Project Report On:
Stress Analysis Using Social Media

[1]Aditi Srivastava, [1]Pavan Kumar Reddy Devulapalle
CS 6375, Computer Science, Prof. Sriraam Natarajan
The University of Texas at Dallas

Introduction

Problem Statement

The aim of the project is to analyze stress in people on social media. The dataset used for analysis has been taken from Kaggle. Internet, specially social media, is a major cause of stress like depression, anxiety, post traumatic stress disorder and other mental health problems. Almost everyone uses the social media in one way or the other. Therefore, data from social media can be very informative for doctors and psychologists in providing traits and patterns a stressed person may show via their posts, likes, comments etc.

Dataset Description

The dataset has a total of 2838 training data and 116 features which is divided into training and testing data for training the model. The classes are:

- 0 (indicating no stress symptoms)
- 1 (indicating stress symptoms)

Methodology

Pre-processing

The data required some sort of pre-processing before training the models. The dataset has many features which are not required by the analysis. So, relevant features were selected and the data was checked for any required resampling. The ratio between the two classes is approximately 11:13 hence the classes are balanced.

```
Percentage of no stress is 47.51322751322751
Percentage of stress is 52.48677248677248
```

Figure 1: Class balance percentage

Models Used

- Logistic Regression
- Gaussian Naive Bayes Classifier

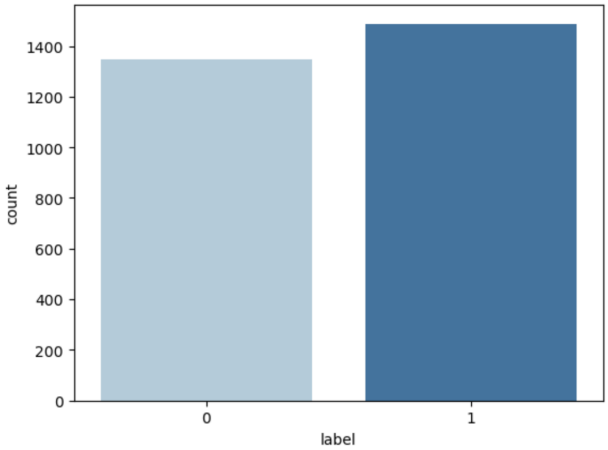

Figure 2: Class Counts Bar Chart

- Decision Tree Classifier
- Support Vector Machines
- Ensemble Methods (Bagging and Boosting)

Results and Discussion

For better understanding of the behavior of these models and the data, the dataset is divided into different feature types and the five classifiers are applied separately to each feature type and then applied on the combined dataset. The different feature types are:

- Linguistic styles and common grammar
- Psychological processes
- Social processes
- Cognitive processes
- Perceptual and Biological Processes
- Time Orientations
- All feature types combined

Gaussian Naive Bayes has classified data with an AUC score of approx. 74 percent because the data was normalised before applying this model. But this model cannot be used

for our predictions because it assumes all features to be independent but the features are practically correlated to each other on different levels hence its results are not accurate.

Decision Trees had a surprisingly lower AUC value in the range 0.5-0.6 which means that the predictions were almost equal to predictions done randomly. This is because decision trees lose valuable information when dealing with continuous values hence the AUC value and is prone to overfitting as the model becomes complex when continuous data is used.

Ensemble methods had varied performances, based on the model used.

Logistic Regression and Support Vector Machines have outdone all other models in every category. They had very minute differences in their performances because the data used is linearly separable and they tend to separate data along linear hyper surfaces, thus reducing complexity.

But, on an average, the classifier that performed the best was Support Vector Machines and the worst performance was of Decision Tree Classifiers.

Linguistic Styles and Common Grammar

1. Logistic Regression

Best performance among all

	precision	recall	f1-score	support
0	0.63	0.62	0.62	386
1	0.69	0.69	0.69	464
accuracy			0.66	850
macro avg	0.66	0.66	0.66	850
weighted avg	0.66	0.66	0.66	850

Figure 3: Classification Report

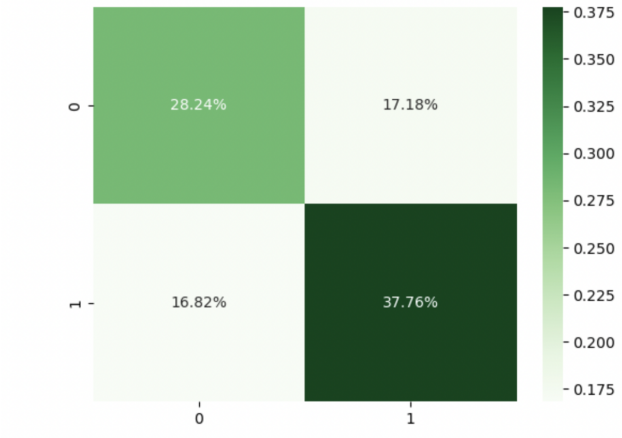

Figure 4: Confusion Matrix

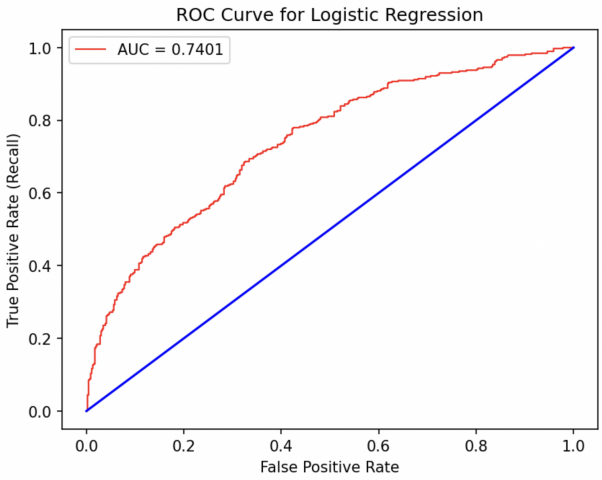

Figure 5: ROC Curve

2. Decision Tree Classifier

Worst performance among all

	precision	recall	f1-score	support
0	0.56	0.57	0.56	386
1	0.63	0.62	0.63	464
accuracy			0.60	850
macro avg	0.60	0.60	0.60	850
weighted avg	0.60	0.60	0.60	850

Figure 6: Classification Report

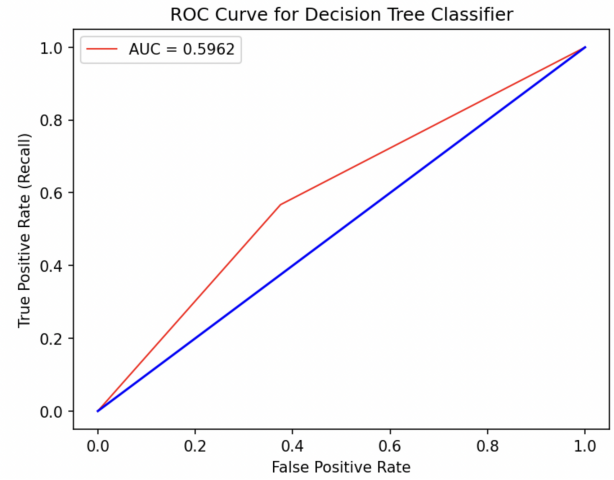

Figure 7: ROC Curve

Psychological Processes

1. Support Vector Machines

Best performance among all

Accuracy of support vector machine on test set: 0.7074

	precision	recall	f1-score	support
0	0.67	0.74	0.70	396
1	0.75	0.68	0.71	455
accuracy			0.71	851
macro avg	0.71	0.71	0.71	851
weighted avg	0.71	0.71	0.71	851

Figure 8: Classification Report

2. Decision Tree Classifier

Worst performance among all

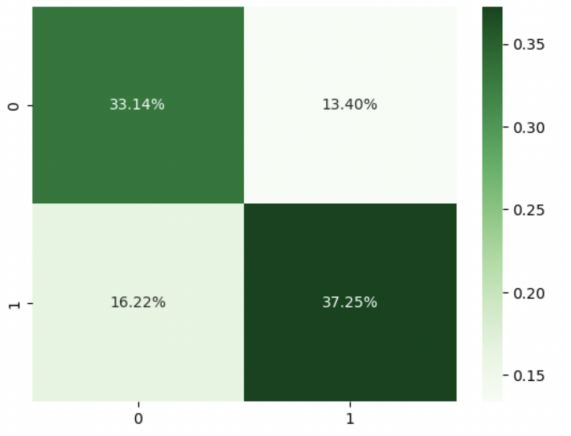

Figure 9: Confusion Matrix

Accuracy of decision tree classifier on test set: 0.6533

	precision	recall	f1-score	support
0	0.62	0.64	0.63	396
1	0.68	0.66	0.67	455
accuracy			0.65	851
macro avg	0.65	0.65	0.65	851
weighted avg	0.65	0.65	0.65	851

Figure 10: Classification Report

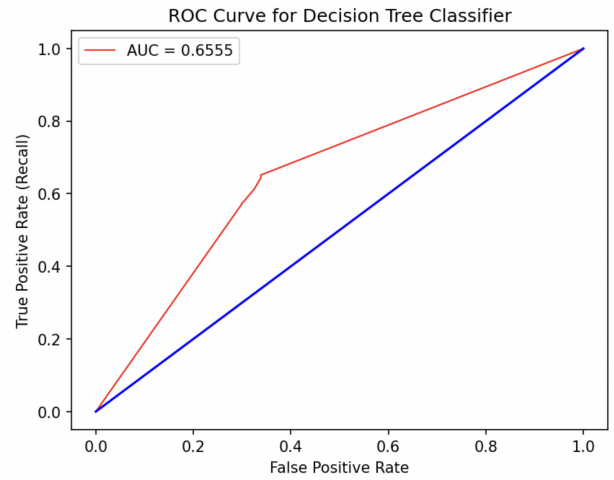

Figure 11: ROC Curve

Social Processes

1. Support Vector Machines

Best performance among all

Accuracy of support vector machine on test set: 0.5558

	precision	recall	f1-score	support
0	0.54	0.18	0.27	389
1	0.56	0.87	0.68	462
accuracy			0.56	851
macro avg	0.55	0.53	0.47	851
weighted avg	0.55	0.56	0.49	851

Figure 12: Classification Report

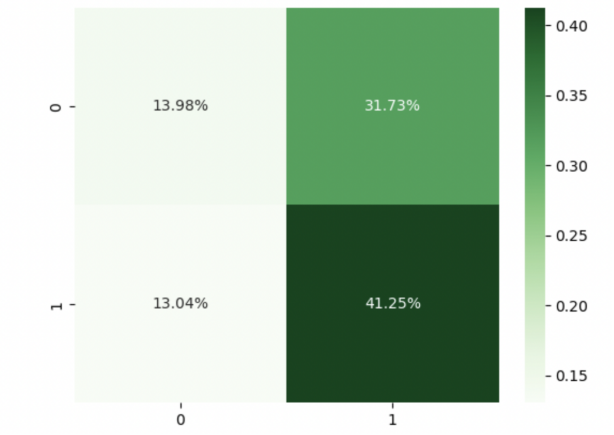

Figure 13: Confusion Matrix

2. Decision Tree Classifier

Worst performance among all

	precision	recall	f1-score	support
0	0.50	0.37	0.42	389
1	0.56	0.69	0.62	462
accuracy			0.54	851
macro avg	0.53	0.53	0.52	851
weighted avg	0.53	0.54	0.53	851

Figure 14: Classification Report

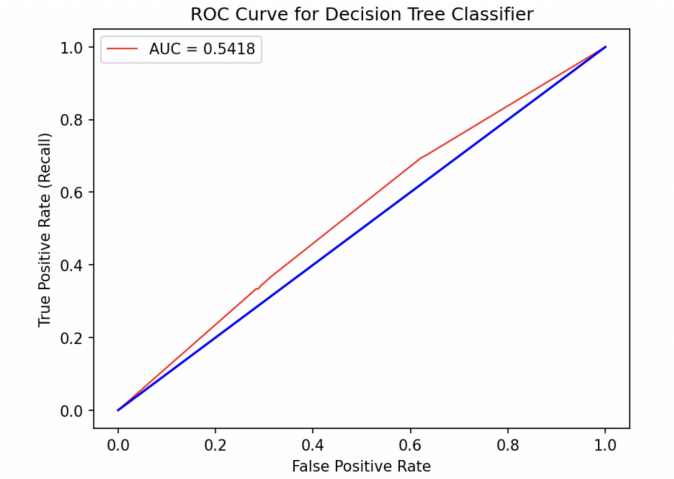

Figure 15: ROC Curve

Cognitive Processes

Best performance among all

1. Logistic Regression

	precision	recall	f1-score	support
0	0.53	0.50	0.51	393
1	0.59	0.62	0.60	458
accuracy			0.56	851
macro avg	0.56	0.56	0.56	851
weighted avg	0.56	0.56	0.56	851

Figure 16: Classification Report

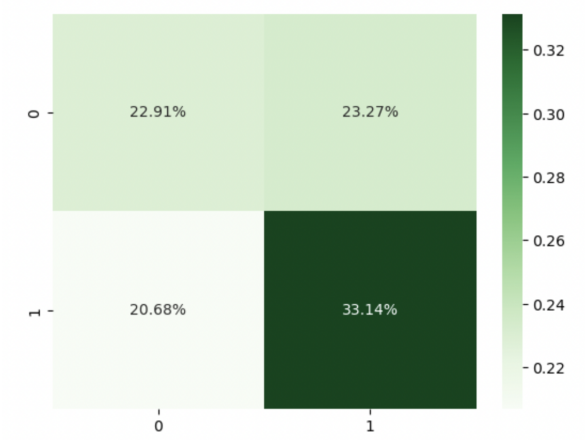

Figure 17: Confusion Matrix

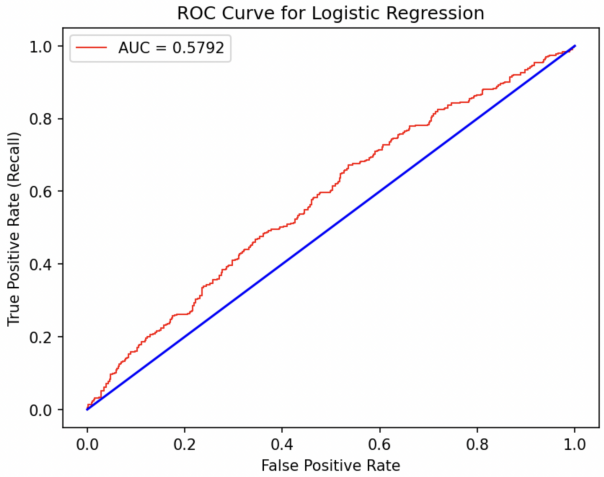

Figure 18: ROC Curve

2. Ensemble methods using Boosting Classifier

Worst performance among all

Accuracy of ensemble method using Boosting Classifier on test set: 0

	precision	recall	f1-score	support
0	0.49	0.47	0.48	393
1	0.56	0.57	0.56	458
accuracy			0.53	851
macro avg	0.52	0.52	0.52	851
weighted avg	0.52	0.53	0.52	851

Figure 19: Classification Report

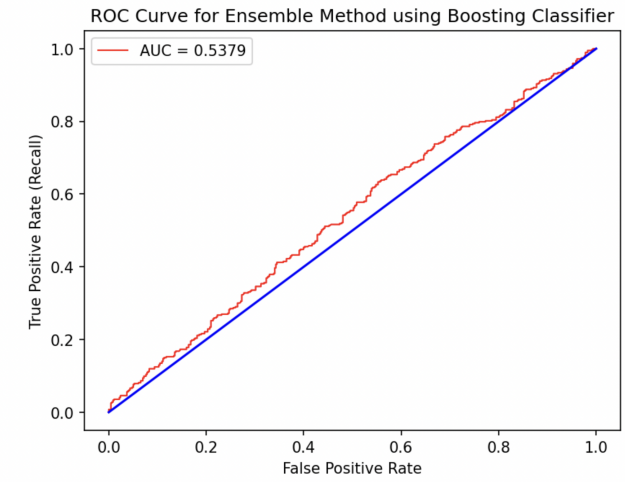

Figure 20: ROC Curve

Perceptual and Biological Processes

Best performance among all

1. Support Vector Machines

Accuracy of support vector machine on test set: 0.6028

	precision	recall	f1-score	support
0	0.56	0.72	0.63	398
1	0.67	0.50	0.57	453
accuracy			0.60	851
macro avg	0.61	0.61	0.60	851
weighted avg	0.62	0.60	0.60	851

Figure 21: Classification Report

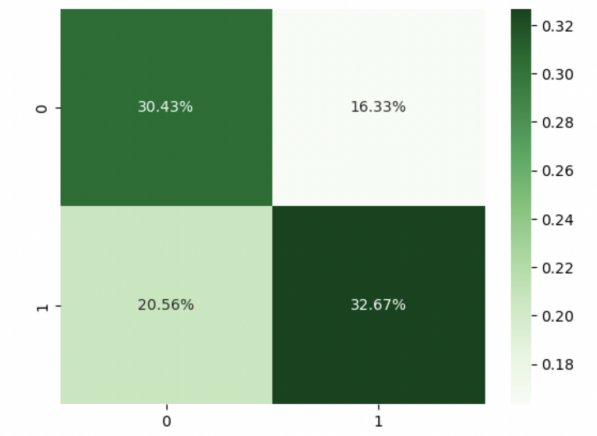

Figure 22: Confusion Matrix

2. Decision Tree Classifier

Worst performance among all

Accuracy of decision tree classifier on test set: 0.5417

	precision	recall	f1-score	support
0	0.51	0.60	0.55	398
1	0.58	0.49	0.53	453
accuracy			0.54	851
macro avg	0.55	0.55	0.54	851
weighted avg	0.55	0.54	0.54	851

Figure 23: Classification Report

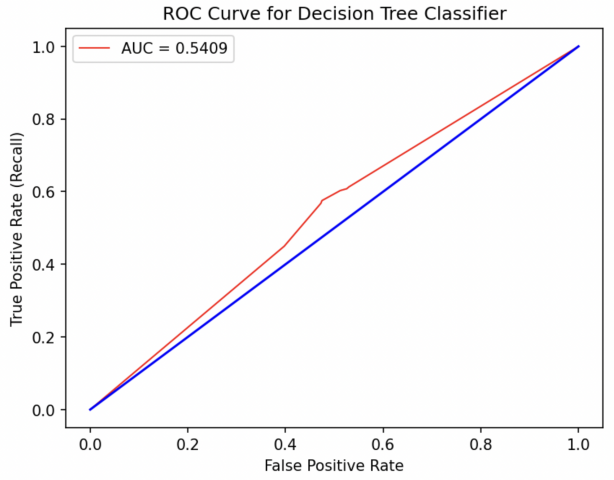

Figure 24: ROC Curve

Time Orientations

1. Ensemble Methods (Gradient Boosting)

Best performance among all

Accuracy of ensemble method using Boosting Classifier on test set: 0

	precision	recall	f1-score	support
0	0.63	0.53	0.57	427
1	0.59	0.68	0.63	424
accuracy			0.61	851
macro avg	0.61	0.61	0.60	851
weighted avg	0.61	0.61	0.60	851

Figure 25: Classification Report

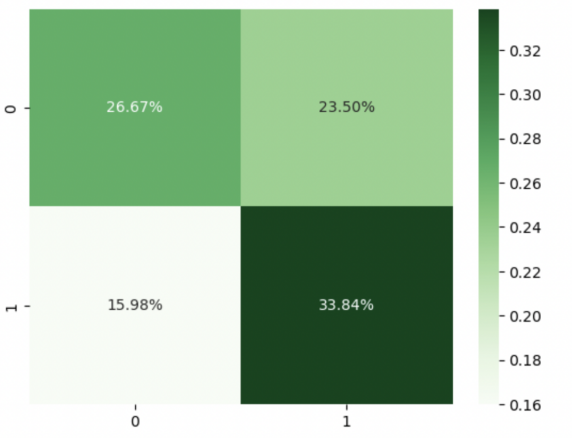

Figure 26: Confusion Matrix

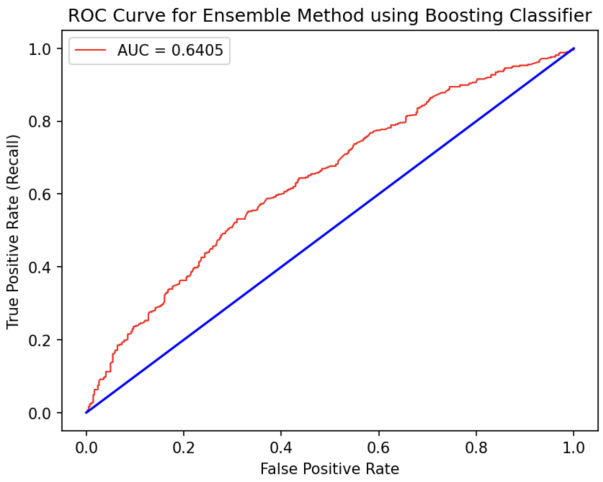

Figure 27: ROC Curve

2. Decision Tree Classifier

Worst performance among all

Accuracy of decision tree classifier on test set: 0.5499

	precision	recall	f1-score	support
0	0.55	0.56	0.55	427
1	0.55	0.54	0.55	424
accuracy			0.55	851
macro avg	0.55	0.55	0.55	851
weighted avg	0.55	0.55	0.55	851

Figure 28: Classification Report

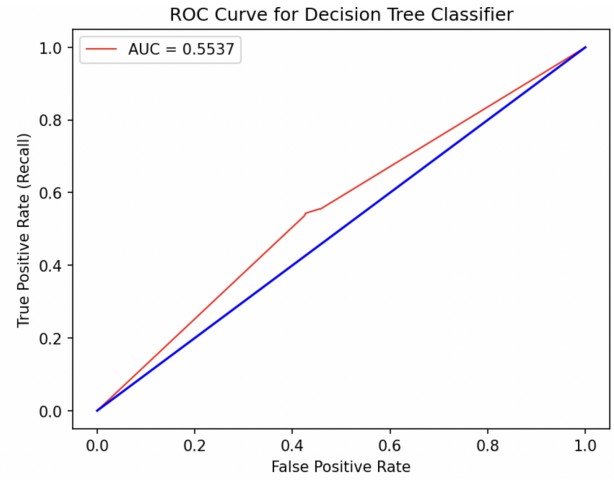

Figure 29: ROC Curve

All Feature Types Combined

1. Logistic Regression

Best performance among all

	precision	recall	f1-score	support
0	0.71	0.74	0.72	387
1	0.77	0.75	0.76	464
accuracy			0.74	851
macro avg	0.74	0.74	0.74	851
weighted avg	0.74	0.74	0.74	851

Figure 30: Classification Report

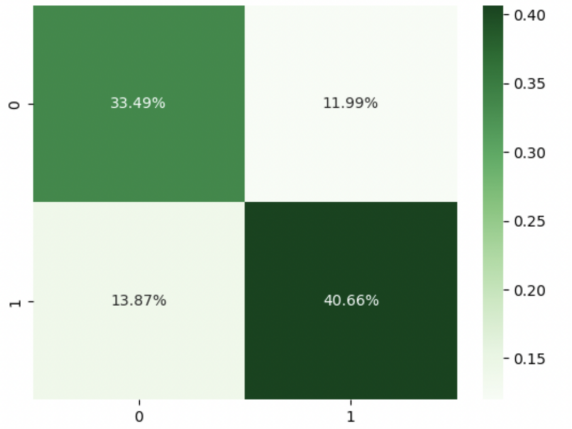

Figure 31: Confusion Matrix

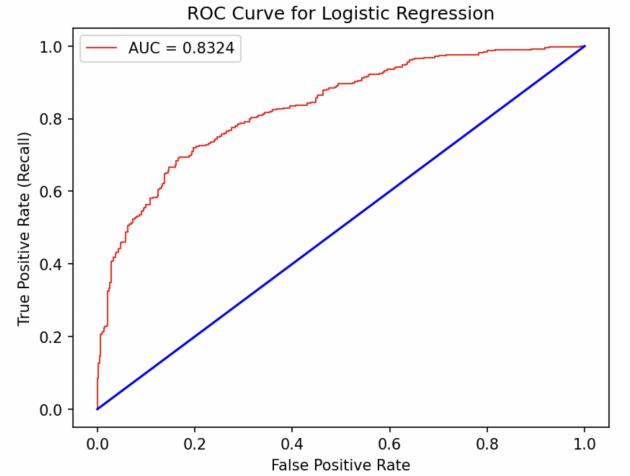

Figure 32: ROC Curve

3. Decision Tree Classifier

Worst performance among all

Accuracy of decision tree classifier on test set: 0.6710

	precision	recall	f1-score	support
0	0.66	0.67	0.66	411
1	0.69	0.67	0.68	440
accuracy			0.67	851
macro avg	0.67	0.67	0.67	851
weighted avg	0.67	0.67	0.67	851

Figure 33: Classification Report

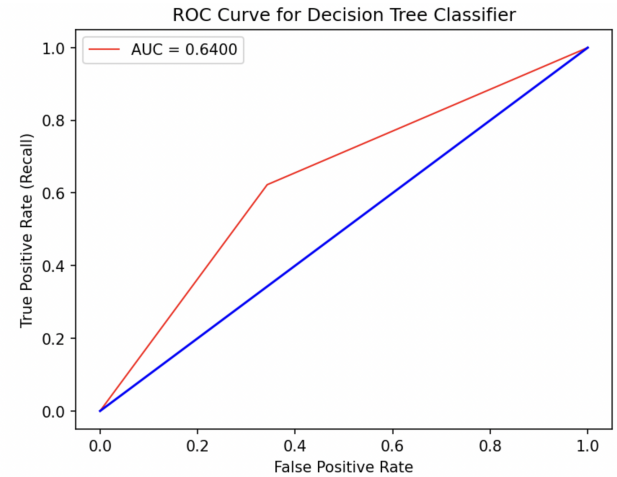

Figure 34: ROC Curve