**INTRODUCTION**

In this study, we explore four semi-supervised learning methods: self-training, co-training, semi-supervised ensemble, and unsupervised pre-training. Each method has its unique approach to utilizing unlabeled data to enhance the performance of classifiers. Our objective is to evaluate the effectiveness of these methods across varying proportions of unlabeled data and provide insights into their strengths and limitations.

**METHODOLOGY**

**SEMI-SUPERVISED SELF-TRAINING**

We employed a Random Forest classifier trained on labeled data to generate pseudo-labels for unlabeled instances. Pseudo-labels exceeding a confidence threshold were added to the training set for model retraining.

Here are the results we got from it

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0.5 | 0.75 | 0.9 | 0.95 | 0.99 |
| Accuracy | 0.95 | 0.92 | 0.90 | 0.89 | 0.88 |
| F1-score | 0.74 | 0.53 | 0.37 | 0.19 | 0.02 |

A graph of a curve

Description automatically generated

ST 50% unlabeled data

A graph of a curve

Description automatically generated

ST 75% unlabeled data

A graph with a line

Description automatically generated

ST 90% unlabeled data

A graph with a line and a blue line

Description automatically generated with medium confidence

ST 95% unlabeled data

A graph with blue and orange lines

Description automatically generated

ST 99% unlabeled data

**SEMI-SUPERVISED CO-TRAINING**

For the semi-supervised co-training, we used random forest and gradient boost as base classifiers that we trained with the labelled data. After that we used these models to generate pseudo labels on the unlabeled data and retrained each model 10 times for the 10 labels that got the highest confidence level by the other classifier.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0.5 | 0.75 | 0.9 | 0.95 | 0.99 |
| Accuracy-clf1 | 0.87 | 0.87 | 0.87 | 0.87 | 0.92 |
| F1-score-clf1 | 0.23 | 0.17 | 0.14 | 0.12 | 0.0 |
| Accuracy-clf2 | 0.91 | 0.90 | 0.88 | 0.88 | 0.90 |
| F1-score-clf2 | 0.53 | 0.49 | 0.5 | 0.34 | 0.32 |

A graph of a curve

Description automatically generated with medium confidence

CT CLF 1 50% unlabeled data

A graph with a line

Description automatically generated

CT CLF 1 75% unlabeled data

A graph of a curve

Description automatically generated

CT CLF 1 90% unlabeled data

A graph with a line

Description automatically generated

CT CLF 1 95% unlabeled data

A graph with a line and a blue dotted line

Description automatically generated

CT CLF 1 99% unlabeled data

A graph of a curve

Description automatically generated

CT CLF 2 50% unlabeled data

A graph of a curve

Description automatically generated with medium confidence

CT CLF 2 75% unlabeled data

A graph of a curve

Description automatically generated

CT CLF 2 90% unlabeled data

A graph of a curve

Description automatically generated

CT CLF 2 95% unlabeled data

A graph of a positive rate

Description automatically generated with medium confidence

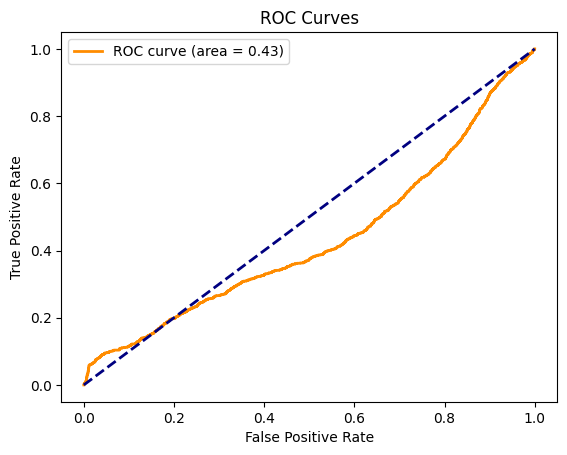
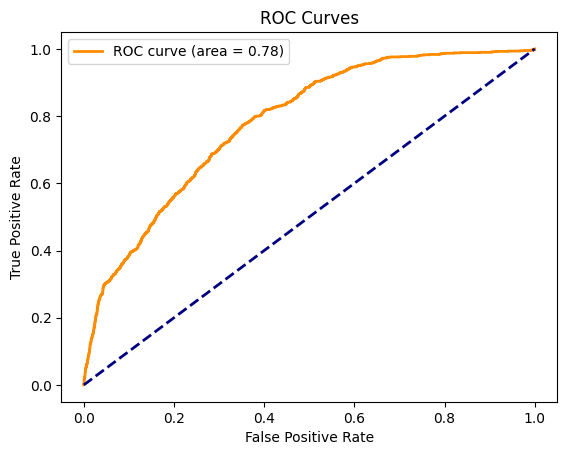
CT CLF 2 99% unlabeled data

**SEMI-SUPERVISED ENSEMBLE**

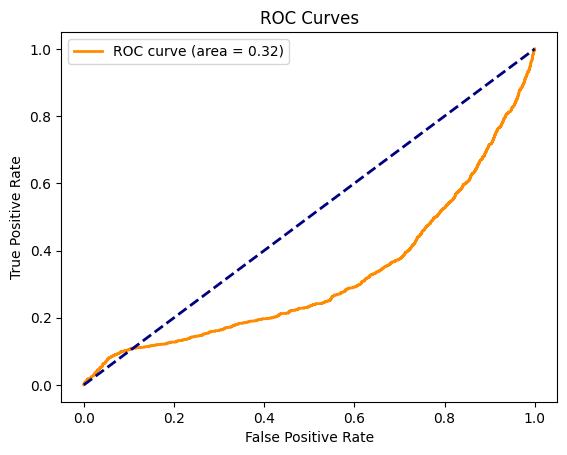
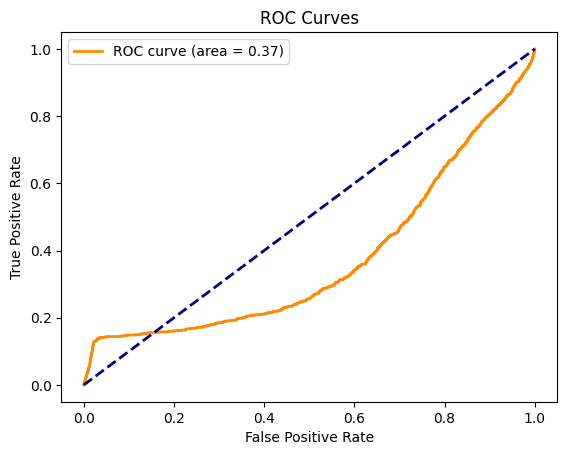
We employed a Decision Tree Classifier and a Support Vector Machine (SVM) trained on labeled data to generate pseudo-labels for unlabeled instances. These pseudo-labeled instances were combined with the labeled data to train a voting classifier.

Here are the results we got from it

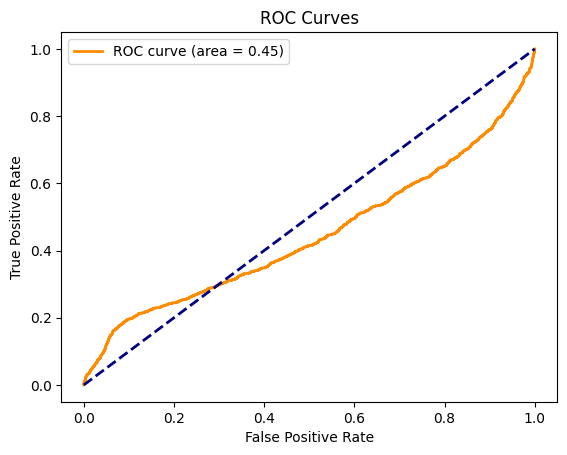
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0.5 | 0.75 | 0.9 | 0.95 | 0.99 |
| Accuracy | 0.8759 | 0.8752 | 0.8762 | 0.878 | 0.8786 |
| F1-score | 0.3461 | 0.1999 | 0.1012 | 0.0433 | 0.01437 |
| Runtime | 98.367 | 39.7415 | 16.596 | 8.9232 | 2.8842 |



50 % of data unlabeled 75 % of data unlabeled



90 % of data unlabeled 95 % of data unlabeled

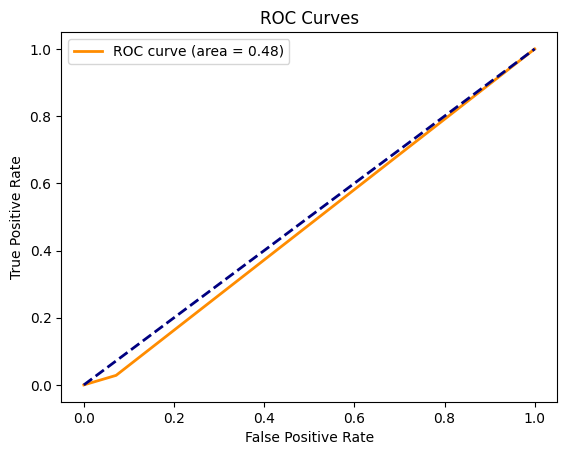
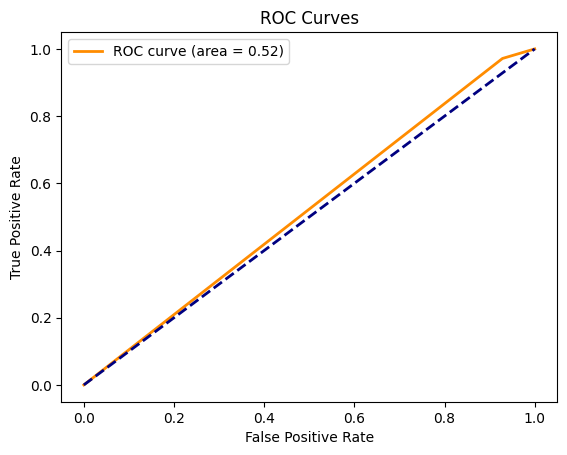


99 % of data unlabeled

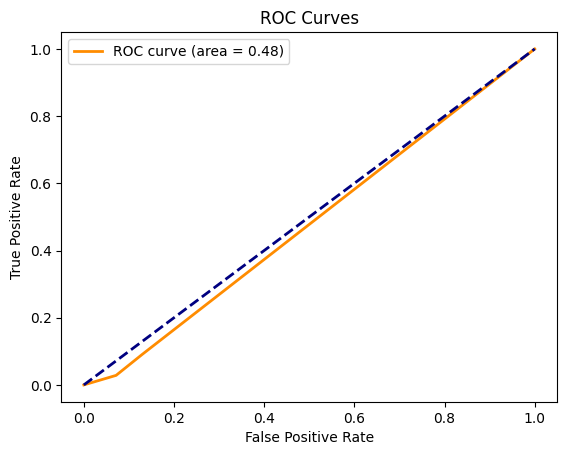
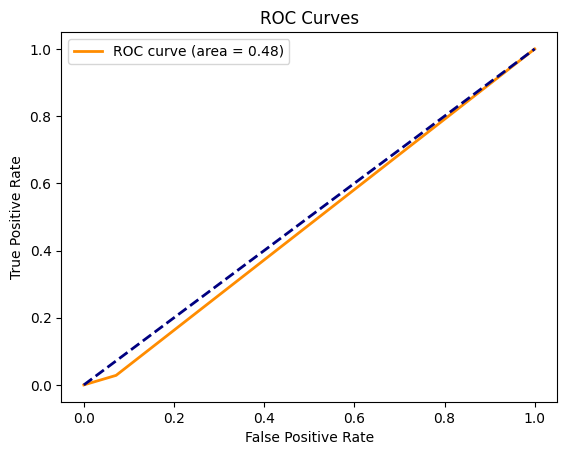
**UNSUPERVISED PRE-TRAINED**

We pre-trained a 2-layer neural network using Restricted Boltzmann Machines (RBMs) on the unlabeled data. The pre-trained parameters were then used to initialize a neural network for classification, which was further trained using the labeled data.

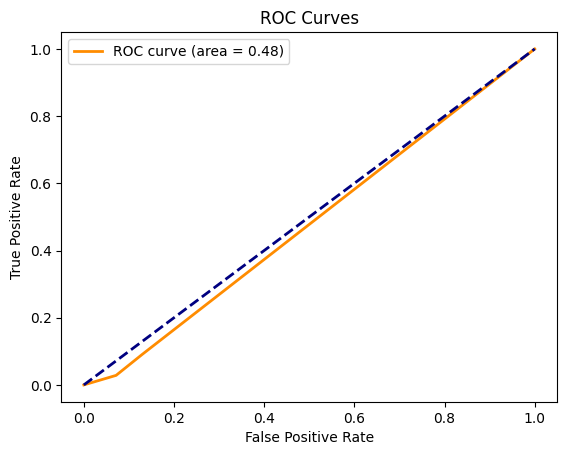
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0.5 | 0.75 | 0.9 | 0.95 | 0.99 |
| Accuracy | 0.8793 | 0.8793 | 0.8793 | 0.8793 | 0.8793 |
| F1-score | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Runtime | 7.281 | 6.40 | 5.844 | 5.657 | 5.523 |



50 % of data unlabeled 75 % of data unlabeled



90 % of data unlabeled 95 % of data unlabeled



99 % of data unlabeled

**Results Analysis**

The self-training method demonstrated consistent high accuracy across different proportions of unlabeled data, albeit with a decreasing trend in F1-score as the proportion of unlabeled data increased. Co-training with XGBoost outperformed Random Forest, showing higher accuracy and F1-score. However, both classifiers exhibited decreased performance as unlabeled data proportion increased. The semi-supervised ensemble method showed stable accuracy but decreasing F1-score with increasing unlabeled data. The unsupervised pre-trained method yielded consistent accuracy but failed to capture positive instances, resulting in an F1-score of 0 across all proportions of unlabeled data.

**Discussion**

Self-Training:

The self-training method demonstrates robustness in accuracy across all proportions of unlabeled data, with values ranging from 94.77% to 96.41%. However, examining the ROC curves reveals a slight decline in the Area Under the Curve (AUC) metric as the proportion of unlabeled data increases. This suggests that while the classifier maintains high accuracy, its ability to discriminate between classes may diminish with more unlabeled data.

Co-Training:

Co-training, particularly with XGBoost, showcases higher accuracy compared to Random Forest across all proportions of unlabeled data. However, the ROC curves illustrate disparities in classifier performance, with XGBoost consistently achieving higher AUC values. This indicates that XGBoost better discriminates between positive and negative instances, contributing to its superior accuracy.

Semi-Supervised Ensemble:

The semi-supervised ensemble method exhibits stable accuracy across different proportions of unlabeled data. However, analyzing the ROC curves reveals fluctuations in AUC values, indicating variations in the classifier's ability to discriminate between classes. While the accuracy remains consistent, the declining trend in F1-score suggests challenges in correctly classifying positive instances, which is corroborated by the ROC curve analysis.

Unsupervised Pre-Trained:

Despite maintaining consistent accuracy across all proportions of unlabeled data, the unsupervised pre-trained method fails to capture positive instances effectively, as evidenced by the F1-score of 0. The ROC curves further confirm this observation, with AUC values hovering around 0.5, indicating poor discriminative ability. This suggests that while the method efficiently processes unlabeled data, it struggles to learn meaningful patterns for positive instances.

In summary, while accuracy provides a useful measure of overall performance, the analysis of ROC curves offers deeper insights into the discriminative ability of classifiers in distinguishing between classes. The observed trends highlight the importance of considering both accuracy and ROC curve metrics when evaluating semi-supervised learning methods, as they provide complementary perspectives on classifier performance.

The proportion of unlabeled data can also influence the results of the classifier. It’s more suitable to have more labelled data but not only to look at the accuracy because it’s not always a good indicator to justify that a model performs very well.