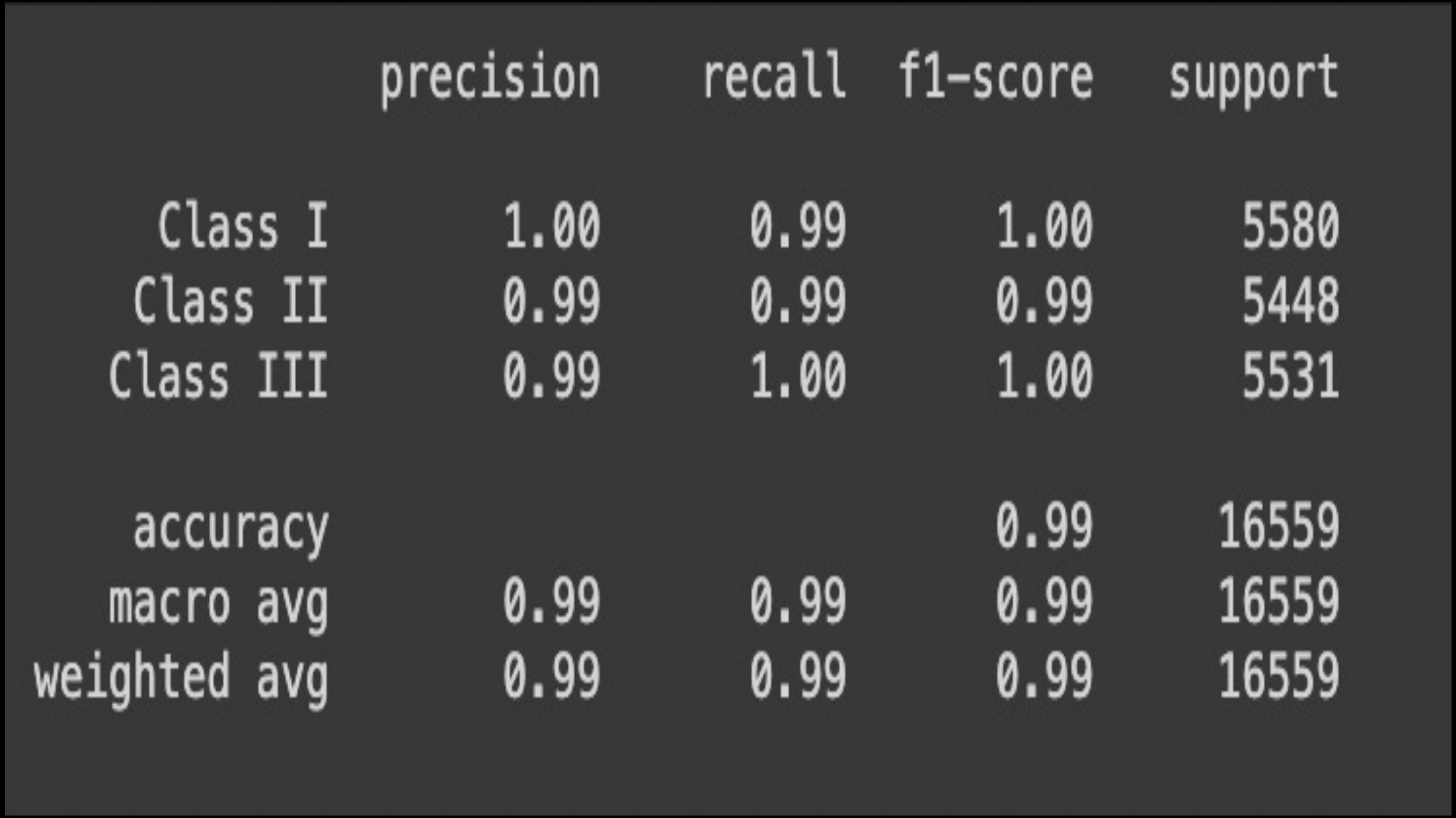
- Hyperparameter Tuning:
  - Used Grid Search CV and cross validation and alpha values for best parameter
  - The best parameter was alpha is 0.1, and accuracy is .954

Random Forest Classifier

Confusion Matrix Heatmap



Classification Report



```
              precision    recall  f1-score   support

     Class I       1.00      0.99      1.00      5580
    Class II       0.99      0.99      0.99      5448
   Class III       0.99      1.00      1.00      5531

    accuracy                           0.99     16559
   macro avg       0.99      0.99      0.99     16559
weighted avg       0.99      0.99      0.99     16559
```

Random Forest Classifier

- Accuracy 0.995
- Precision: 0.937
- F1 Score: 0.995
- Recall: 0.995

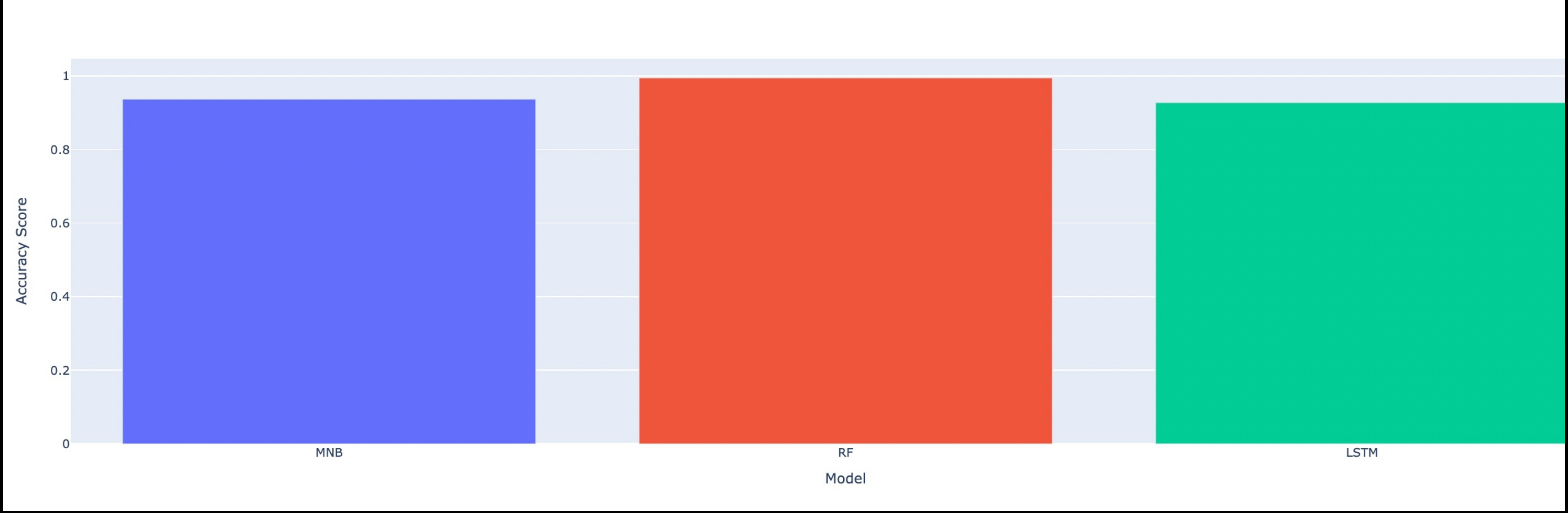
- Hyperparameter Tuning:
  - Used Randomized Search and cross validation to find the best parameters
  - Best Parameters: n_estimators: 100, min_samples_split: 5, min_samples_leaf: 1, max_depth: None
  - Accuracy: 0.995

LSTM Model

- Accuracy .929
- num_epochs = 10
- batch_size = 32

- Validation Loss: 0.248
- Validation Accuracy: 0.928

Models Accuracy Bar Chart



Conclusion

- Random Forest performed the best with accuracy of .995
- Cross Validation- implemented during hyperparameter tuning to prevent overfitting
- SMOTE-to prevent overfitting due to imbalanced dataset
- The Random Forest classifier-reliable to classify FDA medical device recalls faster than manual classification.