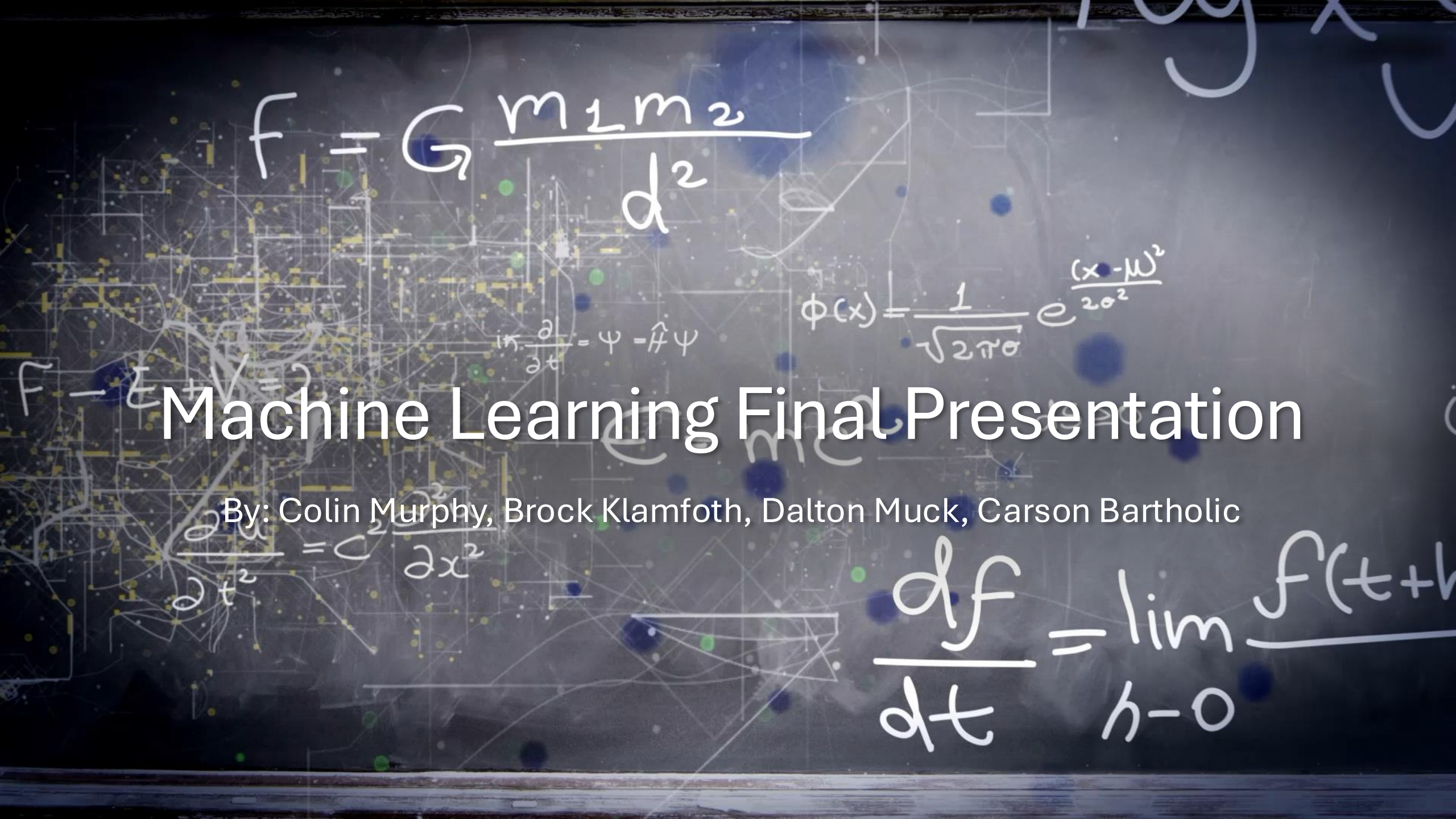


Machine Learning Final Presentation

By: Colin Murphy, Brock Klamfoth, Dalton Muck, Carson Bartholic

Research Question

- Does combinational linear separability of individual features reduced to two dimensions yield an increase in accuracy when applied to supervised learning?

https://drive.google.com/file/d/1ihIpkdWM_BPVOFXVoz63fD0dn3xw08mE/view

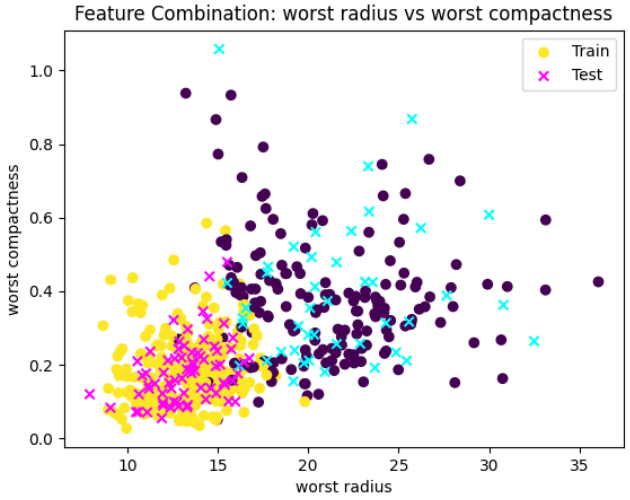
Flowchart

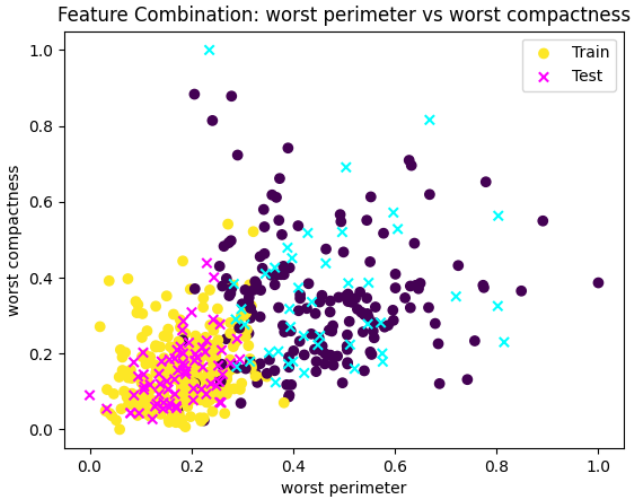
Feature Selection

- Analyze **linear separability** between every 2 features.
- Train **Perceptron** model on every combination, then evaluate accuracy.
- Define an accuracy threshold = **0.96**.
- Use **accuracy threshold** to select features that were part of a comparison >= to the accuracy threshold.

Best Accuracy

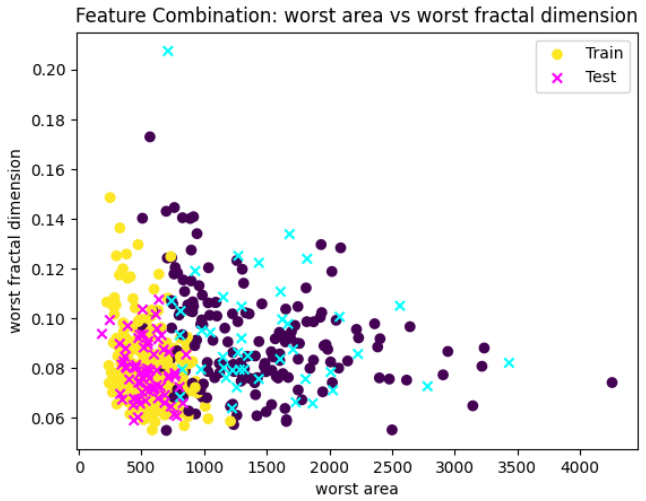
- No Scaling: Worst radius vs. Worst compactness -> 0.96
- Min-Max Scaling: Worst perimeter vs. Worst compactness -> 0.99

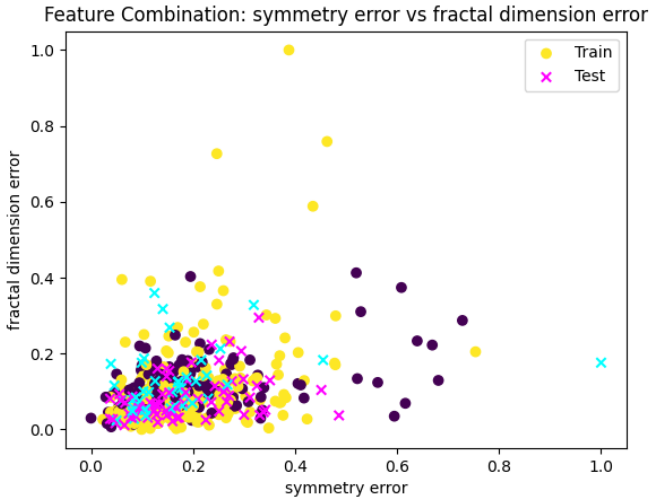




Worst Accuracy

- No Scaling: Worst area vs. Worst fractal dimension -> 0.37
- Min-Max Scaling: Symmetry error vs. Fractal dimension error -> 0.37



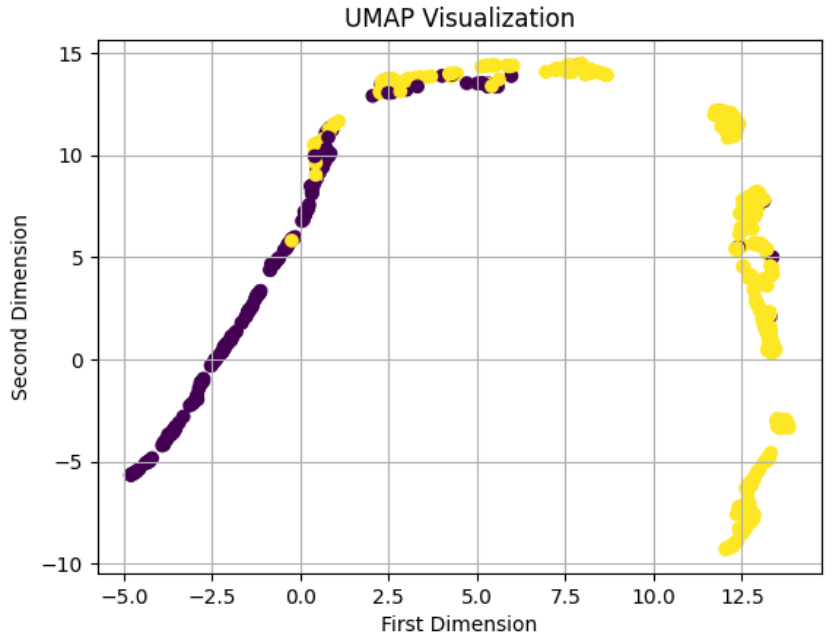


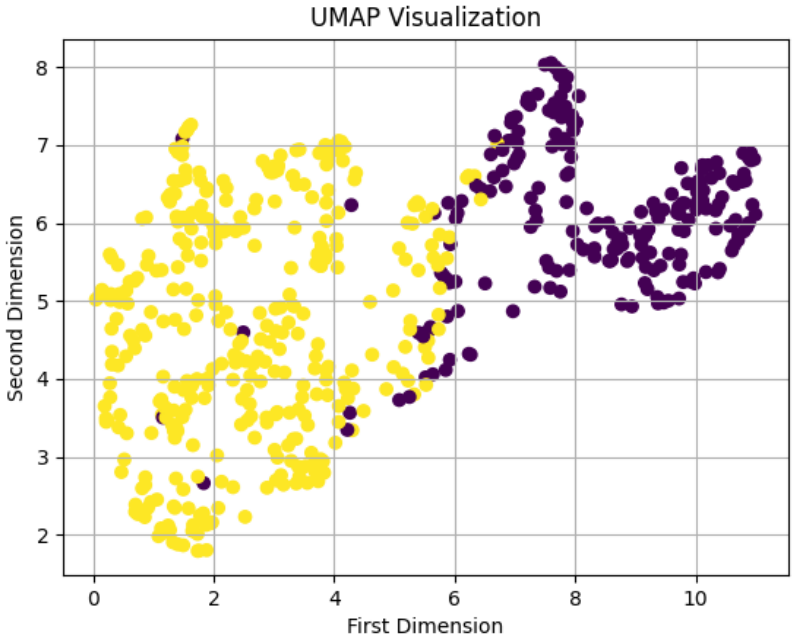
Dimensionality Reduction

- Make models of **UMAP**, **PCA**, and **TSNE** to use for dimensionality reduction.
- Fit each model with the data after the data has been through the feature selection process.
- The following values led to the best/most appropriate results:
  - Choose n_components = **2**
  - Choose random_state = **42** (For UMAP only)
- The main takeaway is that the data is reduced to **2** dimensions.
- Results were both stored as graphs for visualization purposes. They were also stored as CSV files to be used in the next process.

UMAP Model

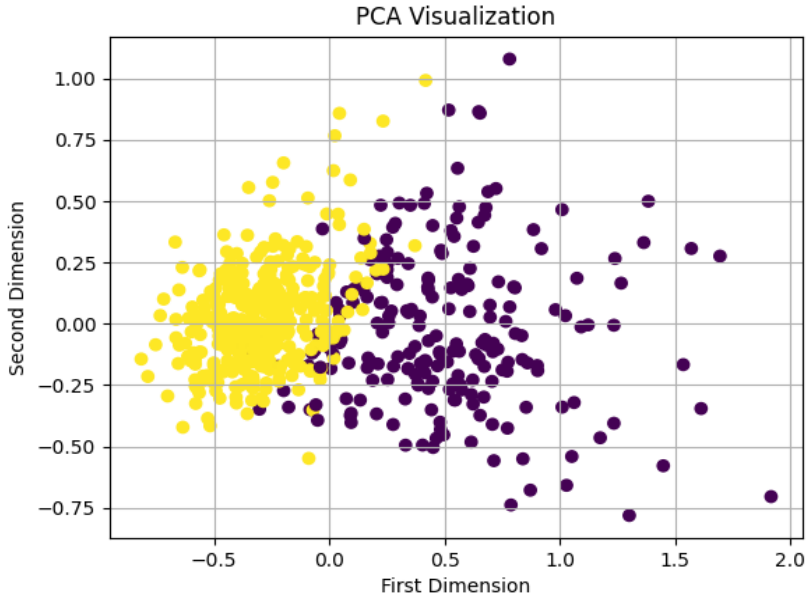
- Using n_compents = 2 and random_state = 42, the following UMAP models were created.
- Graph Visualizations show the effects of no scaling (Top Graph) and min/max scaling (Bottom Graph).

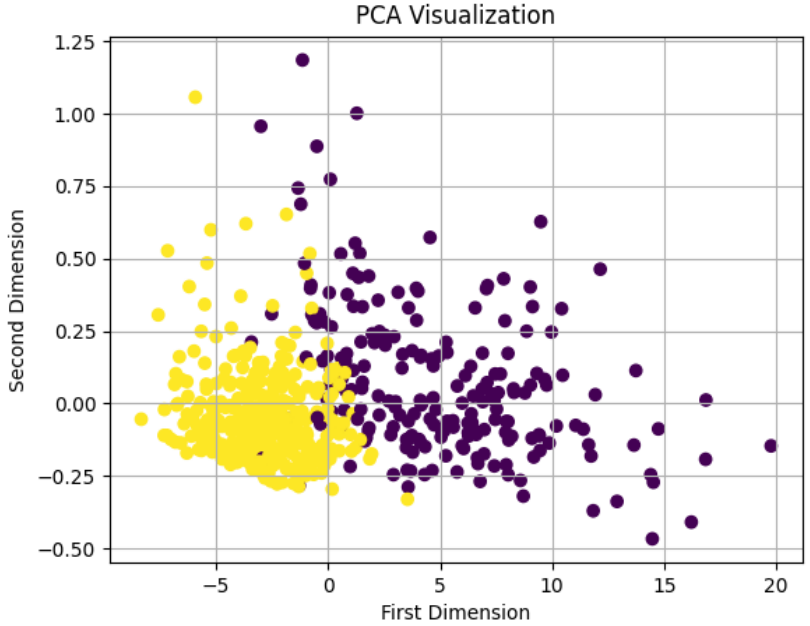




PCA Model

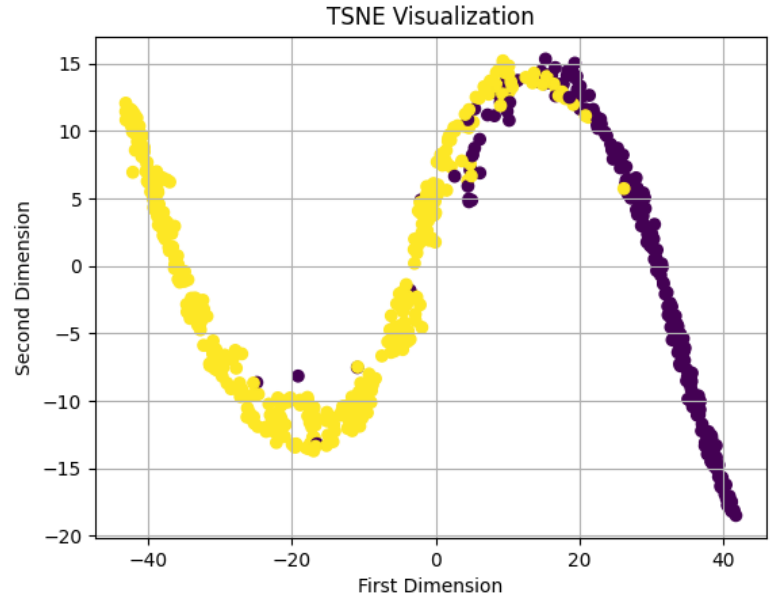
- Using n_compents = 2, the following PCA models were created.
- Graph Visualizations show the effects of no scaling (Top Graph) and min/max scaling (Bottom Graph).

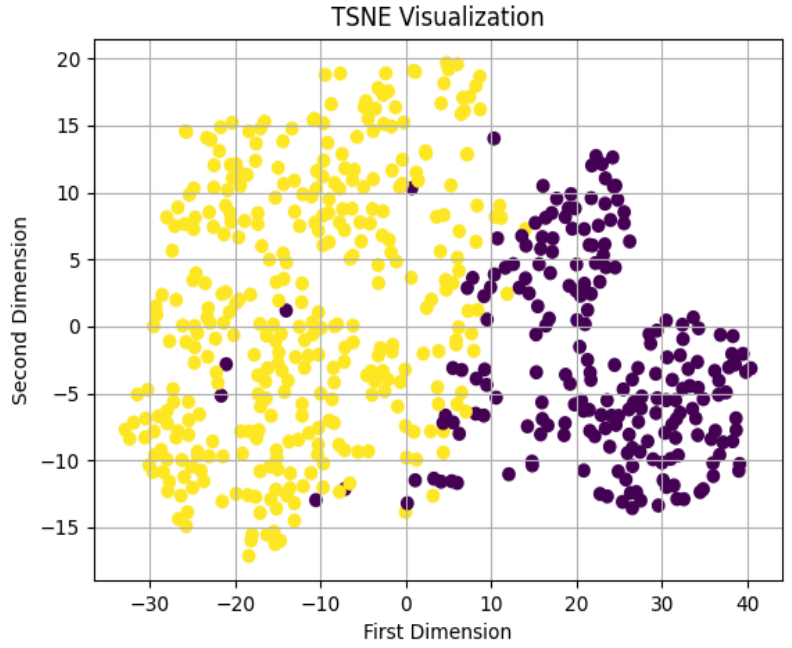




TSNE Model

- Using n_compents = 2, the following TSNE models were created.
- Graph Visualizations show the effects of no scaling (Top Graph) and min/max scaling (Bottom Graph).





Grid Search SVM

```
Best SVM Parameters: {'C': 1, 'coef0': 3.0, 'degree': 4, 'gamma': 'auto', 'kernel': 'poly', 'tol': 0.0001}
############################Grid Search in ./results/umap_reduced_data.csv############################
Best SVM Score From Grid Search: 96.92307692307693 %
              precision    recall  f1-score   support

           0       0.92      0.85      0.89        41
           1       0.92      0.96      0.94        73

    accuracy                           0.92       114
   macro avg       0.92      0.91      0.91       114
weighted avg       0.92      0.92      0.92       114

Best SVM Parameters: {'C': 1, 'coef0': 2.0, 'degree': 4, 'gamma': 0.05, 'kernel': 'poly', 'tol': 0.0001}
############################Grid Search in ./results/pca_reduced_data.csv############################
Best SVM Score From Grid Search: 94.28571428571428 %
              precision    recall  f1-score   support

           0       0.95      0.85      0.90        41
           1       0.92      0.97      0.95        73

    accuracy                           0.93       114
   macro avg       0.93      0.91      0.92       114
weighted avg       0.93      0.93      0.93       114
```

Accuracy SVM Default

Before Dimensionality Reduction, Feature Selection, and parameter optimization

Accuracy = 93.86%

```
#########################SVM with Default Data#########################
{'C': 1.0, 'break_ties': False, 'cache_size': 200, 'class_weight': None, 'coef0': 0.0, 'decision_function_shape': 'ovr', 'degree': 3, 'gamma': 'scale', 'kernel': 'rbf', 'max_iter': -1, 'proba
bility': False, 'random_state': None, 'shrinking': True, 'tol': 0.001, 'verbose': False}
Accuracy: 93.86%
```

Accuracy SVM tSNE Reduction

T-distributed Stochastic Embedding (TSNE) Before Scaling: 92.89%

Best Scores After Scaling (Achieved during grid search): 92.98%

```
Processing file: ./results/tsne_reduced_data.csv
############################SVM in ./results/tsne_reduced_data.csv############################
SVM Parameters: {'C': 1, 'coef0': 3.0, 'gamma': 'auto', 'kernel': 'poly', 'tol': 0.0001}
Accuracy for ./results/tsne_reduced_data.csv: 92.98%
```

*The accuracy data in the image above was output using random state 35
This state was chosen to keep all data splits consistent across different runs

Accuracy SVM PCA Reduction

Principal Component Analysis (PCA) Before Scaling: 92.89%

Best Scores After Scaling (Achieved during grid search): 94.74%

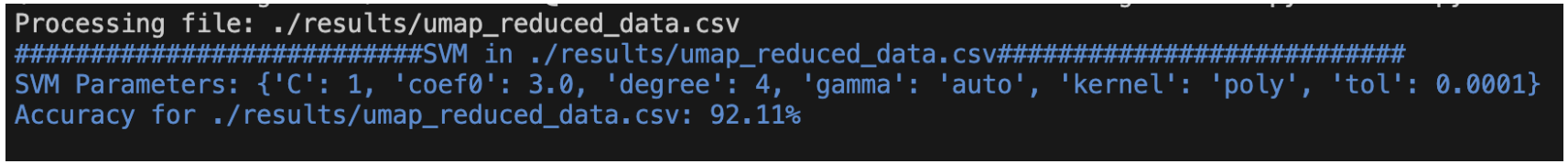
```
Processing file: ./results/pca_reduced_data.csv
############################SVM in ./results/pca_reduced_data.csv############################
SVM Parameters: {'C': 1, 'coef0': 2.0, 'degree': 4, 'gamma': '0.05', 'kernel': 'poly', 'tol': 0.0001}
Accuracy for ./results/pca_reduced_data.csv: 94.74%
```

*The accuracy data in the image above was output using random state 35
This state was chosen to keep all data splits consistent across different runs

Accuracy SVM UMAP Reduction

Uniform Manifold Approximation and Projection (Umap) Before Scaling: 92.89%

Best Scores After Scaling (Achieved during grid search): 96.92%



*The lower accuracy data in the image above was output using random state 35
This state was chosen to keep all data splits consistent across different runs

Accuracy SVM Summary

Before Dimensionality Reduction, Feature Selection, and parameter optimization : *93.86%*

Before Scaling

T-distributed Stochastic Embedding (TSNE) : *92.89%*

Principal Component Analysis (PCA) : *92.89%*

Uniform Manifold Approximation and Projection (Umap) : *92.89%*

After Scaling

T-distributed Stochastic Embedding (TSNE) : *92.89%*

Principal Component Analysis (PCA) : *94.74%*

Uniform Manifold Approximation and Projection (Umap) : *96.92%*

Accuracy Perceptron Summary

Before Dimensionality Reduction and Feature Selection

89.47%

Before Scaling	After Scaling	After Grid Search
T- distributed Stochastic Embedding (TSNE) 94.74	TSNE 92.98%	TSNE 92.98%
Principal Component Analysis (PCA) 92.98	PCA 93.86%	PCA 93.86%
Uniform Manifold Approximationand Projection (Umap) 79.82%	UMAP 91.23%	UMAP 93.86%

Grid Search Perceptron

```
Best Parameters for ./results/umap_reduced_data.csv: {'alpha': 0.01, 'eta0': 0.001, 'max_iter': 500, 'penalty': 'elasticnet', 'tol': 0.0001}
Accuracy for ./results/umap_reduced_data.csv with best parameters: 93.86%
Classification Report for ./results/umap_reduced_data.csv:
              precision    recall  f1-score   support

           0     0.9048    0.9268    0.9157        41
           1     0.9583    0.9452    0.9517        73

    accuracy                         0.9386       114
   macro avg     0.9315    0.9360    0.9337       114
weighted avg     0.9391    0.9386    0.9388       114


Processing file: ./results/tsne_reduced_data.csv
Best Parameters for ./results/tsne_reduced_data.csv: {'alpha': 0.0001, 'eta0': 1.0, 'max_iter': 500, 'penalty': None, 'tol': 0.0001}
Accuracy for ./results/tsne_reduced_data.csv with best parameters: 92.98%
Classification Report for ./results/tsne_reduced_data.csv:
              precision    recall  f1-score   support

           0     0.8837    0.9268    0.9048        41
           1     0.9577    0.9315    0.9444        73

    accuracy                         0.9298       114
   macro avg     0.9207    0.9292    0.9246       114
weighted avg     0.9311    0.9298    0.9302       114


Processing file: ./results/pca_reduced_data.csv
Best Parameters for ./results/pca_reduced_data.csv: {'alpha': 0.0001, 'eta0': 1.0, 'max_iter': 500, 'penalty': None, 'tol': 0.01}
Accuracy for ./results/pca_reduced_data.csv with best parameters: 93.86%
Classification Report for ./results/pca_reduced_data.csv:
              precision    recall  f1-score   support

           0     1.0000    0.8293    0.9067        41
           1     0.9125    1.0000    0.9542        73

    accuracy                         0.9386       114
   macro avg     0.9563    0.9146    0.9305       114
weighted avg     0.9440    0.9386    0.9371       114
```