ENRON

POI-CLASSIFIER

MACHINE LEARNING

ASHAY PANCHAL

NORTHEASTERN UNIVERSITY

CS 6140
Github: https://github.com/Ashay1301/Enron-POI.git

TABLE OF CONTENTS

ASHAY PANCHAL

INTRODUCTION

Enron Corporation was an American energy, commodities, and services company based in Houston, Texas. It was founded by Kenneth Lay in 1985 as a merger between Lay's Houston Natural Gas and InterNorth, both relatively small regional companies. Before its bankruptcy on December 2, 2001, Enron employed approximately 20,600 staff and was a major electricity, natural gas, communications, and pulp and paper company, with claimed revenues of nearly $101 billion during 2000. Fortune named Enron "America's Most Innovative Company" for six consecutive years.

At the end of 2001, it was revealed that Enron's reported financial condition was sustained by an institutionalized, systematic, and creatively planned accounting fraud, known since as the Enron scandal. Enron has become synonymous with willful corporate fraud and corruption. The scandal also brought into question the accounting practices and activities of many corporations in the United States and was a factor in the enactment of the Sarbanes–Oxley Act of 2002. The scandal also affected the greater business world by causing, together with even larger fraudulent bankruptcy WorldCom, the dissolution of the Arthur Andersen accounting firm, which had been Enron and WorldCom's main auditor for years.

Enron filed for bankruptcy in the Southern District of New York in late 2001 and selected Weil, Gotshal & Manges as its bankruptcy counsel. It ended its bankruptcy in November 2004, pursuant to a court-approved plan of reorganization. A new board of directors changed the name of Enron to, Enron Creditors Recovery Corp., and emphasized reorganizing and liquidating certain operations and assets of the pre-bankruptcy Enron. On September 7, 2006, Enron sold its last remaining subsidiary. It is the largest bankruptcy, due specifically to fraud, of all time.

BACKGROUND

PREVIOUS SOLUTIONS

In the past, researchers and data scientists have used the Enron Email Dataset for various purposes, including sentiment analysis, social network analysis, and email classification. However, the majority of these studies have focused on descriptive analysis and lacked predictive or prescriptive capabilities. Some researchers have attempted to apply machine learning techniques, but there remains a significant untapped potential for more advanced predictive modeling.

DRAWBACKS OF PREVIOUS SOLUTIONS

Existing solutions have limitations when it comes to harnessing the full potential of the Enron Email Dataset. They often fail to predict fraudulent activities or to provide actionable insights for fraud prevention. Additionally, many studies have not adequately addressed the complexities of corporate communication data.

PROBLEM STATEMENT

The Enron Email Dataset, a treasure trove of emails exchanged by employees of the Enron Corporation, is a valuable resource for conducting machine learning analysis, particularly in the areas of natural language processing (NLP) and social network analysis. This project aims to leverage this dataset to extract meaningful insights, patterns, and potentially develop predictive models to better understand and potentially predict corporate misconduct and fraud.

PROBLEM SIGNIFICANCE

The Enron scandal of the early 2000s remains one of the most infamous corporate fraud cases in history. An analysis of the email communications leading up to the scandal could provide significant insights into the precursors and patterns of fraudulent activities. Machine learning models applied to this dataset may help in identifying red flags and anomalies in corporate communications, thus assisting in the prevention of future corporate misconduct.

PROPOSED SOLUTION

ADDRESSING DRAWBACKS

This project aims to address the drawbacks of existing solutions by developing advanced machine learning models and frameworks for the Enron Email Dataset. The focus will be on predictive modeling to identify patterns and anomalies associated with fraudulent activities. Additionally, we will explore the application of NLP techniques to gain a deeper understanding of the sentiment, intent, and content of the emails.

APPROACH

The proposed solution involves the following steps:

1. **Data Preprocessing:** Cleaning and structuring the Enron Email Dataset to prepare it for analysis. This includes text preprocessing, data augmentation, and network data preparation.
2. **Feature Engineering:** Extracting meaningful features from the emails, such as sentiment scores, topic modeling, email network metrics, and temporal patterns.
3. **Machine Learning Models:** Developing and training machine learning models, including classification, clustering, and anomaly detection, to predict fraudulent activities or other significant patterns.
4. **Evaluation:** Evaluating the models' performance using appropriate metrics, such as precision, recall, and, F1 score.
5. **Interpretation and Insights:** Providing actionable insights and visualizations based on the model results to assist in the detection and prevention of corporate misconduct.

DATASET

DESCRIPTION

The Enron Email Dataset contains approximately 500,000 emails generated by employees of Enron Corporation. It includes various types of data, such as text, timestamps, sender and recipient information, and attachment details. The dataset is publicly available and provides a comprehensive archive of corporate communication.

The dataset contains records of 146 people (thus 146 records of for each feature (including missing values)).

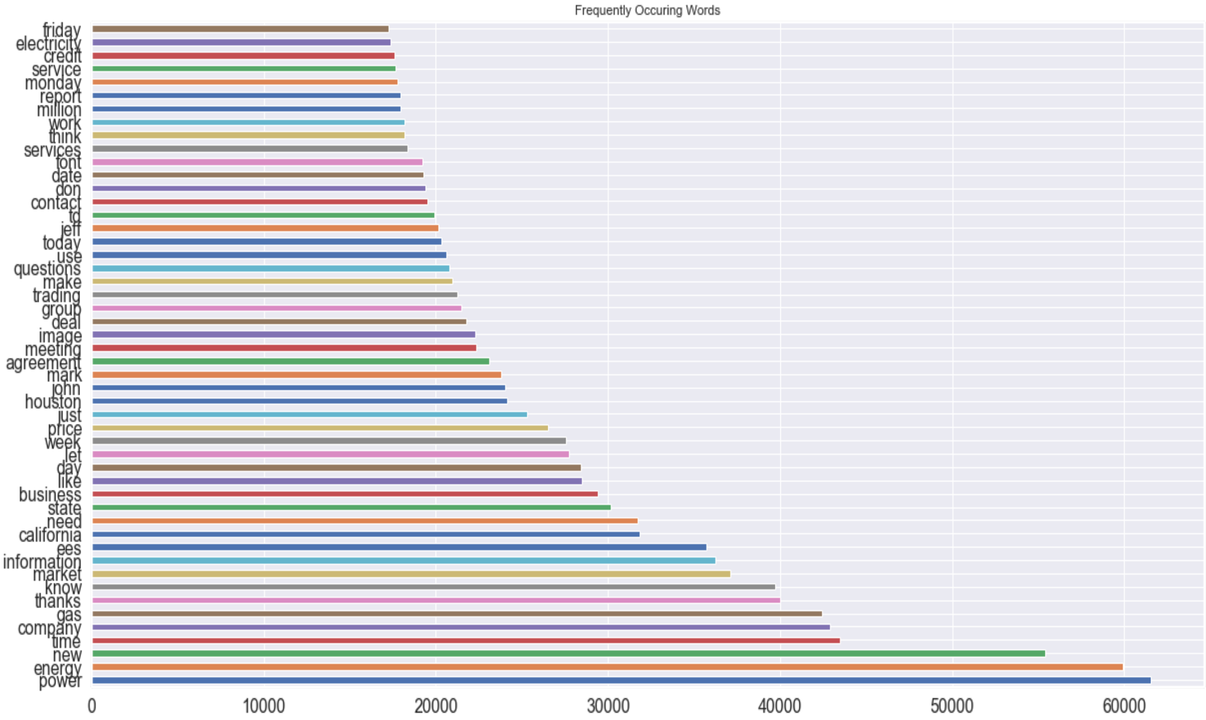
There are 143 records with 20 features and a binary classification "poi".

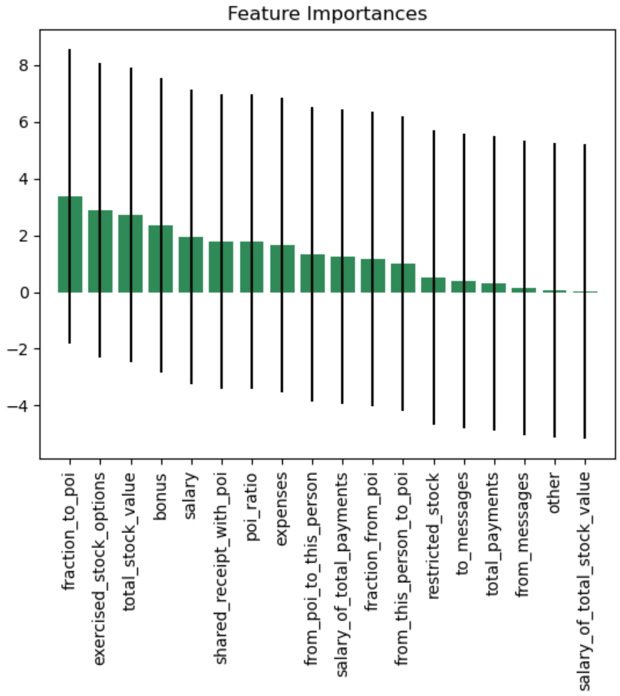
The 146 records are split in 18 'poi' and 128 'non-poi'

There are considerable amounts of missing values can be observable in all most every feature (for example: the features salary, bonus and to_messages have 51, 64 and 50 missing values, respectively).

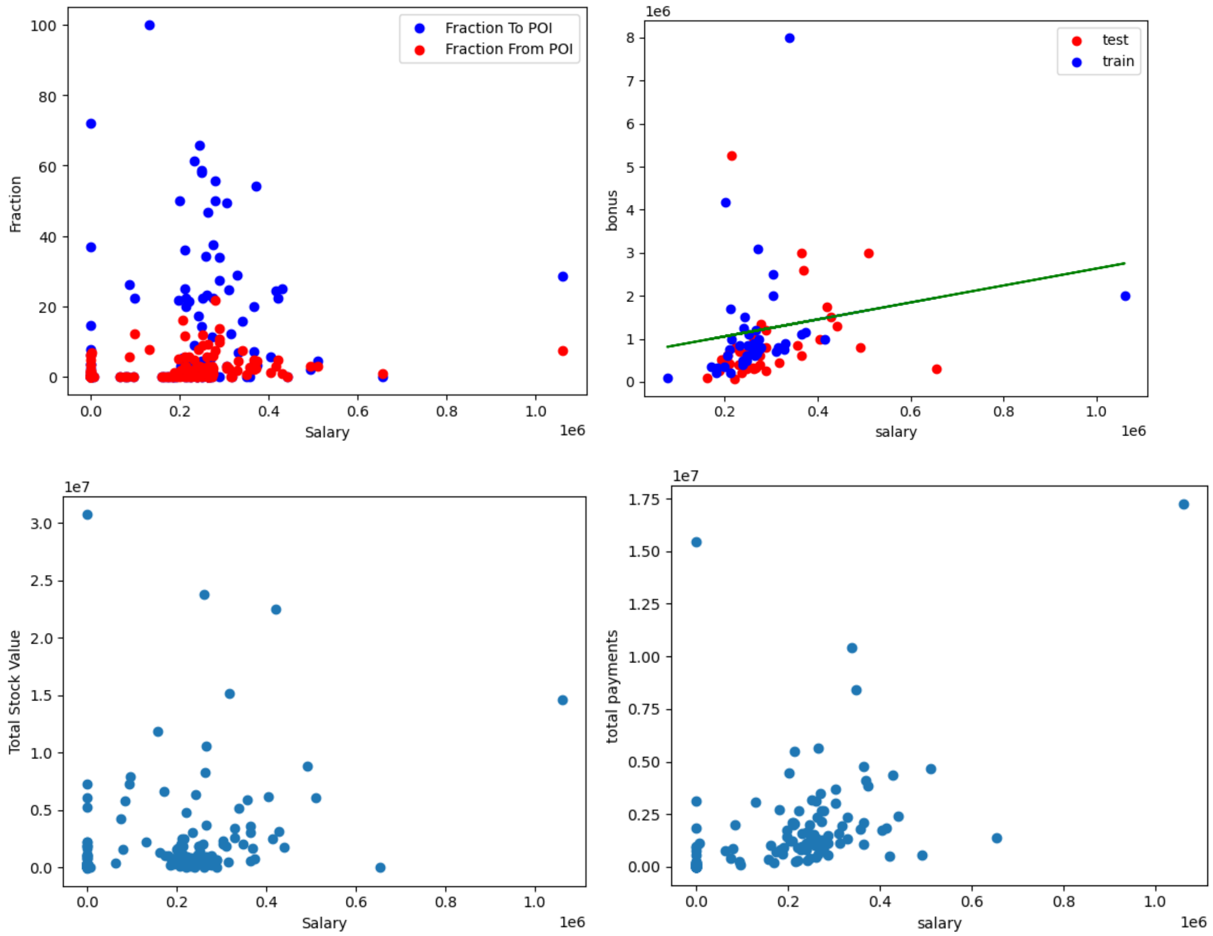
This project will focus on a subset of the dataset, which will be selected based on relevance to the research goals and the availability of labels or annotations related to fraudulent activities. We will also consider ethical and privacy considerations when handling and analyzing the data.

DATA VISUALIZATION

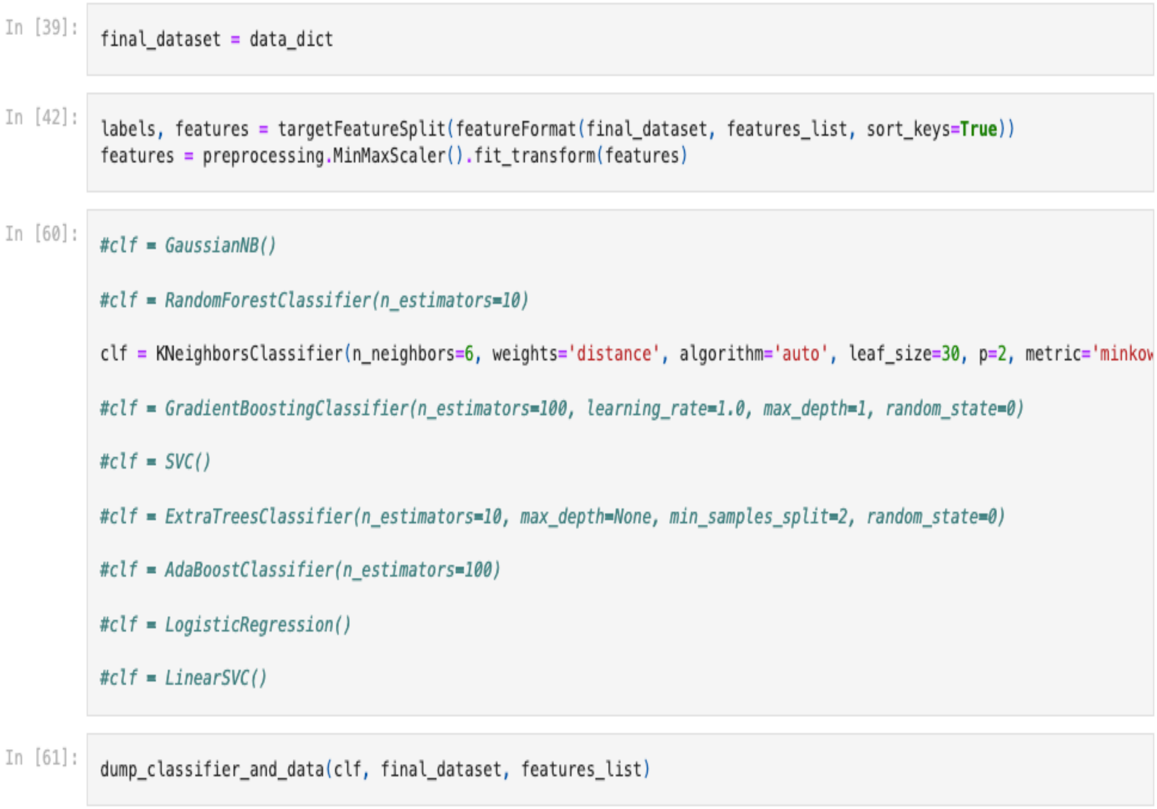




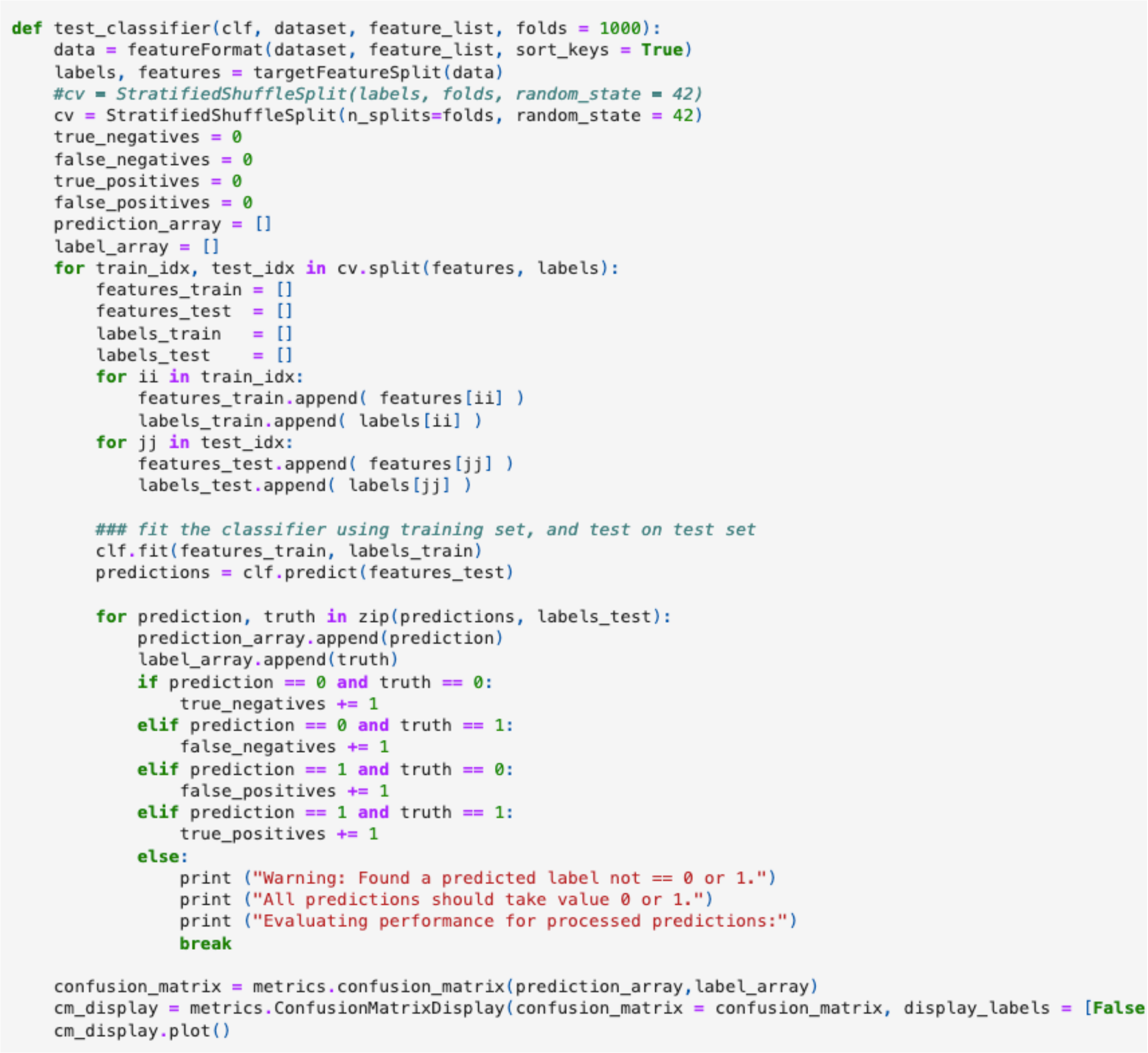
FEATURE COMPARISONS



CODE:

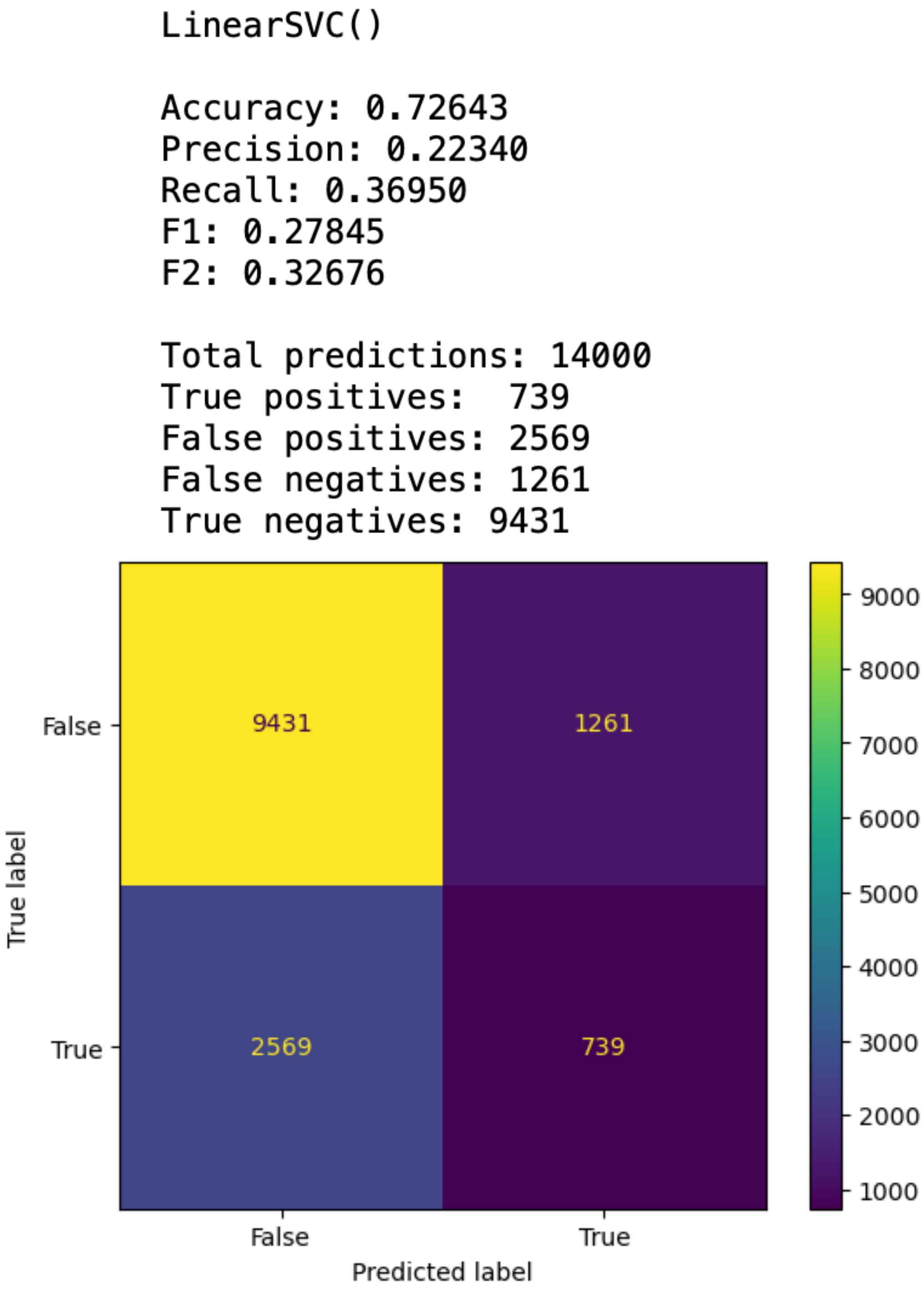


```
In [39]: final_dataset = data_dict

In [42]: labels, features = targetFeatureSplit(featureFormat(final_dataset, features_list, sort_keys=True))
         features = preprocessing.MinMaxScaler().fit_transform(features)

In [60]: #clf = GaussianNB()

         #clf = RandomForestClassifier(n_estimators=10)

         clf = KNeighborsClassifier(n_neighbors=6, weights='distance', algorithm='auto', leaf_size=30, p=2, metric='minkow

         #clf = GradientBoostingClassifier(n_estimators=100, learning_rate=1.0, max_depth=1, random_state=0)

         #clf = SVC()

         #clf = ExtraTreesClassifier(n_estimators=10, max_depth=None, min_samples_split=2, random_state=0)

         #clf = AdaBoostClassifier(n_estimators=100)

         #clf = LogisticRegression()

         #clf = LinearSVC()

In [61]: dump_classifier_and_data(clf, final_dataset, features_list)
```

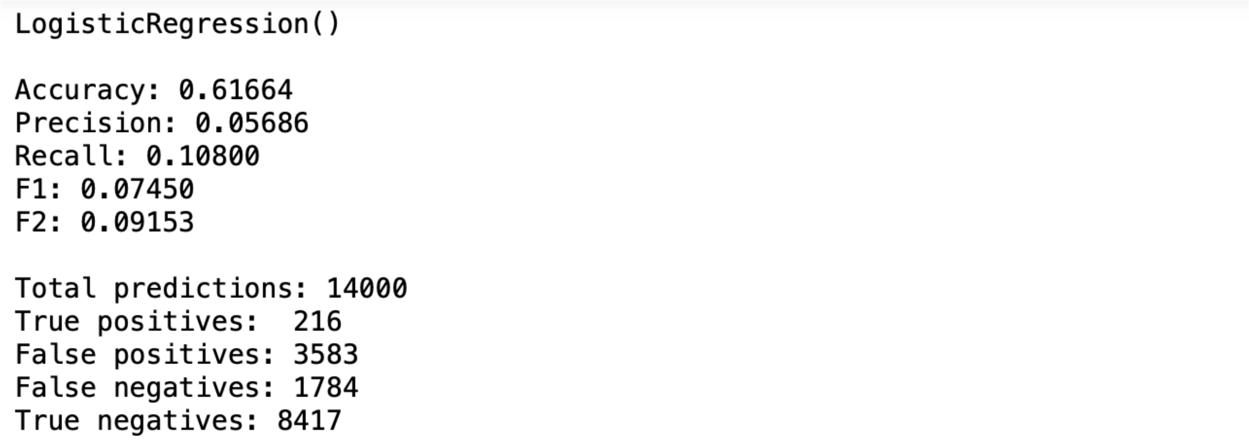


```
def test_classifier(clf, dataset, feature_list, folds = 1000):
    data = featureFormat(dataset, feature_list, sort_keys = True)
    labels, features = targetFeatureSplit(data)
    #cv = StratifiedShuffleSplit(labels, folds, random_state = 42)
    cv = StratifiedShuffleSplit(n_splits=folds, random_state = 42)
    true_negatives = 0
    false_negatives = 0
    true_positives = 0
    false_positives = 0
    prediction_array = []
    label_array = []
    for train_idx, test_idx in cv.split(features, labels):
        features_train = []
        features_test  = []
        labels_train   = []
        labels_test    = []
        for ii in train_idx:
            features_train.append( features[ii] )
            labels_train.append( labels[ii] )
        for jj in test_idx:
            features_test.append( features[jj] )
            labels_test.append( labels[jj] )

        ### fit the classifier using training set, and test on test set
        clf.fit(features_train, labels_train)
        predictions = clf.predict(features_test)

        for prediction, truth in zip(predictions, labels_test):
            prediction_array.append(prediction)
            label_array.append(truth)
            if prediction == 0 and truth == 0:
                true_negatives += 1
            elif prediction == 0 and truth == 1:
                false_negatives += 1
            elif prediction == 1 and truth == 0:
                false_positives += 1
            elif prediction == 1 and truth == 1:
                true_positives += 1
            else:
                print ("Warning: Found a predicted label not == 0 or 1.")
                print ("All predictions should take value 0 or 1.")
                print ("Evaluating performance for processed predictions:")
                break

    confusion_matrix = metrics.confusion_matrix(prediction_array,label_array)
    cm_display = metrics.ConfusionMatrixDisplay(confusion_matrix = confusion_matrix, display_labels = [False
    cm_display.plot()
```

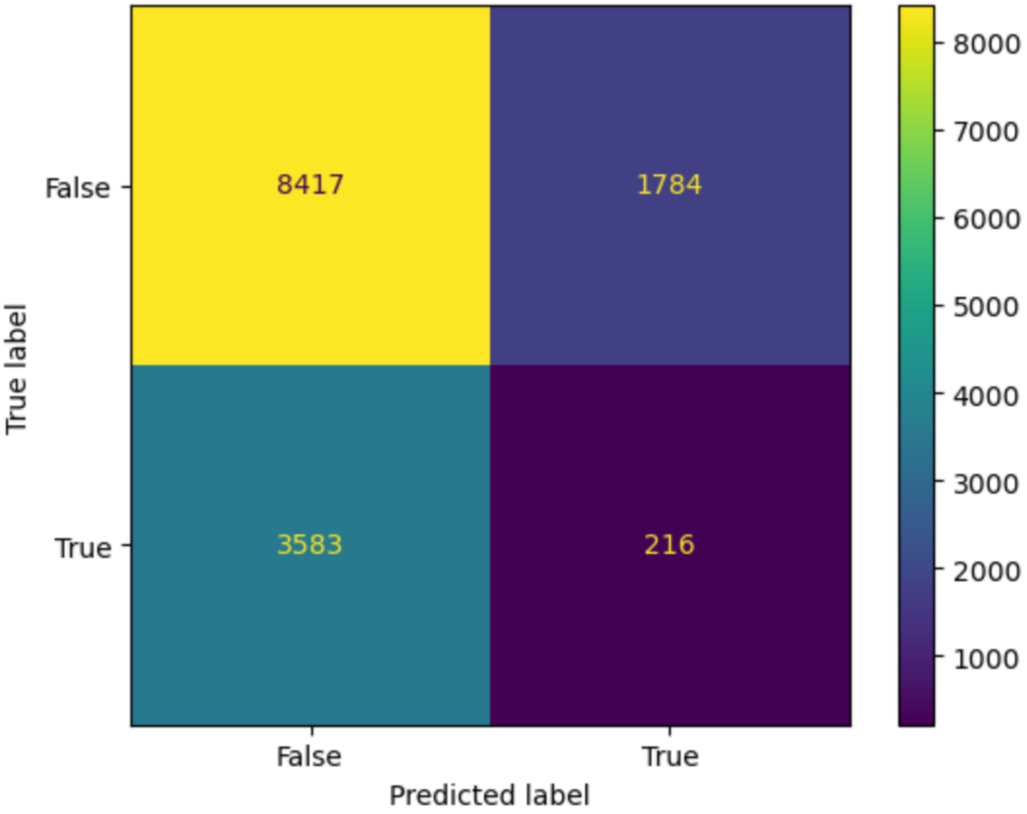
OUTPUT:

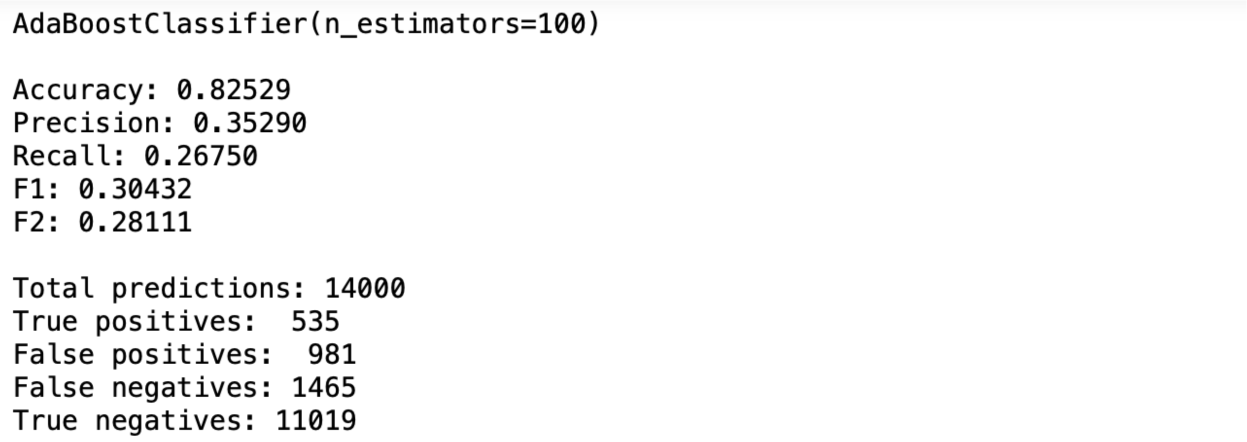


```
LinearSVC()

Accuracy: 0.72643
Precision: 0.22340
Recall: 0.36950
F1: 0.27845
F2: 0.32676

Total predictions: 14000
True positives:  739
False positives: 2569
False negatives: 1261
True negatives: 9431
```

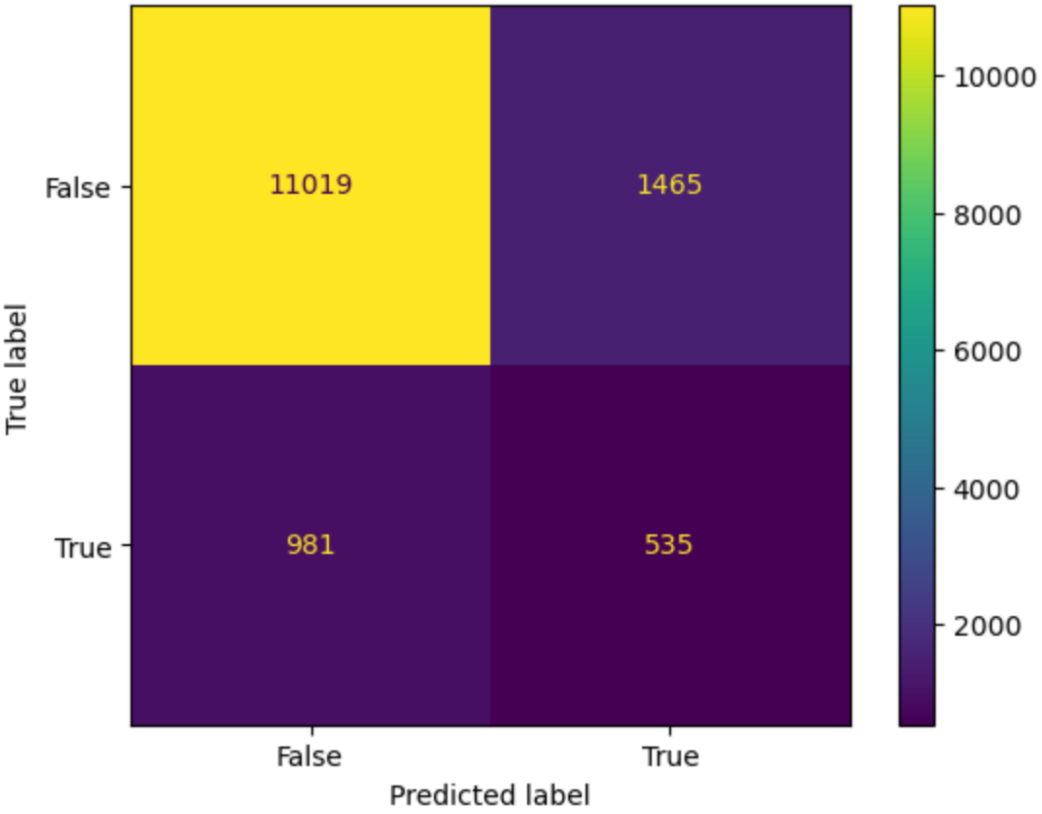


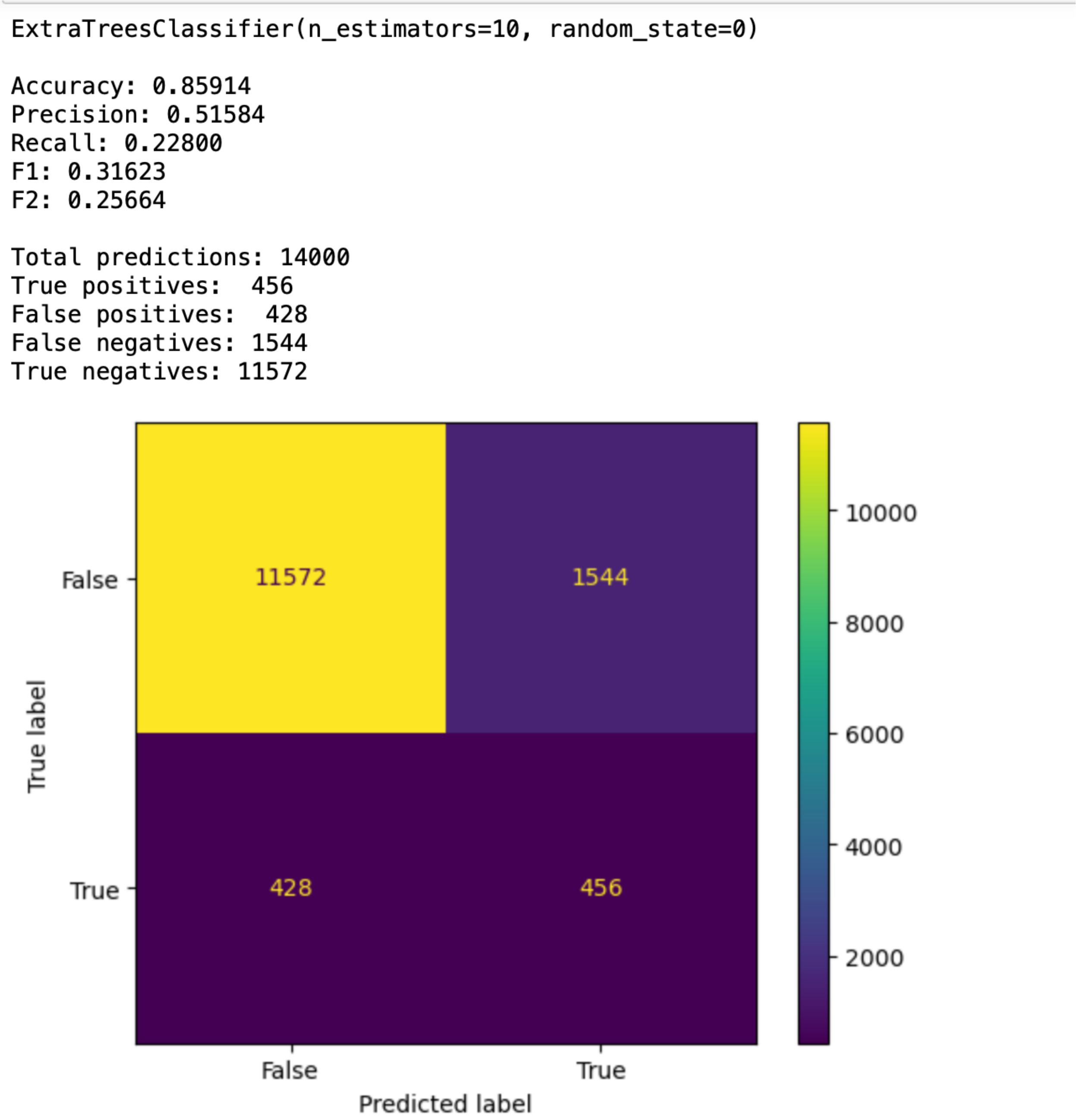
```
LogisticRegression()

Accuracy: 0.61664
Precision: 0.05686
Recall: 0.10800
F1: 0.07450
F2: 0.09153

Total predictions: 14000
True positives:  216
False positives: 3583
False negatives: 1784
True negatives: 8417
```



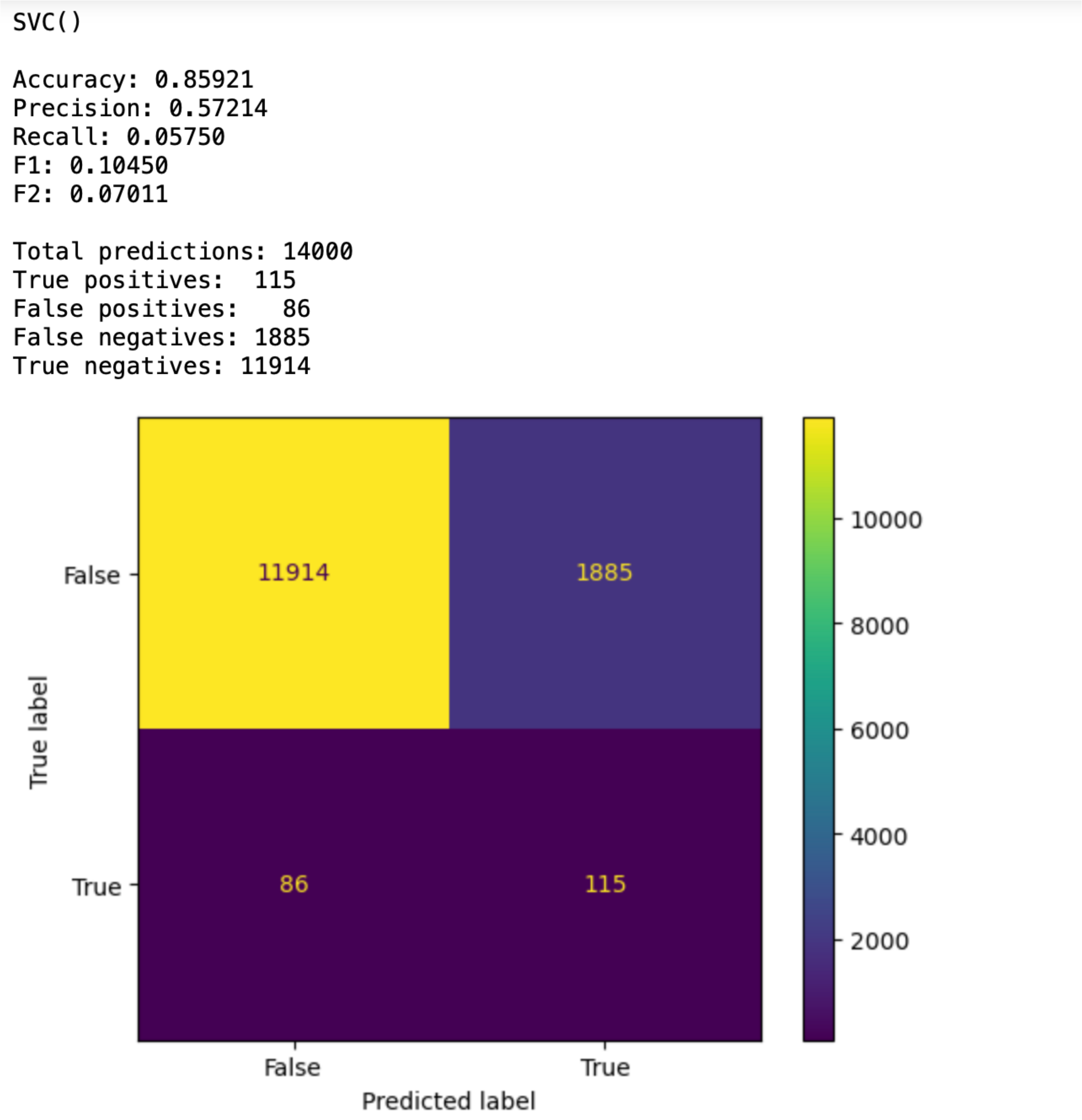


```
AdaBoostClassifier(n_estimators=100)

Accuracy: 0.82529
Precision: 0.35290
Recall: 0.26750
F1: 0.30432
F2: 0.28111

Total predictions: 14000
True positives:  535
False positives:  981
False negatives: 1465
True negatives: 11019
```

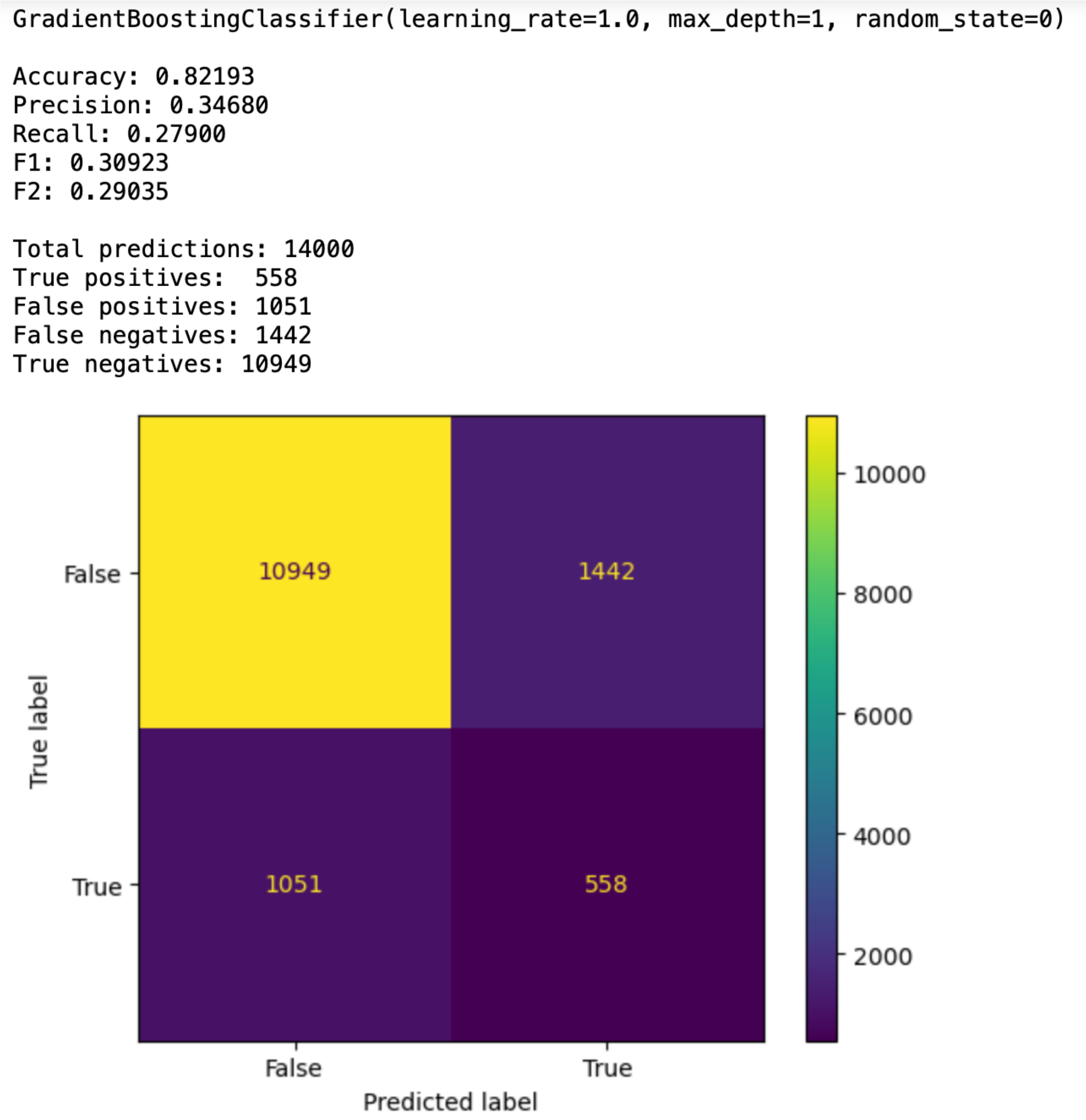




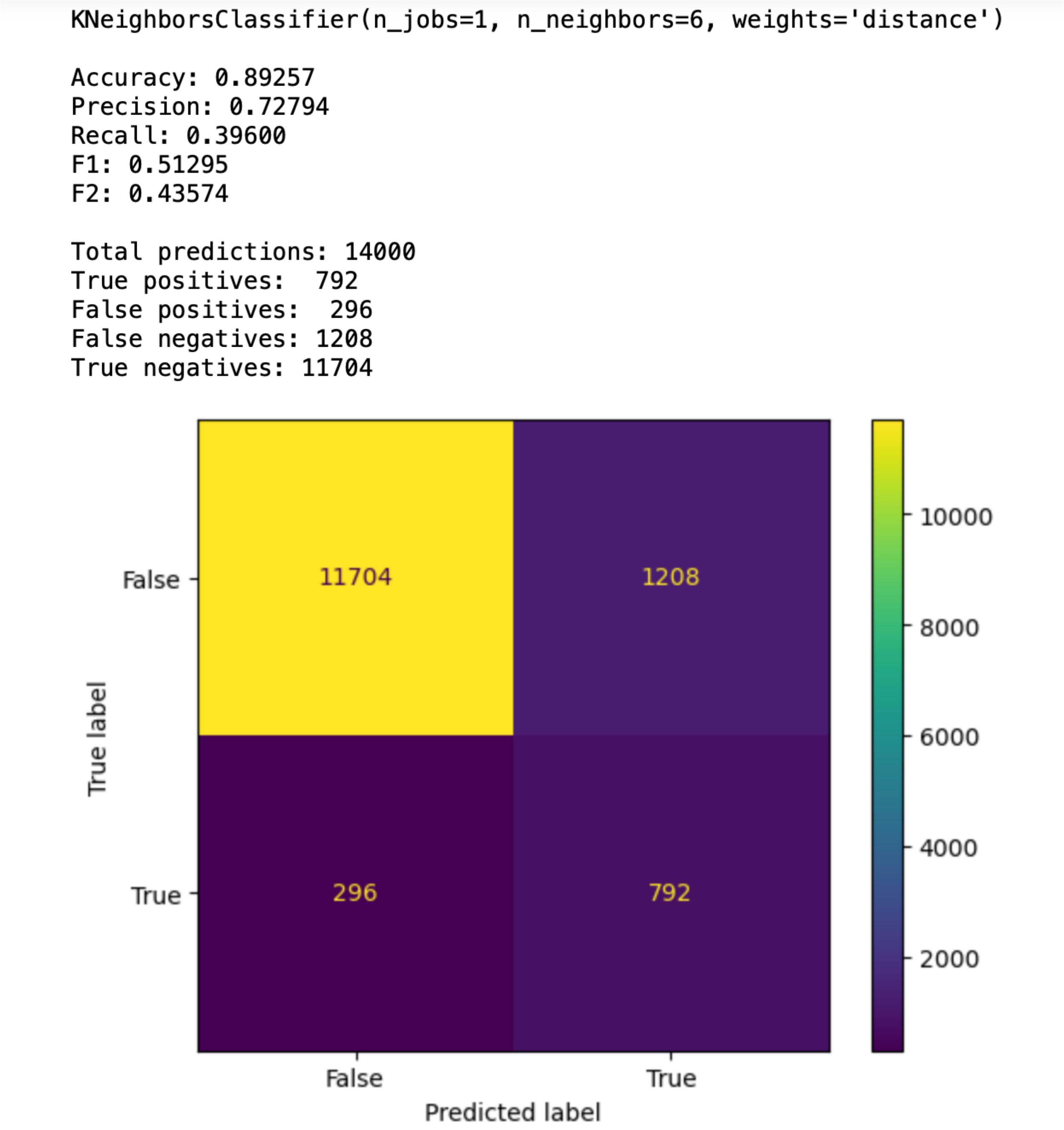
```
ExtraTreesClassifier(n_estimators=10, random_state=0)

Accuracy: 0.85914
Precision: 0.51584
Recall: 0.22800
F1: 0.31623
F2: 0.25664

Total predictions: 14000
True positives:  456
False positives:  428
False negatives: 1544
True negatives: 11572
```



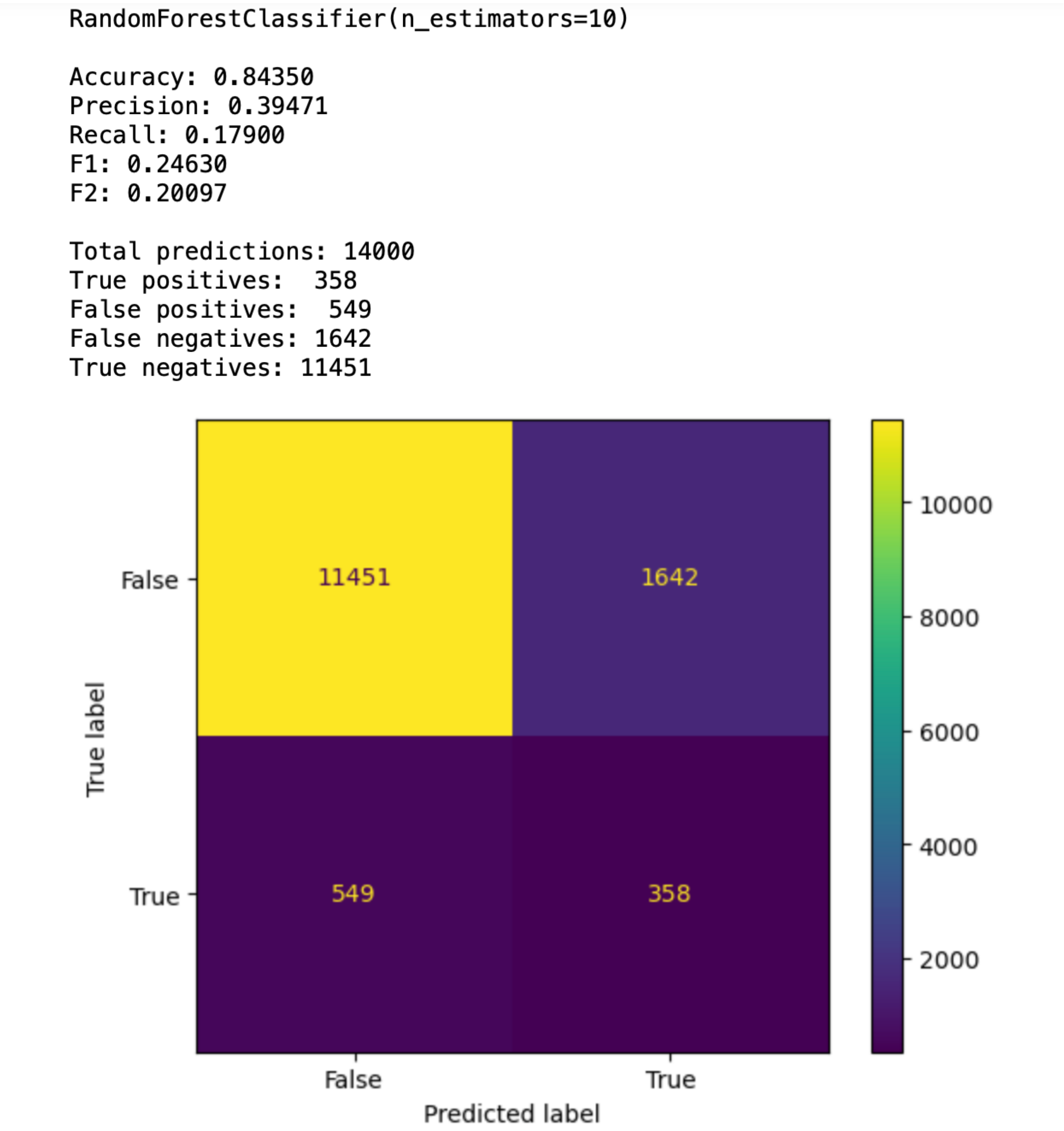
```
SVC()

Accuracy: 0.85921
Precision: 0.57214
Recall: 0.05750
F1: 0.10450
F2: 0.07011

Total predictions: 14000
True positives:  115
False positives:   86
False negatives: 1885
True negatives: 11914
```



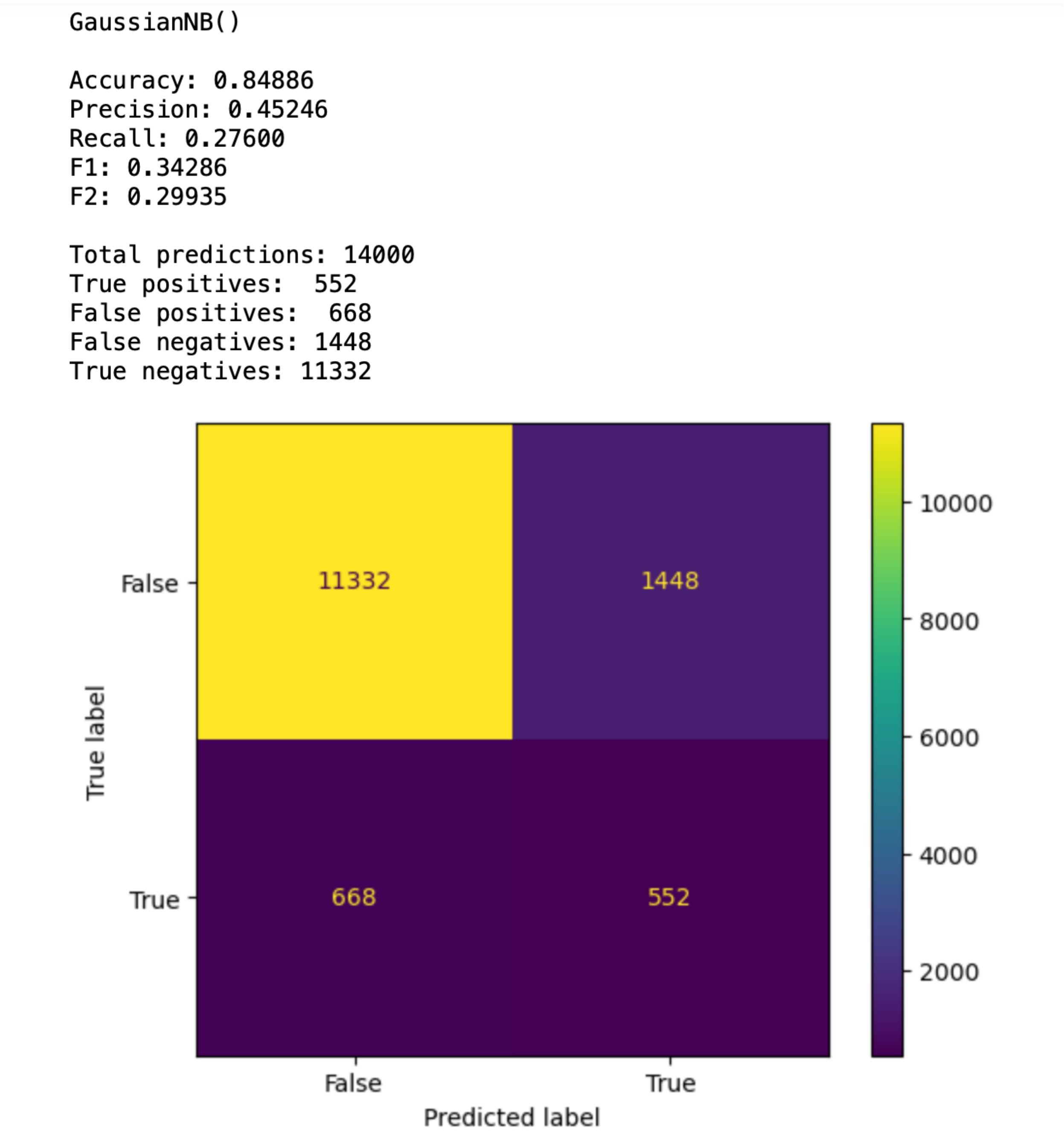
```
GradientBoostingClassifier(learning_rate=1.0, max_depth=1, random_state=0)

Accuracy: 0.82193
Precision: 0.34680
Recall: 0.27900
F1: 0.30923
F2: 0.29035

Total predictions: 14000
True positives:  558
False positives: 1051
False negatives: 1442
True negatives: 10949
```



```
KNeighborsClassifier(n_jobs=1, n_neighbors=6, weights='distance')

Accuracy: 0.89257
Precision: 0.72794
Recall: 0.39600
F1: 0.51295
F2: 0.43574

Total predictions: 14000
True positives:  792
False positives:  296
False negatives: 1208
True negatives: 11704
```



```
RandomForestClassifier(n_estimators=10)

Accuracy: 0.84350
Precision: 0.39471
Recall: 0.17900
F1: 0.24630
F2: 0.20097

Total predictions: 14000
True positives:  358
False positives:  549
False negatives: 1642
True negatives: 11451
```



```
GaussianNB()

Accuracy: 0.84886
Precision: 0.45246
Recall: 0.27600
F1: 0.34286
F2: 0.29935

Total predictions: 14000
True positives:  552
False positives:  668
False negatives: 1448
True negatives: 11332
```

CONCLUSION

this project proposal represents a significant opportunity to harness the wealth of information contained within the Enron Email Dataset and apply cutting-edge machine learning techniques to address an enduring issue of corporate misconduct and fraud. By delving into the depths of this dataset, we seek to achieve several key objectives:

DETECTION AND PREVENTION

The primary goal of this project is to develop predictive models capable of detecting early signs and patterns of fraudulent activities within the corporate communications of Enron. This has the potential to be a valuable tool for companies and regulators alike, helping to prevent future corporate scandals of similar magnitude.

CORPORATE UNDERSTANDING

We aim to gain a deeper understanding of the dynamics within corporate communications, including the subtleties of sentiment, intent, and linguistic patterns. This understanding will not only inform our predictive models but also contribute to the broader field of natural language processing and social network analysis.

ACTIONABLE INSIGHTS

The project intends to deliver actionable insights to corporate governance and compliance teams. By pinpointing communication patterns indicative of wrongdoing, our research could guide organizations in improving their internal monitoring and compliance procedures, ultimately reducing the risk of misconduct.

Refrences :

https://www.cs.cmu.edu/~./enron/
https://ceas.cc/
https://link.springer.com/chapter/10.1007/978-3-540-30115-8_22