Optimization of Nonconvex Penalties

16:954:588:01 Financial Data Mining

Team 6
Adish Golechha
FNU Vasureddy
Pradhyumna Kasula
Shreyash Kalal

Introduction

- In various scientific fields, like biology, solving the problem of selecting the right variables for regression is vital. This means picking out the most important predictors from a big group of variables to accurately predict outcomes.
- To tackle this, nonconvex penalty functions such as the Smoothly Clipped Absolute Deviation (SCAD) and Minimax Concave Penalty (MCP) are utilized due to their favorable theoretical properties.
- Efficient optimization algorithms are introduced to fit models with these penalties, offering both theoretical convergence assurances and notable speed advantages over alternative methods

Non-Convex Penalties

Nonconvex penalties are functions used in regression to enforce sparsity in the coefficient estimates. They penalize large coefficient values more severely than small ones, encouraging simpler models with fewer predictors. This encourages variable selection by effectively shrinking less important coefficients to zero, thus excluding irrelevant predictors from the model.

- Smoothly Clipped Absolute Deviation (SCAD)
- Minimax Concave Penalty (MCP)

In this project, we will focus on these 2 non convex penalties and explore the advantages and disadvantages of using them.

SCAD

1

- The SCAD penalty is a nonconvex regularization method used in regression analysis to promote sparsity in coefficient estimates.
- Unlike convex penalties, SCAD smoothly transitions from penalizing small coefficients, similar to the L1 penalty, to imposing a more severe penalty on larger coefficients.
- SCAD's mathematical formulation, controlled by a regularization parameter (λ), enables efficient optimization despite its nonconvex nature. Its utility extends across various fields, making it a popular choice for balancing bias reduction, sparsity, and computational efficiency in regression modeling tasks.

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{p} \psi_\lambda(|\beta_j|)$$

MCP

2

- MCP is a regularization method commonly employed in regression analysis to promote sparsity in coefficient estimates while maintaining flexibility in model fitting.
- It smoothly adjusts penalties, favoring smaller coefficients while imposing stricter penalties on larger ones, regulated by a regularization and a tuning parameter.
- Despite its nonconvex nature, MCP penalties can be efficiently optimized, ensuring computational scalability in high-dimensional data settings.

$$P_\lambda(\beta) = \lambda \sum_{j=1}^{p} \left(\sqrt{1 + \left(\frac{\beta_j}{\gamma\lambda}\right)^2} - 1 \right)$$

Optimization Algorithms

Optimization algorithms are computational methods used to minimize or maximize a mathematical function. They iteratively adjust the parameters of the function to find the best solution, often referred to as the optimal solution.

These algorithms tailored for nonconvex penalty functions enable efficient and accurate parameter estimation in regression models, crucial for variable selection and model fitting in data analysis.

- Coordinate Descent Algorithm
- Stochastic Descent Algorithm
- Proximal Descent Algorithm

Coordinate Descent

1

- Coordinate Descent is an optimization algorithm used to minimize a function by iteratively updating one variable while keeping the others fixed.
- Coordinate Descent is particularly useful when the objective function is non-convex and when optimizing with respect to each variable separately is more efficient than optimizing with respect to all variables simultaneously.

$$\beta_j^{(t+1)} = \arg\min_{\beta_j} f(\beta_1^{(t)}, \beta_2^{(t)}, \ldots, \beta_j, \ldots, \beta_p^{(t)})$$

Stochastic Descent

2

- Stochastic Descent is an optimization algorithm used to minimize a function by randomly sampling data points and updating the parameters based on the gradient of the objective function with respect to that sampled data point.
- Instead of computing the gradient using the entire dataset, SGD uses a single or a small subset of data points to compute an approximate gradient.

$$\beta^{(t+1)} = \beta^{(t)} - \eta \nabla f_i(\beta^{(t)})$$

Proximal Descent

3

- Proximal Descent is an optimization algorithm used to minimize a function with a combination of smooth and non-smooth terms.
- Proximal Descent updates the parameters by taking a gradient descent step and then applying the proximal operator to enforce penalties or constraints.
- This approach allows Proximal Descent to efficiently handle non-smooth or non-convex objective functions, making it suitable for a wide range of optimization problems

$$\beta^{(t+1)} = \mathrm{prox}_{\lambda g}(\beta^{(t)} - \eta \nabla f(\beta^{(t)}))$$

Dataset

The dataset comprises features extracted from digitized images of fine needle aspirates (FNA) of breast masses. These features characterize the nuclei present in the images. The dataset includes 10 real-valued features for each cell nucleus, such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Features are computed based on the mean, standard error, and "worst" values, resulting in a total of 30 features per image. The dataset consists of 569 instances, with 357 benign and 212 malignant diagnoses. The dataset is publicly available through the UCI Machine Learning Repository.

SCAD Results

Hyperparameter Tuning

```
Params: {'alpha': 0.001, 'gamma': 1.0}, Accuracy: 0.9298
Params: {'alpha': 0.001, 'gamma': 1.5}, Accuracy: 0.9211
Params: {'alpha': 0.001, 'gamma': 2.0}, Accuracy: 0.9123
Params: {'alpha': 0.001, 'gamma': 2.5}, Accuracy: 0.9123
Params: {'alpha': 0.001, 'gamma': 3}, Accuracy: 0.9123
Params: {'alpha': 0.001, 'gamma': 3.5}, Accuracy: 0.9123
Params: {'alpha': 0.001, 'gamma': 4}, Accuracy: 0.9123
Params: {'alpha': 0.01, 'gamma': 1.0}, Accuracy: 0.9298
Params: {'alpha': 0.01, 'gamma': 1.5}, Accuracy: 0.9211
Params: {'alpha': 0.01, 'gamma': 2.0}, Accuracy: 0.9123
Params: {'alpha': 0.01, 'gamma': 2.5}, Accuracy: 0.9123
Params: {'alpha': 0.01, 'gamma': 3}, Accuracy: 0.9123
Params: {'alpha': 0.01, 'gamma': 3.5}, Accuracy: 0.9123
Params: {'alpha': 0.01, 'gamma': 4}, Accuracy: 0.9123
Params: {'alpha': 0.1, 'gamma': 1.0}, Accuracy: 0.9298
Params: {'alpha': 0.1, 'gamma': 1.5}, Accuracy: 0.9211
Params: {'alpha': 0.1, 'gamma': 2.0}, Accuracy: 0.9123
Params: {'alpha': 0.1, 'gamma': 2.5}, Accuracy: 0.9035
Params: {'alpha': 0.1, 'gamma': 3}, Accuracy: 0.9035
Params: {'alpha': 0.1, 'gamma': 3.5}, Accuracy: 0.9035
Params: {'alpha': 0.1, 'gamma': 4}, Accuracy: 0.9035
Best Parameters: {'alpha': 0.001, 'gamma': 1.0}
Best Accuracy: 0.9298245614035088
```

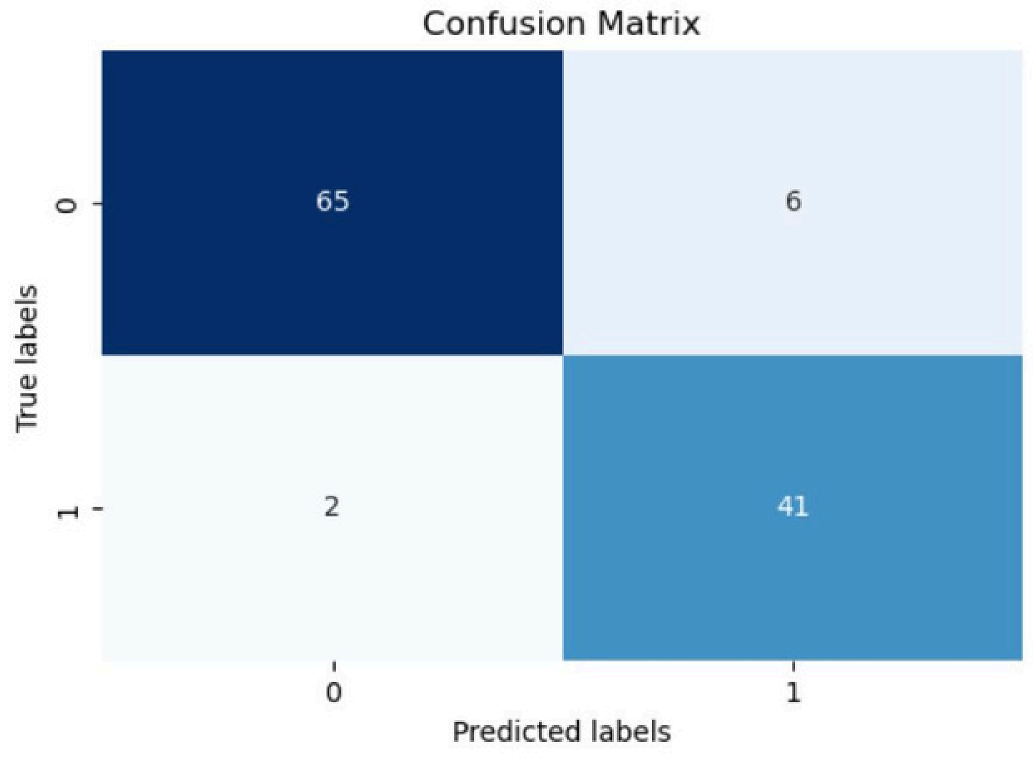
Classification Report

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.92	0.94	71
1	0.87	0.95	0.91	43
accuracy			0.93	114
macro avg	0.92	0.93	0.93	114
weighted avg	0.93	0.93	0.93	114

Confusion Matrix

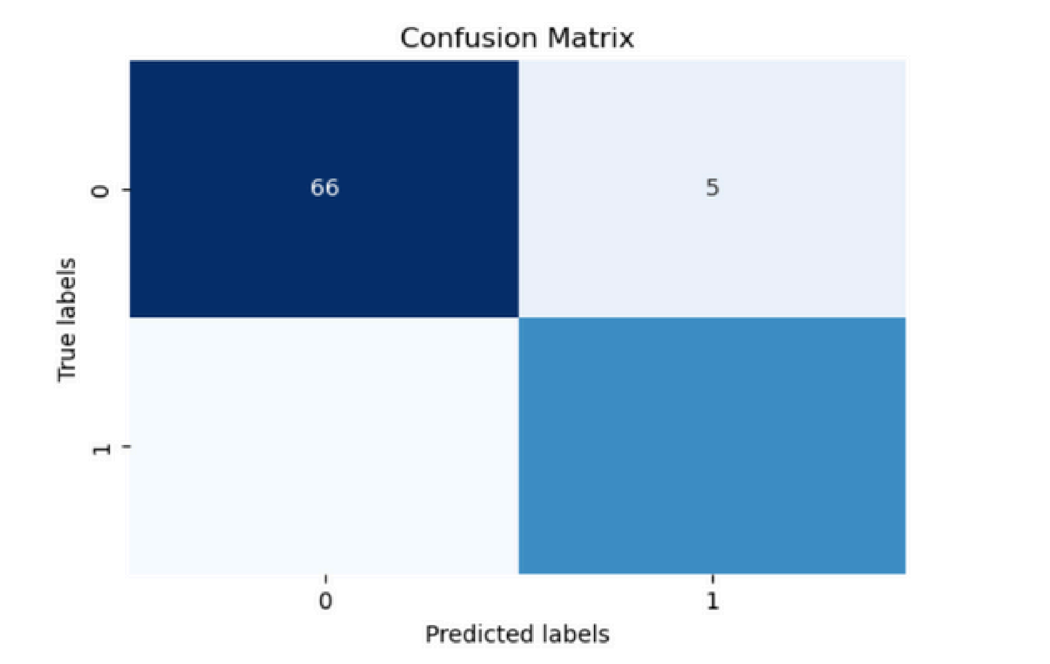
```
Confusion Matrix:
[[65  6]
 [ 2 41]]
```



SCAD with Optimization

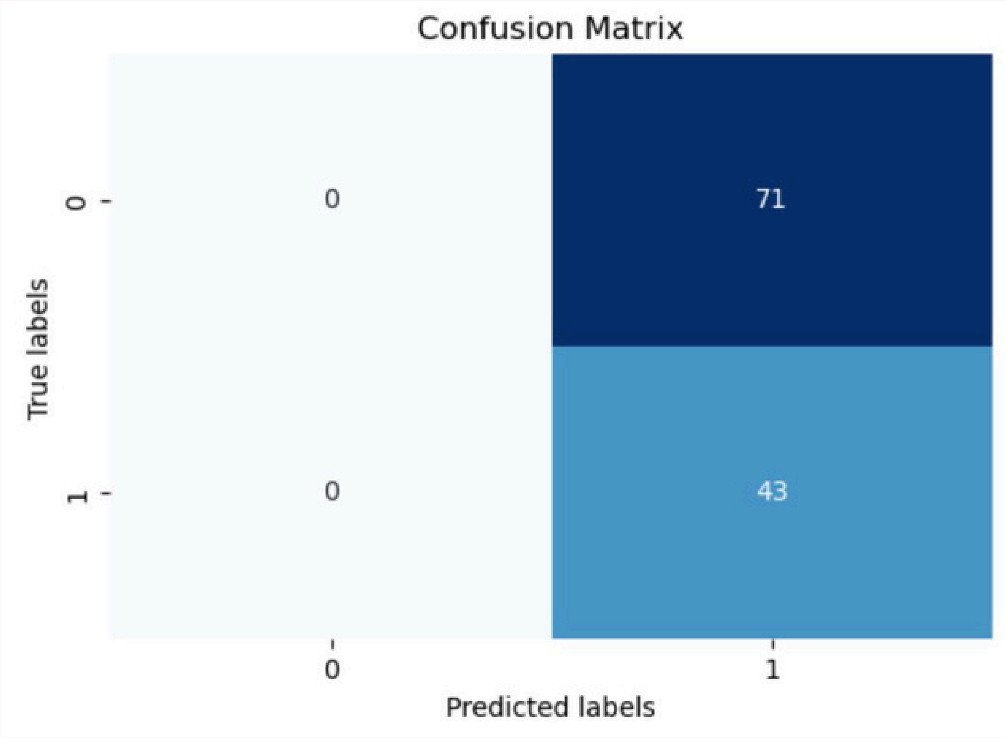
Coordinate Descent

```
Coordinate Descent:
Accuracy: 0.9473684210526315
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.93      0.96        71
           1       0.89      0.98      0.93        43

    accuracy                           0.95       114
   macro avg       0.94      0.95      0.94       114
weighted avg       0.95      0.95      0.95       114
```



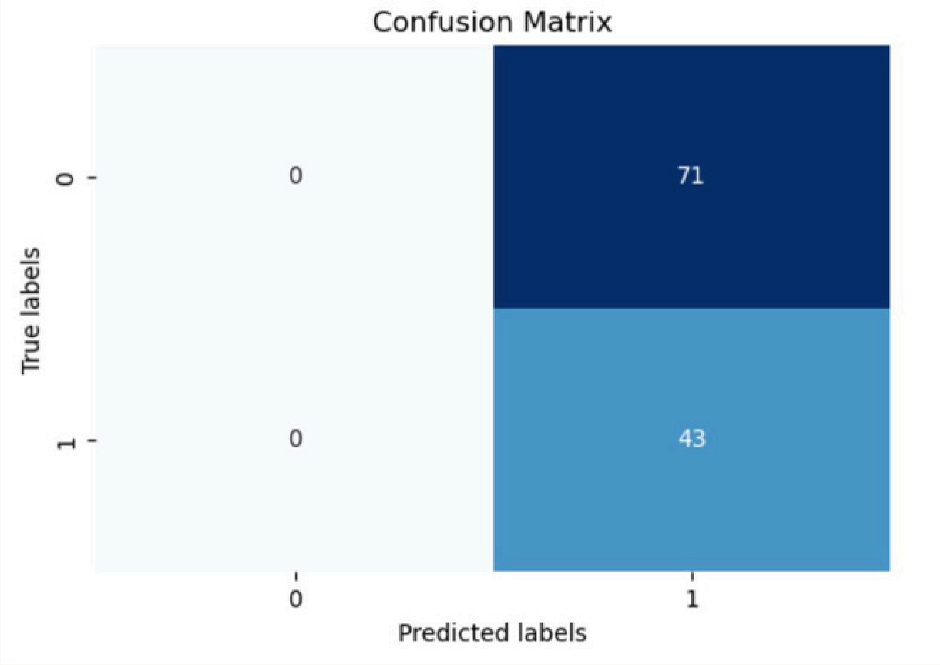
Stochastic Descent

```
Stochastic Gradient Descent:
Accuracy: 0.37719298245614036
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        71
           1       0.38      1.00      0.55        43

    accuracy                           0.38       114
   macro avg       0.19      0.50      0.27       114
weighted avg       0.14      0.38      0.21       114
```



Proximal Descent

```
Proximal Gradient Descent:
Accuracy: 0.37719298245614036
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        71
           1       0.38      1.00      0.55        43

    accuracy                           0.38       114
   macro avg       0.19      0.50      0.27       114
weighted avg       0.14      0.38      0.21       114
```



MCP Results

Hyperparameter Tuning

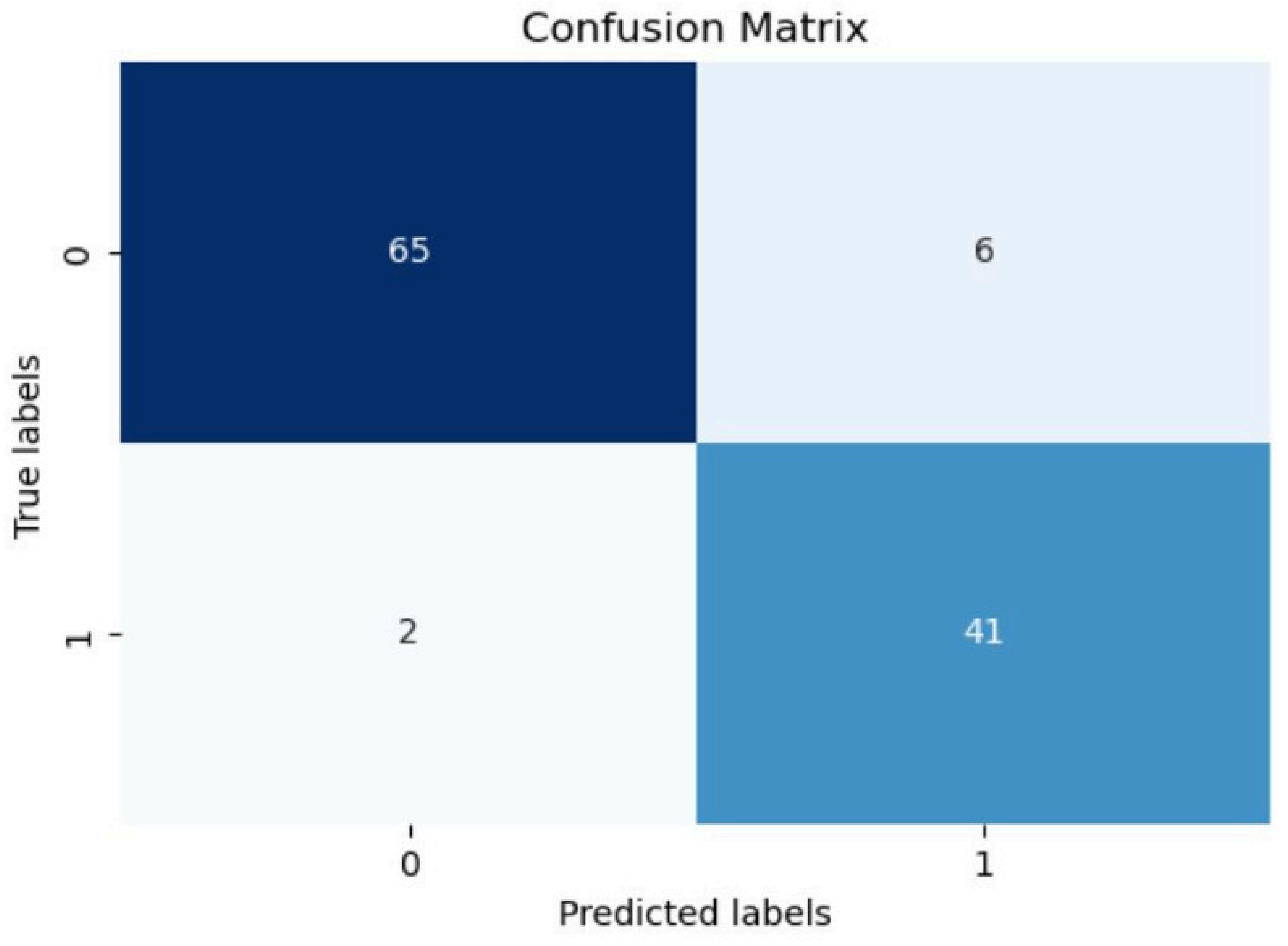
```
Params: {'alpha': 0.01, 'gamma': 2.0}, Accuracy: 0.8947
Params: {'alpha': 0.01, 'gamma': 3.0}, Accuracy: 0.8947
Params: {'alpha': 0.01, 'gamma': 4.0}, Accuracy: 0.8947
Params: {'alpha': 0.1, 'gamma': 2.0}, Accuracy: 0.8772
Params: {'alpha': 0.1, 'gamma': 3.0}, Accuracy: 0.8860
Params: {'alpha': 0.1, 'gamma': 4.0}, Accuracy: 0.8947
Params: {'alpha': 1.0, 'gamma': 2.0}, Accuracy: 0.9123
Params: {'alpha': 1.0, 'gamma': 3.0}, Accuracy: 0.9123
Params: {'alpha': 1.0, 'gamma': 4.0}, Accuracy: 0.9123
Best Parameters: {'alpha': 1.0, 'gamma': 2.0}
Best Accuracy: 0.9122807017543859
```

Classification Report

```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.92      0.94        71
           1       0.87      0.95      0.91        43

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

Confusion Matrix

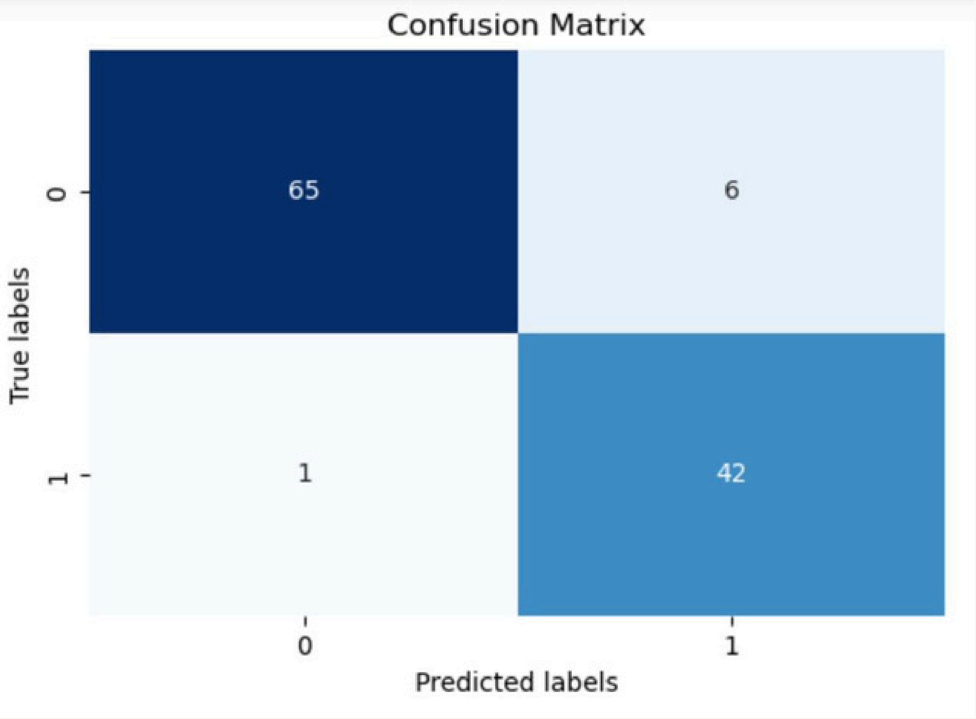
```
Confusion Matrix:
[[65  6]
 [ 2 41]]
```



MCP with Optimization

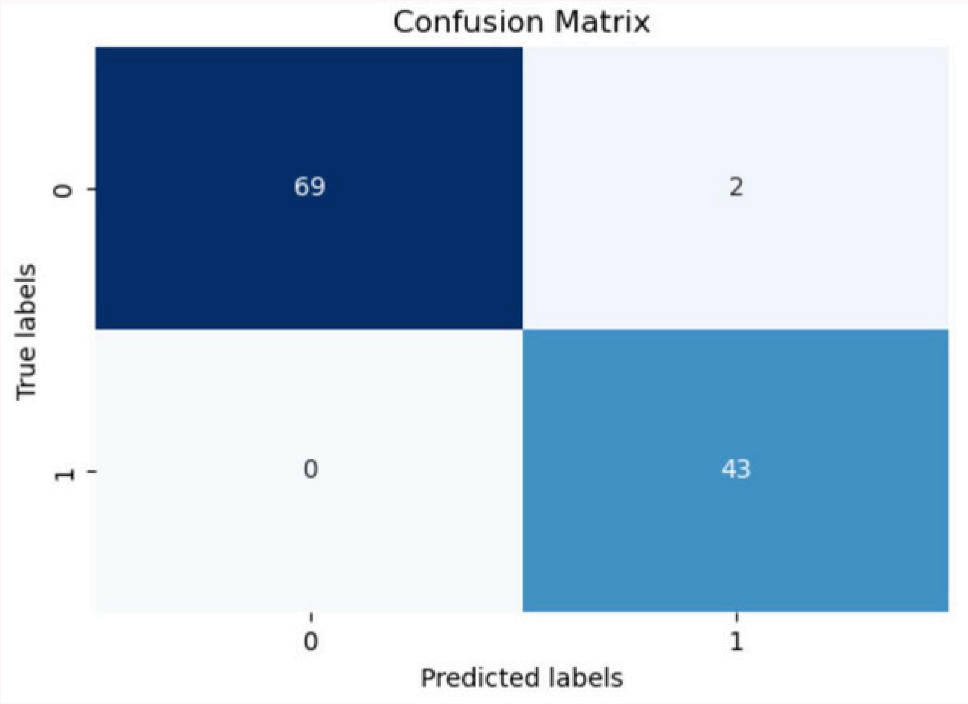
Coordinate Descent

```
Coordinate Descent with MCP Penalty:
Accuracy: 0.9385964912280702
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.92      0.95        71
           1       0.88      0.98      0.92        43

    accuracy                           0.94       114
   macro avg       0.93      0.95      0.94       114
weighted avg       0.94      0.94      0.94       114
```



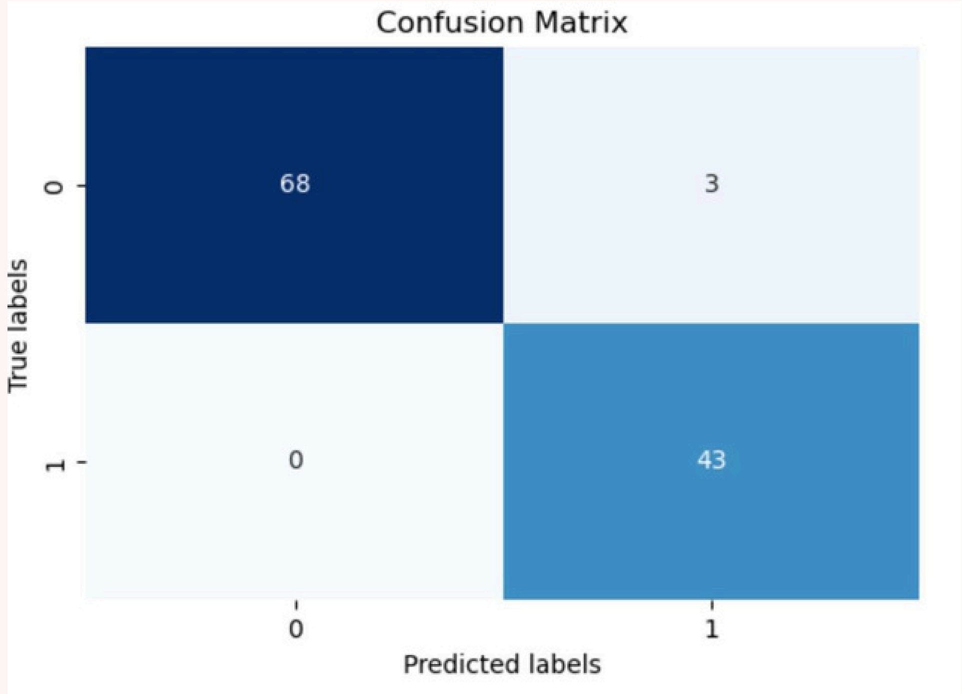
Stochastic Descent

```
Stochastic Gradient Descent with MCP Penalty:
Accuracy: 0.9824561403508771
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.97      0.99        71
           1       0.96      1.00      0.98        43

    accuracy                           0.98       114
   macro avg       0.98      0.99      0.98       114
weighted avg       0.98      0.98      0.98       114
```



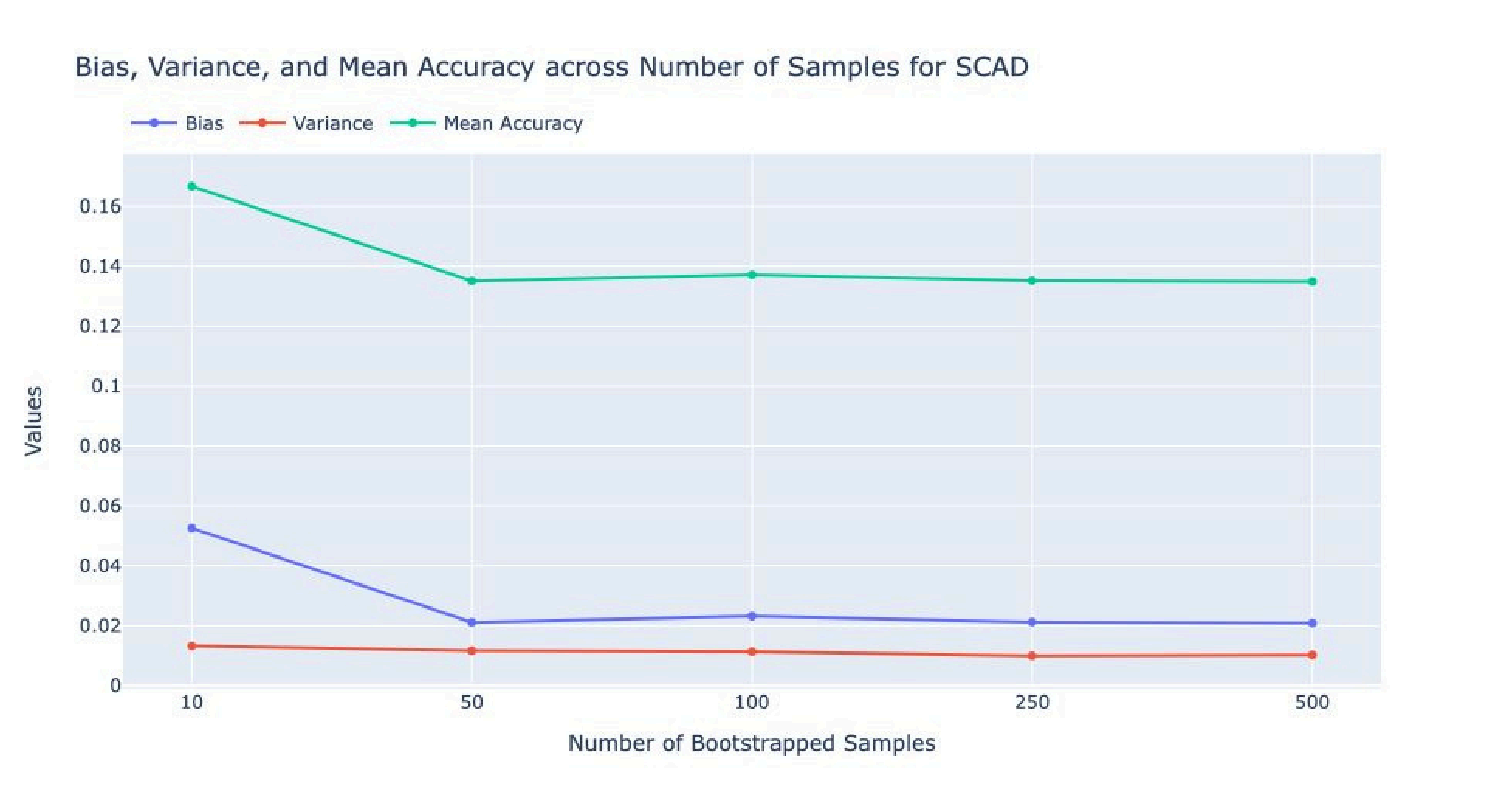
Proximal Descent

```
Proximal Gradient Descent with MCP Penalty:
Accuracy: 0.9736842105263158
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.96      0.98        71
           1       0.93      1.00      0.97        43

    accuracy                           0.97       114
   macro avg       0.97      0.98      0.97       114
weighted avg       0.98      0.97      0.97       114
```

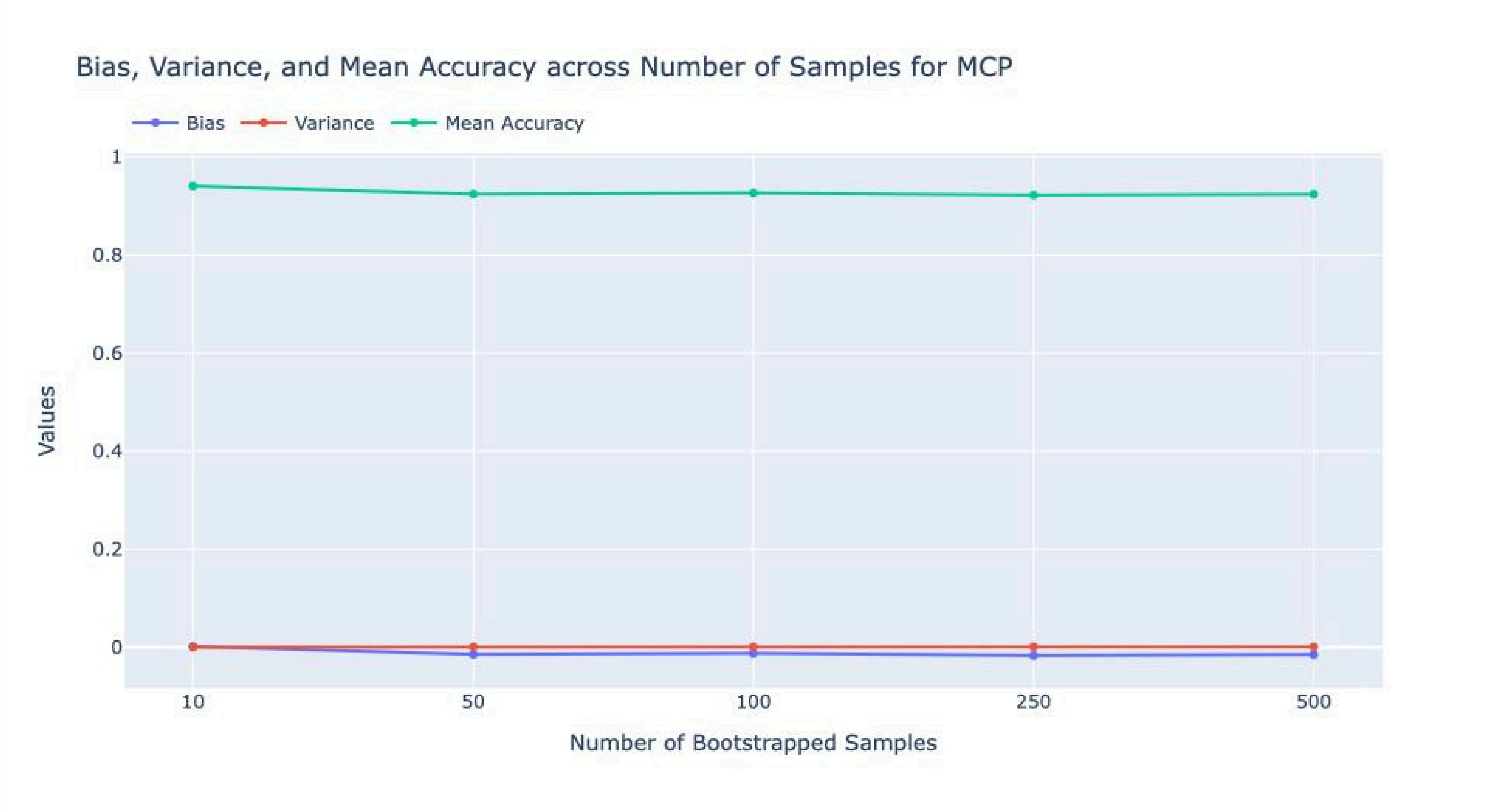


Bootstrap

SCAD



MCP



Conclusion

- In conclusion, our project has explored the intricacies of breast cancer diagnosis using nonconvex penalties and optimization algorithms. We've scrutinized penalties like SCAD and MCP, alongside optimization methods including Coordinate Descent, Stochastic Descent, and Proximal Descent.

- Through meticulous evaluation, Coordinate Descent emerged as the most effective algorithm for our dataset, offering superior accuracy and computational efficiency. This highlights the importance of selecting appropriate methods for optimal model performance, advancing our understanding of breast cancer diagnosis and paving the way for improved patient care.

Thank You