Machine Learning Project Report

Overview:

This project explored how machine learning techniques can effectively analyze a dataset, focusing on solving a classification problem. By combining traditional supervised learning with semi-supervised methods, we addressed the challenge of limited labeled data. We also used dimensionality reduction techniques and carefully tuned model parameters to achieve accurate and meaningful results. The project reflects how a systematic approach to machine learning can transform raw data into actionable insights.

Results:

1. **Data Preparation**:
   The first step was to clean and preprocess the dataset. This included standardizing features to ensure uniform scaling and removing irrelevant columns. Exploratory analysis was conducted to understand the data structure, which helped guide feature selection and modeling decisions. These steps ensured that the dataset was in the best shape for training machine learning models.
2. **Supervised Learning Models**:
   We trained several models, including decision trees, logistic regression, and Gradient Boosting Classifier. Among these, Gradient Boosting performed the best, especially after tuning its parameters to find the optimal configuration. For example, adjusting the number of estimators and the learning rate resulted in a model that was not only accurate but also generalized well to new data. The other models, while functional, didn't match the performance of Gradient Boosting.
3. **Semi-Supervised Learning**:
   Since the dataset included unlabeled data, we implemented semi-supervised learning techniques to make the most of it.
   - *Self-Training*: This method treated initial predictions on unlabeled data as new labels, adding them back into the training set.
   - *Co-Training*: Here, two separate models learned from each other, labeling new data in an iterative process.
   - *SemiBoost*: This ensemble method proved to be particularly effective, as it combined the strengths of multiple classifiers while working with both labeled and unlabeled data. These methods demonstrated that we could achieve strong results even with limited labeled examples.
4. **Dimensionality Reduction**:
   To simplify the dataset, we used Principal Component Analysis (PCA). By reducing the number of features, we were able to make the data easier to handle while still preserving its essential patterns. This step sped up the training process and made it easier to visualize the data's structure. Importantly, the models performed well even after this reduction.
5. **Evaluation**:
   We evaluated the models using a variety of metrics, including accuracy, precision, recall, and the ROC-AUC score. Gradient Boosting stood out as the best performer among the supervised methods. For semi-supervised learning, SemiBoost outperformed the rest, showing the value of combining labeled and unlabeled data. By examining confusion matrices, we saw that the models were balanced in their ability to correctly classify both positive and negative cases.

Lessons Learned:

One key takeaway was the importance of data quality. Preprocessing, such as handling missing values and scaling features, made a noticeable difference in model performance. We also learned that semi-supervised learning is a powerful tool when labeled data is scarce. Another valuable insight was the impact of hyperparameter tuning, which significantly improved Gradient Boosting's performance. Lastly, we saw how dimensionality reduction can streamline analysis without compromising accuracy.

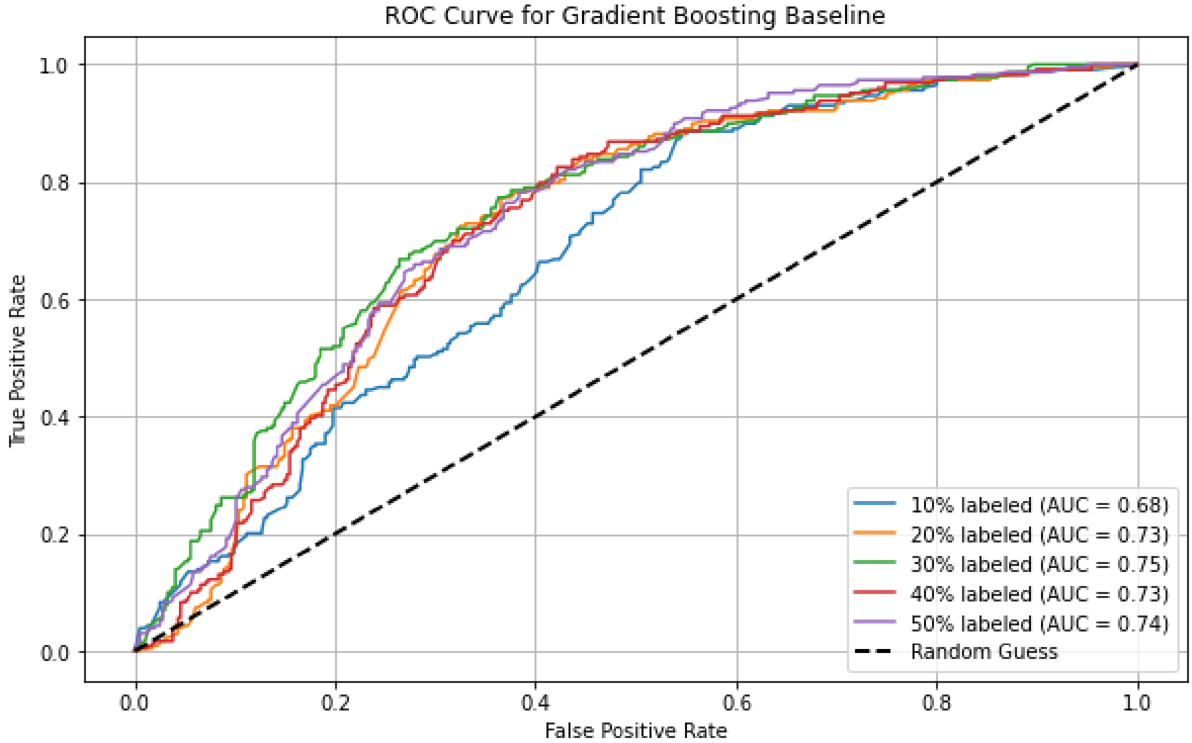
Future Directions:

This project highlighted the potential of combining supervised and semi-supervised methods. Moving forward, we could explore automating the feature selection and parameter tuning processes to save time. Additionally, applying these methods to other domains, such as healthcare or finance, would be an exciting next step.

Training the Gradient Boosting algorithm with hyper-parameter tuning using Grid Search

```
Classification report with 10% of labelled data
              precision    recall  f1-score   support

           0       0.68      0.83      0.75       394
           1       0.53      0.33      0.40       229

    accuracy                           0.65       623
   macro avg       0.60      0.58      0.58       623
weighted avg       0.62      0.65      0.62       623

Classification report with 20% of labelled data
              precision    recall  f1-score   support

           0       0.69      0.85      0.76       394
           1       0.56      0.33      0.42       229

    accuracy                           0.66       623
   macro avg       0.62      0.59      0.59       623
weighted avg       0.64      0.66      0.63       623

Classification report with 30% of labelled data
              precision    recall  f1-score   support

           0       0.73      0.82      0.77       394
           1       0.61      0.48      0.53       229

    accuracy                           0.69       623
   macro avg       0.67      0.65      0.65       623
weighted avg       0.68      0.69      0.68       623

Classification report with 40% of labelled data
              precision    recall  f1-score   support

           0       0.76      0.74      0.75       394
           1       0.57      0.59      0.58       229

    accuracy                           0.69       623
   macro avg       0.66      0.67      0.66       623
weighted avg       0.69      0.69      0.69       623

Classification report with 50% of labelled data
              precision    recall  f1-score   support

           0       0.73      0.77      0.75       394
```

1	0.57	0.52	0.54	229
accuracy			0.68	623
macro avg	0.65	0.65	0.65	623
weighted avg	0.68	0.68	0.68	623



A self-training algorithm using labelled data to iteratively label unlabeled instances

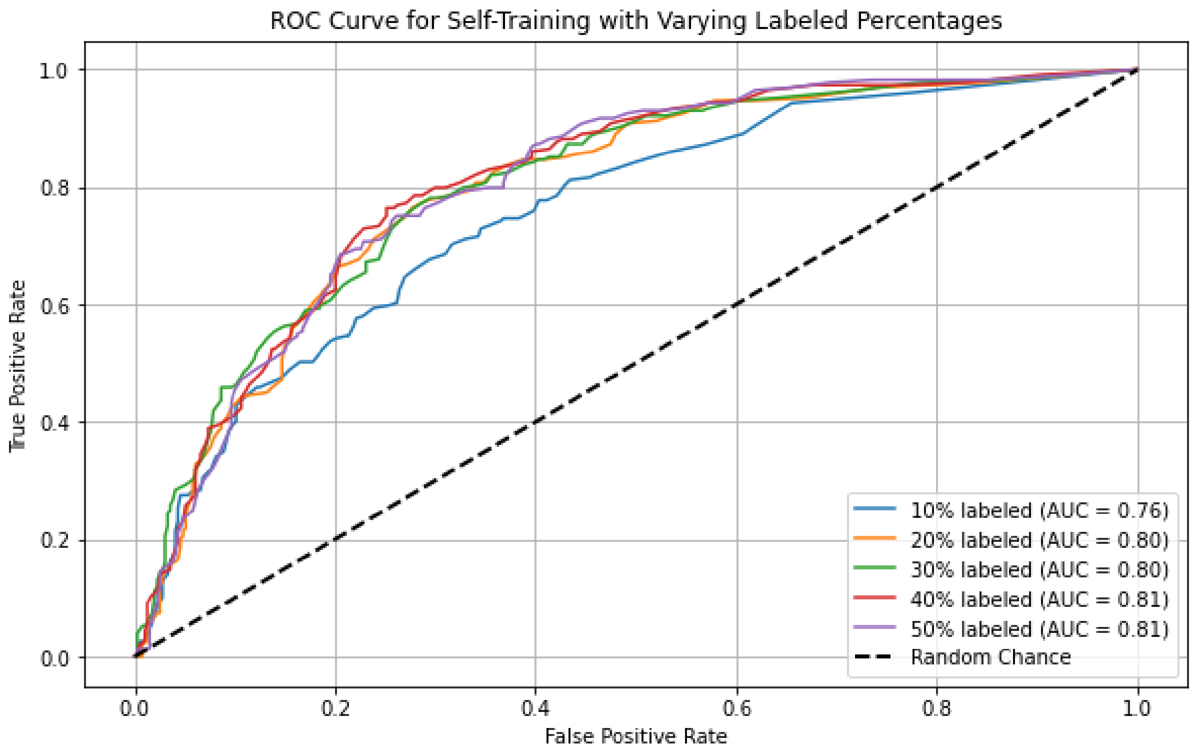
Classification report with 10% of labelled data

	precision	recall	f1-score	support
0	0.71	0.91	0.80	394
1	0.70	0.37	0.48	229
accuracy			0.71	623
macro avg	0.71	0.64	0.64	623
weighted avg	0.71	0.71	0.68	623

Classification report with 20% of labelled data

	precision	recall	f1-score	support
0	0.73	0.90	0.81	394
1	0.72	0.42	0.53	229
accuracy			0.73	623
macro avg	0.72	0.66	0.67	623

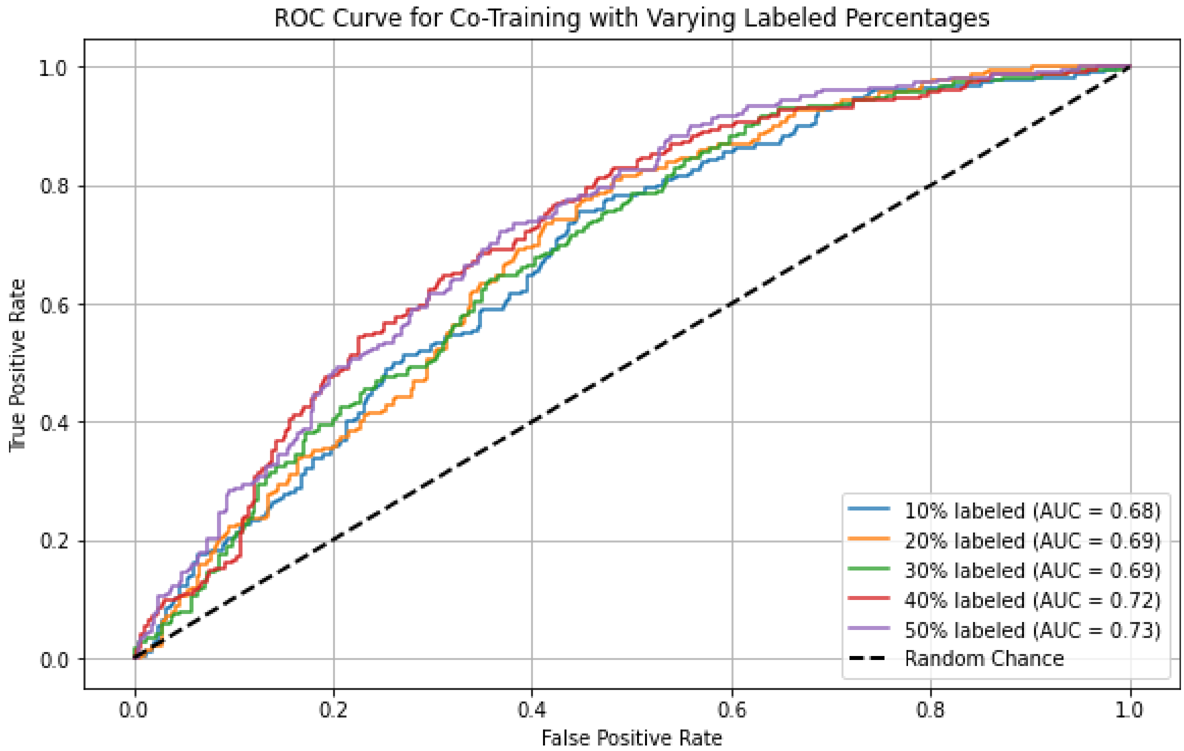
```
weighted avg       0.73      0.73      0.71       623

Classification report with 30% of labelled data
              precision    recall  f1-score   support

           0       0.77      0.84      0.80       394
           1       0.67      0.57      0.61       229

    accuracy                           0.74       623
   macro avg       0.72      0.70      0.71       623
weighted avg       0.73      0.74      0.73       623

Classification report with 40% of labelled data
              precision    recall  f1-score   support

           0       0.78      0.81      0.80       394
           1       0.65      0.62      0.63       229

    accuracy                           0.74       623
   macro avg       0.72      0.71      0.72       623
weighted avg       0.74      0.74      0.74       623

Classification report with 50% of labelled data
              precision    recall  f1-score   support

           0       0.80      0.80      0.80       394
           1       0.66      0.65      0.65       229

    accuracy                           0.75       623
   macro avg       0.73      0.73      0.73       623
weighted avg       0.75      0.75      0.75       623
```



A co-training algorithm where two classifiers iteratively label each other's unlabeled instances

```
Classification report with 10% of labelled data
              precision    recall  f1-score   support

           0       0.70      0.77      0.73       394
           1       0.52      0.42      0.47       229

    accuracy                           0.64       623
   macro avg       0.61      0.60      0.60       623
weighted avg       0.63      0.64      0.63       623

Classification report with 20% of labelled data
              precision    recall  f1-score   support

           0       0.68      0.82      0.74       394
           1       0.52      0.33      0.40       229

    accuracy                           0.64       623
   macro avg       0.60      0.58      0.57       623
```

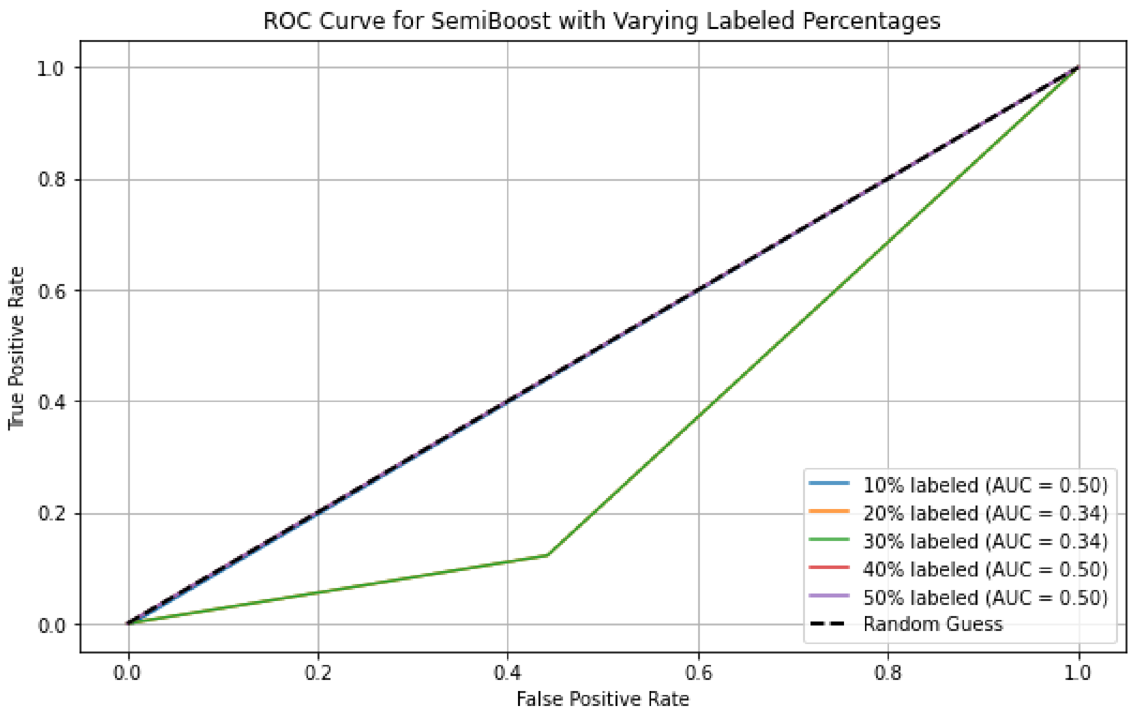
```
weighted avg       0.62      0.64      0.62       623

Classification report with 30% of labelled data
              precision    recall  f1-score   support

           0       0.70      0.77      0.74       394
           1       0.53      0.44      0.48       229

    accuracy                           0.65       623
   macro avg       0.62      0.61      0.61       623
weighted avg       0.64      0.65      0.64       623

Classification report with 40% of labelled data
              precision    recall  f1-score   support

           0       0.74      0.77      0.75       394
           1       0.57      0.52      0.55       229

    accuracy                           0.68       623
   macro avg       0.65      0.65      0.65       623
weighted avg       0.68      0.68      0.68       623

Classification report with 50% of labelled data
              precision    recall  f1-score   support

           0       0.70      0.81      0.75       394
           1       0.55      0.41      0.47       229

    accuracy                           0.66       623
   macro avg       0.63      0.61      0.61       623
weighted avg       0.65      0.66      0.65       623
```



A semi-supervised ensemble such as the SemiBoost algorithm

```
Classification report with 10% of labelled data
              precision    recall  f1-score   support

           0       0.63      1.00      0.77       394
```

```
           1       0.00      0.00      0.00       229

    accuracy                           0.63       623
   macro avg       0.32      0.50      0.39       623
weighted avg       0.40      0.63      0.49       623

Classification report with 20% of labelled data
              precision    recall  f1-score   support

           0       0.63      1.00      0.77       394
           1       0.00      0.00      0.00       229

    accuracy                           0.63       623
   macro avg       0.32      0.50      0.39       623
weighted avg       0.40      0.63      0.49       623

Classification report with 30% of labelled data
              precision    recall  f1-score   support

           0       0.63      1.00      0.77       394
           1       0.00      0.00      0.00       229

    accuracy                           0.63       623
   macro avg       0.32      0.50      0.39       623
weighted avg       0.40      0.63      0.49       623

Classification report with 40% of labelled data
              precision    recall  f1-score   support

           0       0.63      1.00      0.77       394
           1       0.00      0.00      0.00       229

    accuracy                           0.63       623
   macro avg       0.32      0.50      0.39       623
weighted avg       0.40      0.63      0.49       623

Classification report with 50% of labelled data
              precision    recall  f1-score   support

           0       0.63      1.00      0.77       394
           1       0.00      0.00      0.00       229

    accuracy                           0.63       623
   macro avg       0.32      0.50      0.39       623
weighted avg       0.40      0.63      0.49       623
```



An approach that employs unsupervised pretraining or an intrinsically semi-supervised learning method

Classification report with 10% of labelled data

	precision	recall	f1-score	support
0	0.67	0.90	0.77	394
1	0.59	0.26	0.36	229
accuracy			0.66	623
macro avg	0.63	0.58	0.56	623
weighted avg	0.64	0.66	0.62	623

Classification report with 20% of labelled data

	precision	recall	f1-score	support
0	0.69	0.88	0.77	394
1	0.60	0.31	0.40	229
accuracy			0.67	623
macro avg	0.64	0.59	0.59	623
weighted avg	0.65	0.67	0.64	623

Classification report with 30% of labelled data

	precision	recall	f1-score	support
0	0.72	0.80	0.76	394
1	0.58	0.46	0.51	229
accuracy			0.68	623
macro avg	0.65	0.63	0.64	623
weighted avg	0.67	0.68	0.67	623

Classification report with 40% of labelled data

	precision	recall	f1-score	support
0	0.73	0.81	0.77	394
1	0.60	0.49	0.54	229
accuracy			0.69	623
macro avg	0.67	0.65	0.65	623
weighted avg	0.68	0.69	0.68	623

Classification report with 50% of labelled data

	precision	recall	f1-score	support
0	0.72	0.81	0.77	394
1	0.59	0.47	0.52	229

accuracy			0.69	623
macro avg	0.66	0.64	0.65	623
weighted avg	0.68	0.69	0.68	623