**Reference**

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. Chapter 8 covers optimization algorithms, including stochastic gradient descent and mini-batch SGD.

Accessed: October 1, 2023 [Online]

Available: <https://www.deeplearningbook.org/contents/optimization.html>

[2] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13, 281-305. This paper discusses the use of random search for hyperparameter optimization, which can be useful when tuning the parameters of mini-batch SGD.

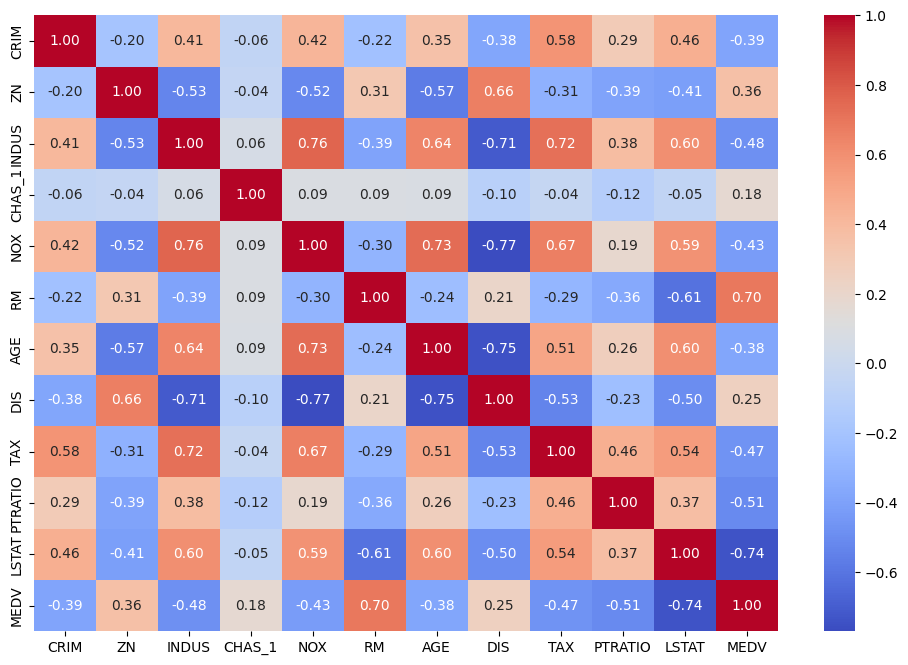
Accessed: October 1, 2023 [Online]

Available: <https://jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf>

**Appendix**

1. **Linear Regression Model**
2. **Feature Correlation Matrix**

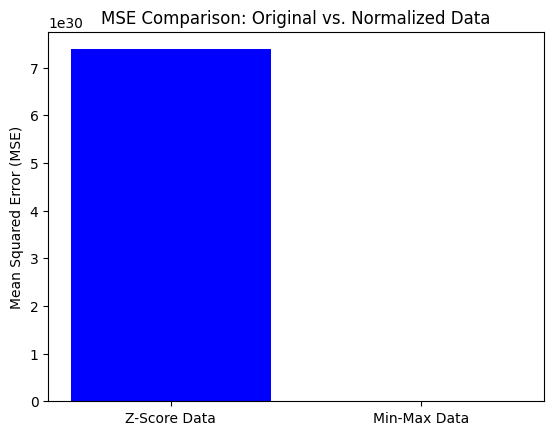
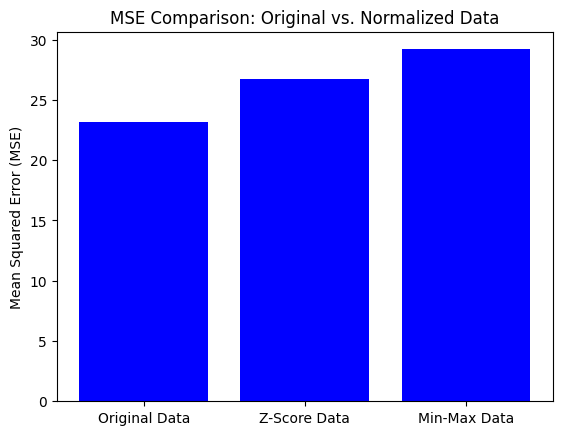
The matrix shows the correlation coefficients between all pairs of features. We can know the importance of each feature on the target output (MEDV)



(Figure A1. The correlation matrix of data set 1 feature and output)

1. **The Influence of Different Normalization**

The graphs show the MSE of different Normalization in the Analytical and Mini-Batch SGD Linear Models, with the original data and the min-max normalized data performing well respectively.



(Figure A2. The Influence of Normalization among analytical and GD Model)

1. **5-Fold Cross Validation of Test and Train Set Complete Table**

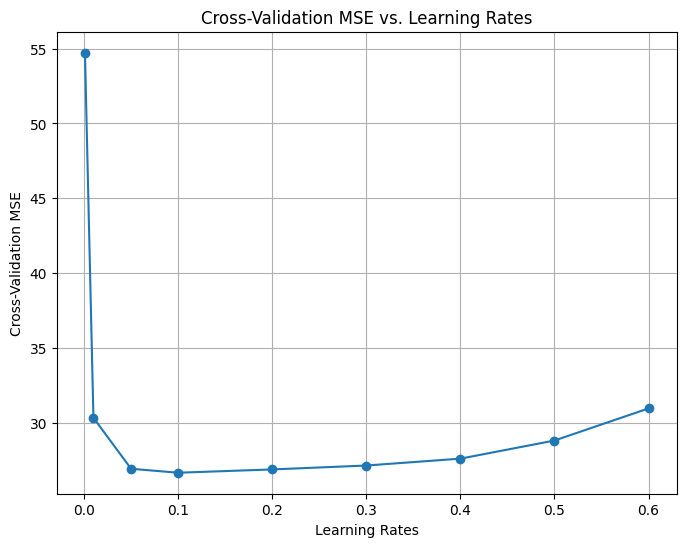
The tables are showing the performance metrics on both the training set and test set for each model.

| Model | Train MSE | Test MSE | Train MAE | Test MAE | Train R^2 | Test  R^2 |
| --- | --- | --- | --- | --- | --- | --- |
| Analytical | 21.5 | 23.3 | 3.26 | 3.40 | 0.75 | 0.72 |
| GD | 22.3 | 23.9 | 3.25 | 3.38 | 0.74 | 0.71 |
| Mini-Batch | 21.6 | 23.4 | 3.24 | 3.39 | 0.74 | 0.72 |

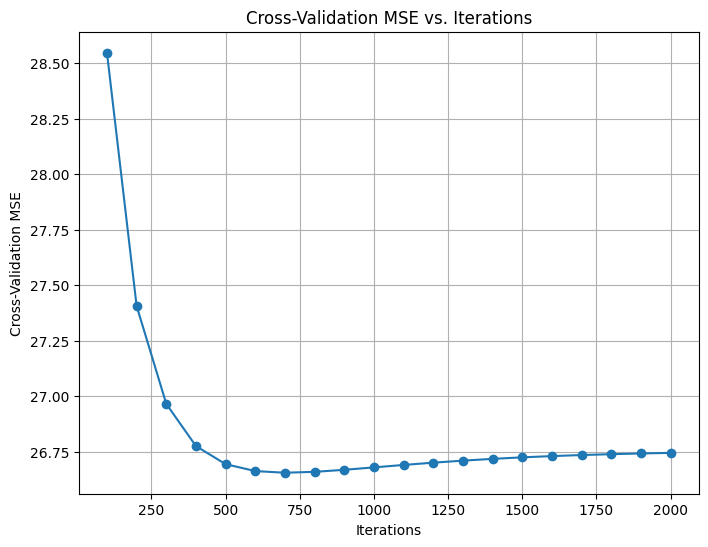
(Figure A3. The 5-Fold Cross Validation Performance Metrics on Train and Test)

1. **5-Fold Cross Validation on Mini-Batch SGD**

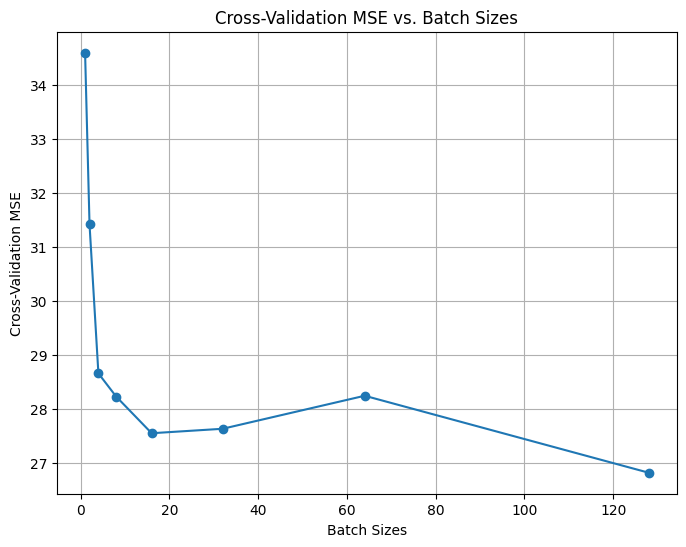
The graphs show the 5-fold validation MSE performance of the Mini-Batch SGD Linear Model, changing by different learning rates, iterations, and batch sizes.



(Figure A4.1 The Mini-Batch CV MSE though increasing learning rate)



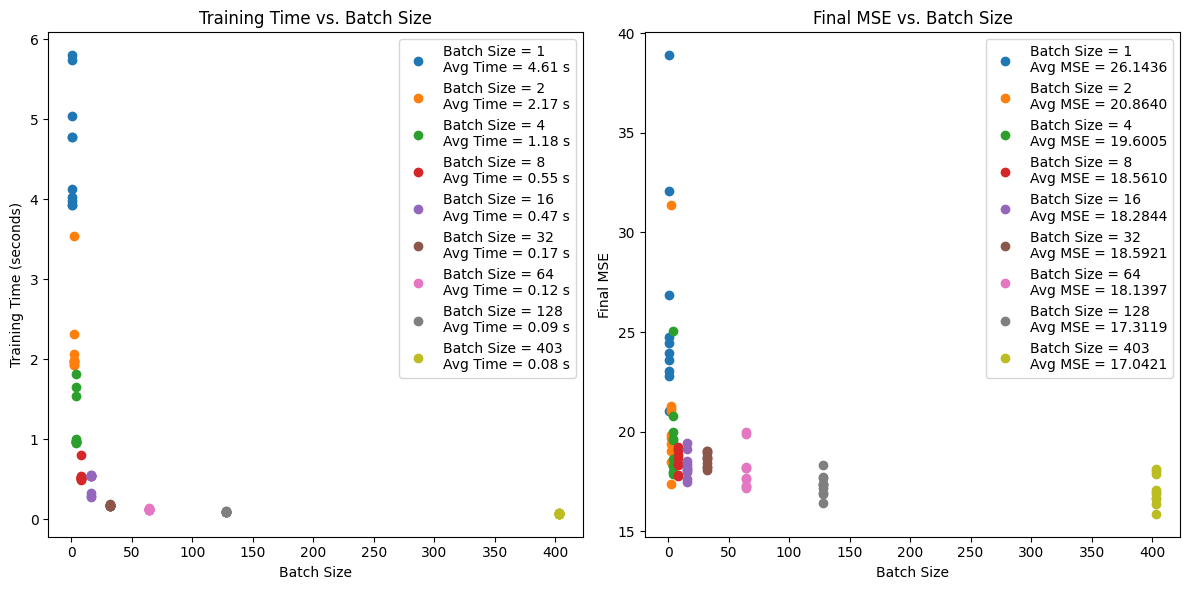
(Figure A4.2 The Mini-Batch CV MSE though increasing Iterations)



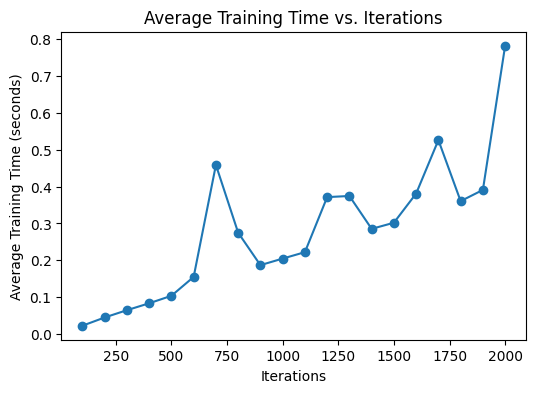
(Figure A4.3 The Mini-Batch CV MSE though increasing batch size)

1. **Additional Analysis of Parameter Change in Mini-Batch**

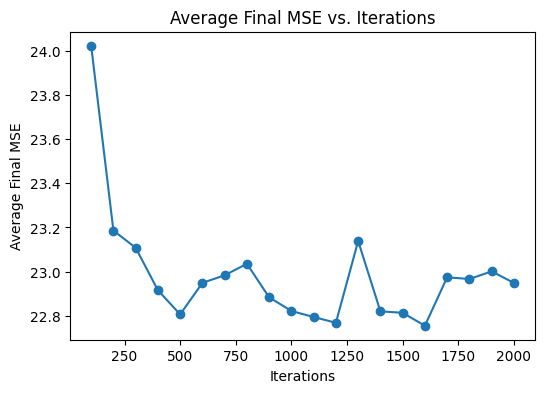
The additional analysis of how different batch sizes and iteration sizes influence the training time and average MSE



(Figure A5.1 The graph analyzes the training time and average MSE performance with different batch sizes)



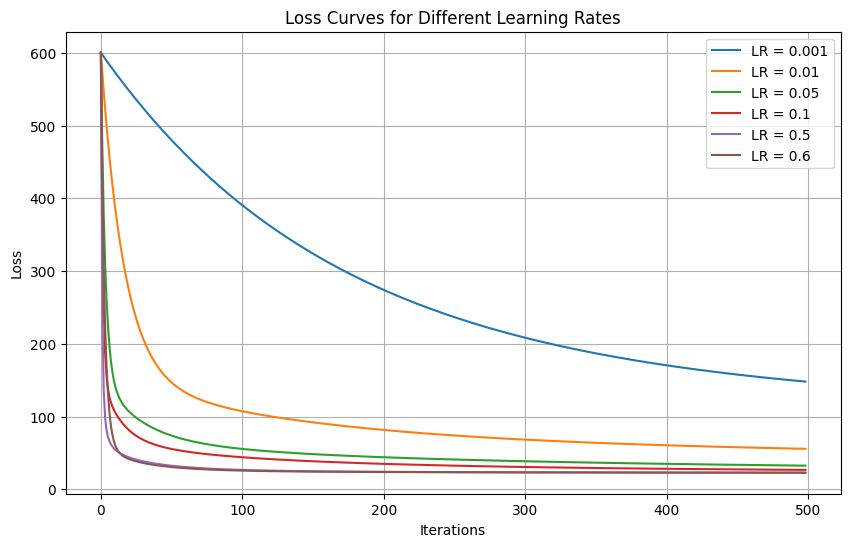
(Figure A5.2 The graph analyzes the training time with different iterations.)



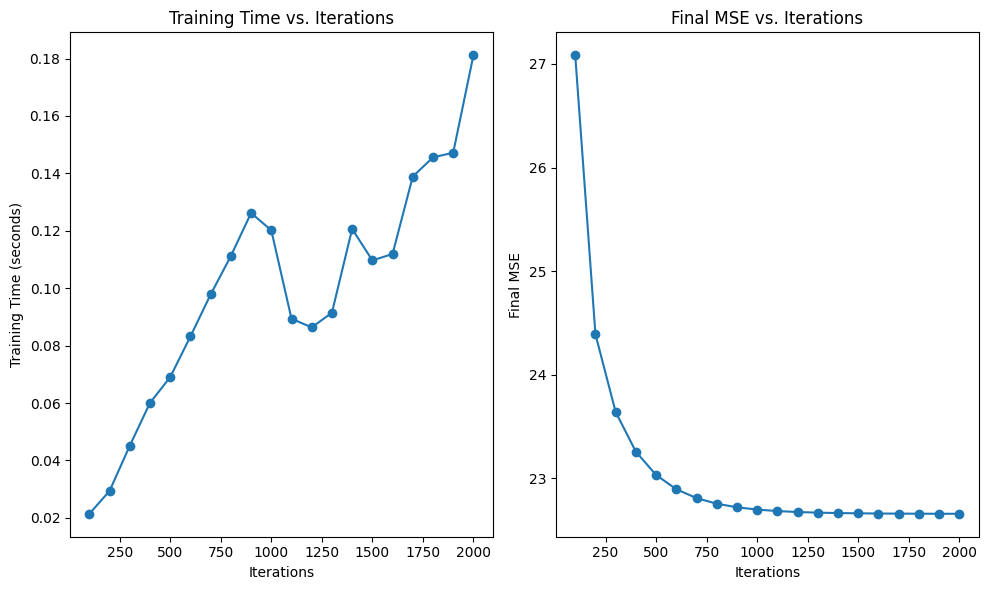
(Figure A5.3 The graph analyzes the average MSE performance with different iterations.)

1. **Additional Analysis of Parameter Change in Gradient Descent**

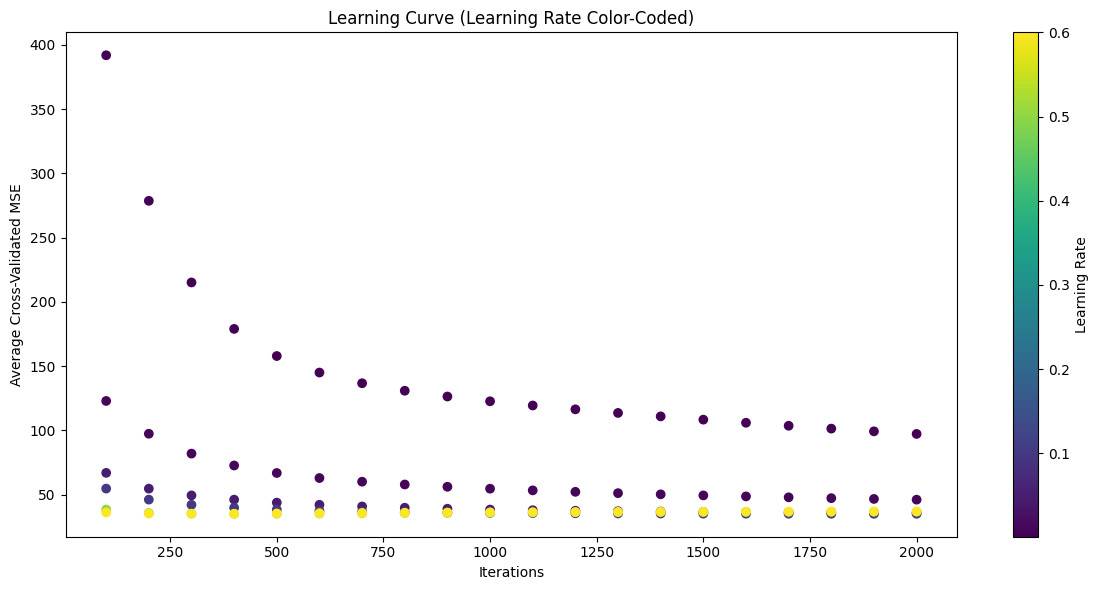
Showing how the different learning rates and iterations affect the performance of the gradient descent model. Finally, hyperparameter tuning finds out the best parameter configurations for the GD linear model.



(Figure A6.1 The graph shows the Loss Curve in GD Linear Model with different Learning Rates)



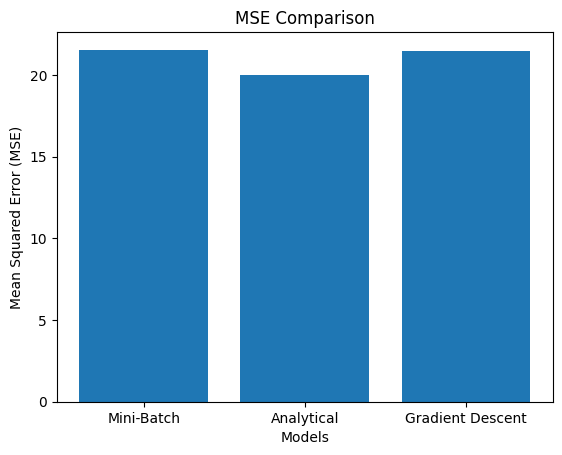
(Figure A6.2 The graph shows the training time and Average MSE in GD Linear Model with different iteration sizes.)



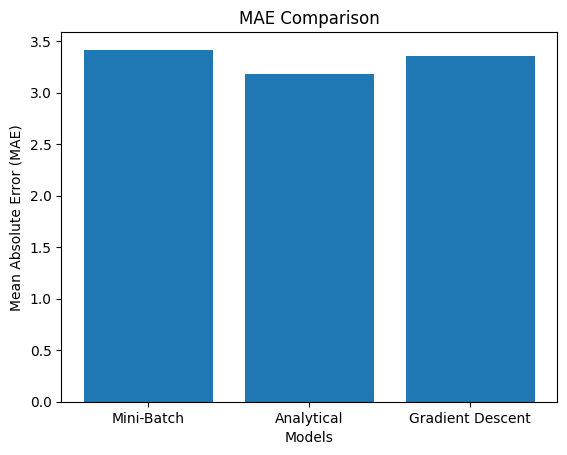
(Figure A6.3 The graph of hyperparameters tuning on the Gradient Descent Linear Model to find out the optimal parameters.)

1. **The Graphs of Model Comparison**

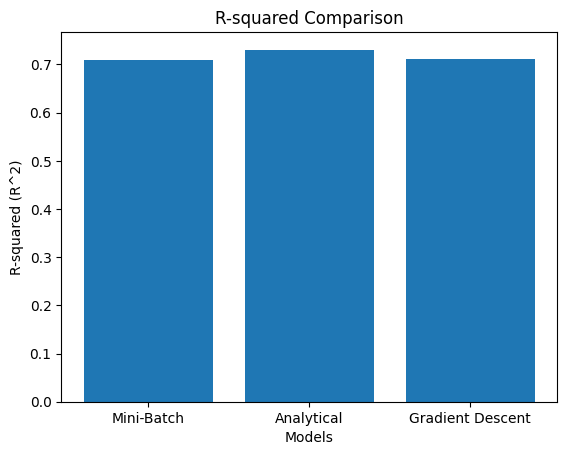
The comparisons of model performance (Analytical, GD, Mini-Batch SGD) in different aspects (MSE, MAE, R squares, Training Time)



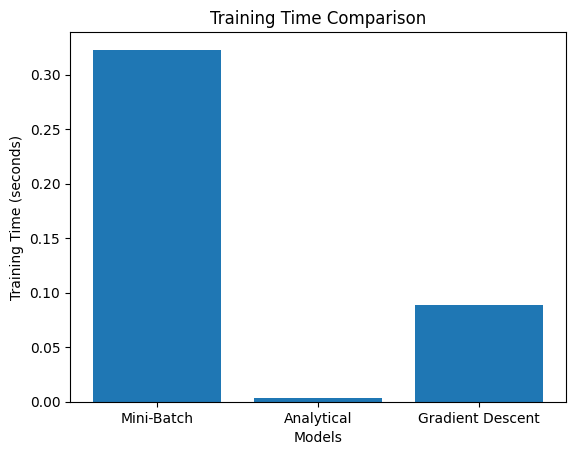
(Figure A7.1 The comparison of Models Performance (Analytical, GD, Mini-Batch SGD) in MSE)



(Figure A7.2 The comparison of Models Performance (Analytical, GD, Mini-Batch SGD) in MAE)



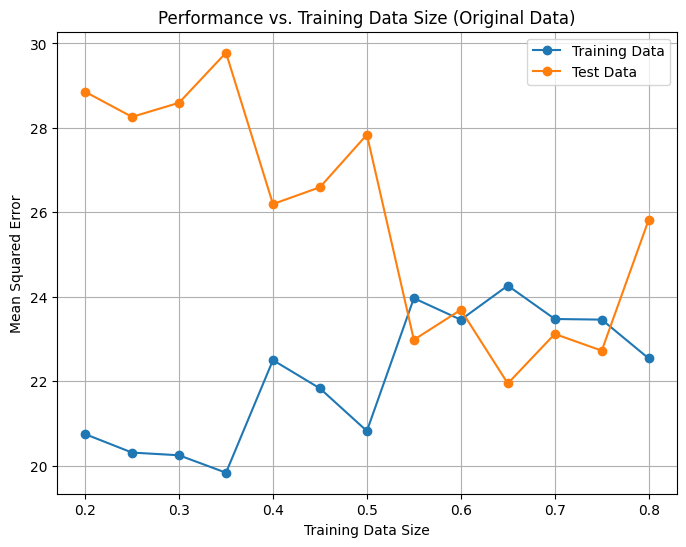
(Figure A7.3 The comparison of Models Performance (Analytical, GD, Mini-Batch SGD) in R squares)



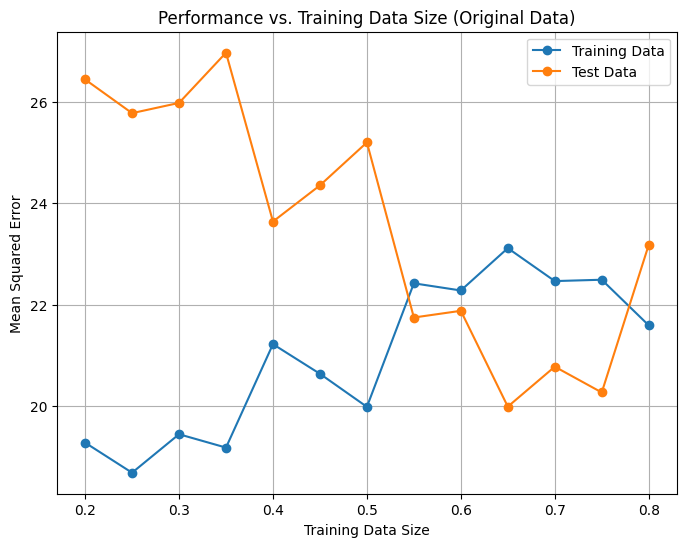
(Figure A7.4 The comparison of Models Performance (Analytical, GD, Mini-Batch SGD) in Training Time)

1. **The L2 Regularization Influence**

With various training sizes, we investigate the difference of the Average MSE of the analytical model with and without L2 Regularization). The graphs with and without L2 are similar shape and both obtain great MSE performance when the train size is 0.65. However, it shows L2 Regularization does not carry significant improvement to lower than the Average MSE, comparing to the situation without any regularization.

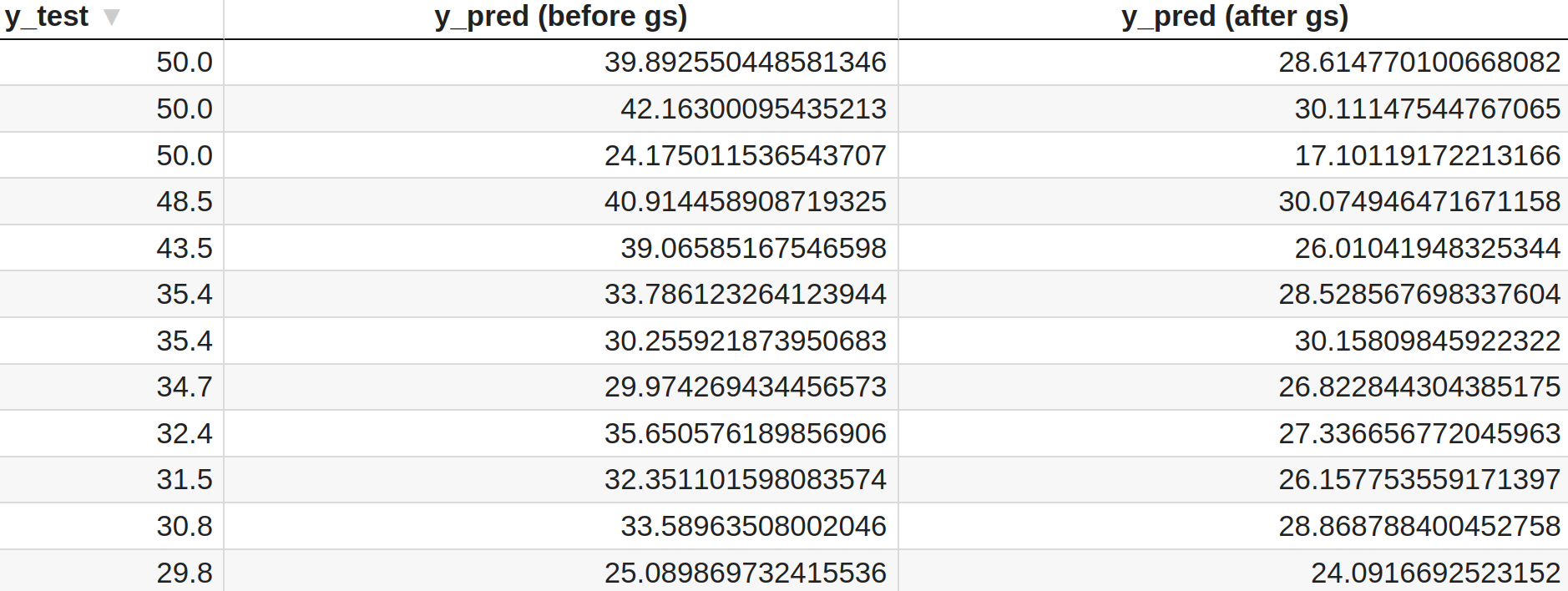
’

(Figure A8.1 The MSE Performance with L2 Regularization in Analytical Model)

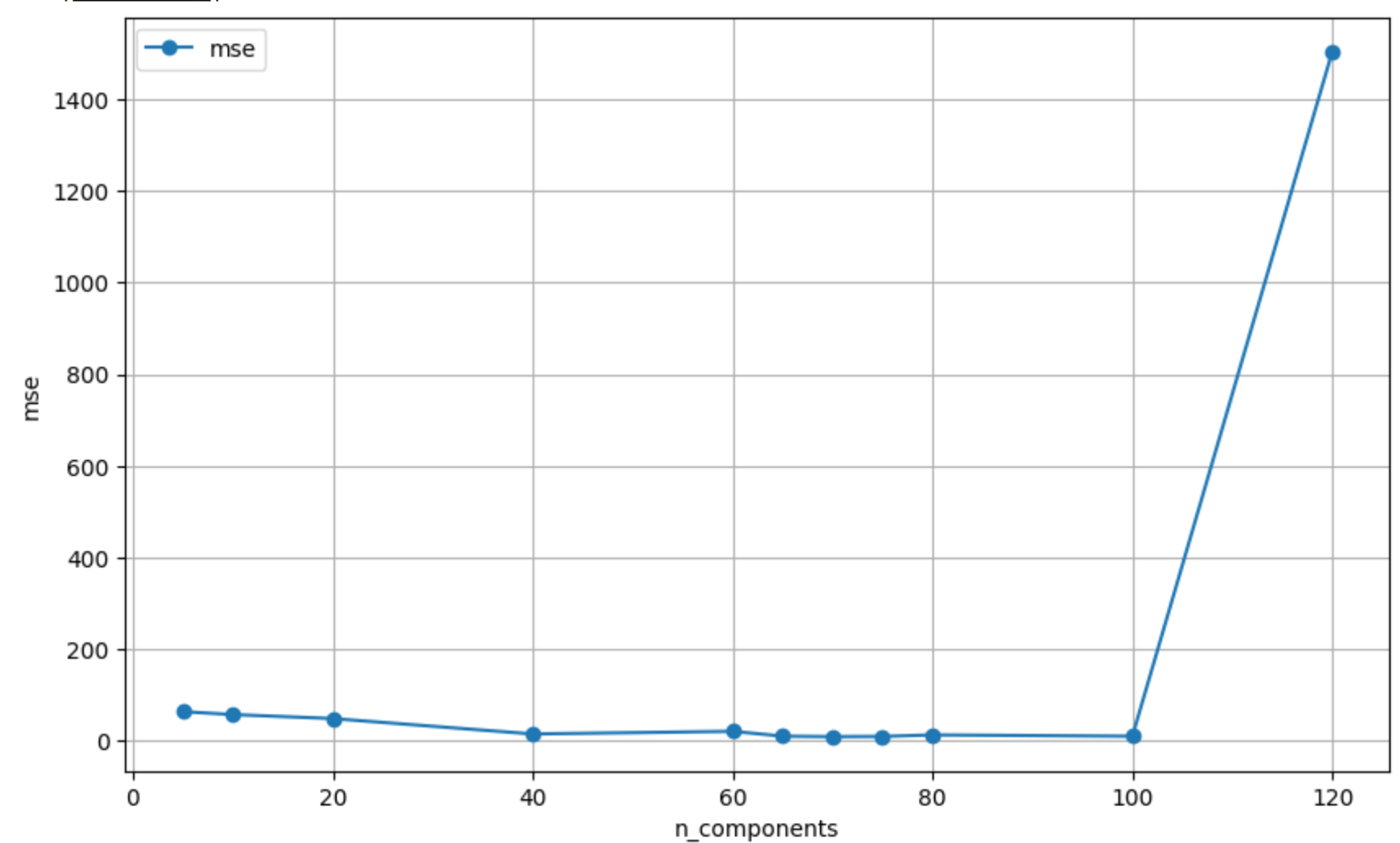


(Figure A8.2 The MSE Performance with no L2 Regularization in Analytical Model)

1. **Predicted value comparison in Feature Extraction using Gaussian basis function**

****

(Figure A9.1 The comparison of the predicted value before and after the feature extraction with target value)



(Figure A9.2 Grid search in training data to find the optimal numbers of Gaussian basis Function)





(Figure A9.3 The comparison of the MSE in test data before and after applying 65 Gaussian basis functions )

B. Logistic Regression Model

B1. Baseline performance for logistic regression

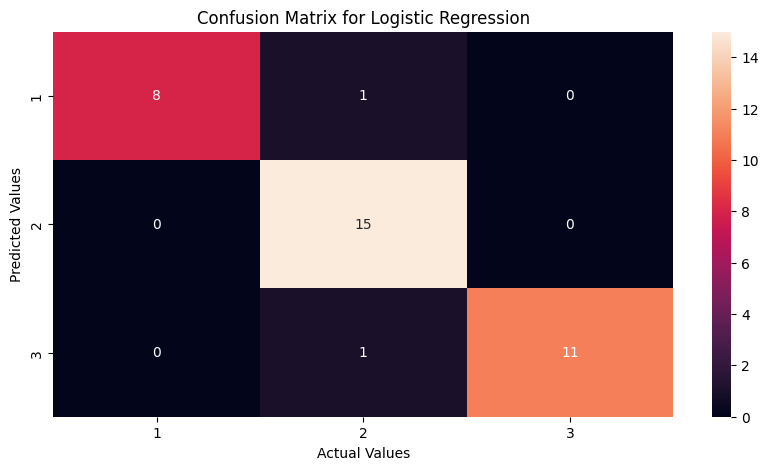


Figure B1.1: Cross validation matrix and its associated computed metrics for batch gradient descent logistic regression. Metrics are as presented for avg\_accuracy: 0.962963, avg\_precision: 0.935185, avg\_recall: 0.960784, avg\_F1: 1.890076

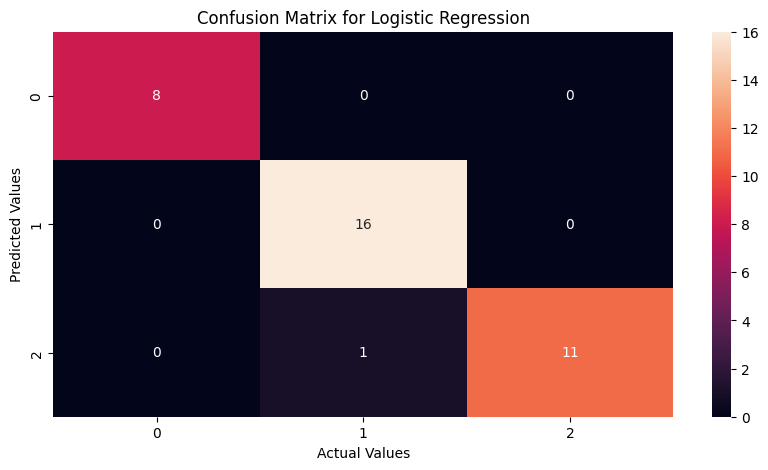


Figure B1.2: Cross validation matrix and its associated computed metrics for mini-batch gradient descent logistic regression. Metrics are as presented for avg\_accuracy: 0.981481, avg\_precision: 0.972222, avg\_recall: 0.980392, avg\_F1: 1.966340

B3. Varying Training Size

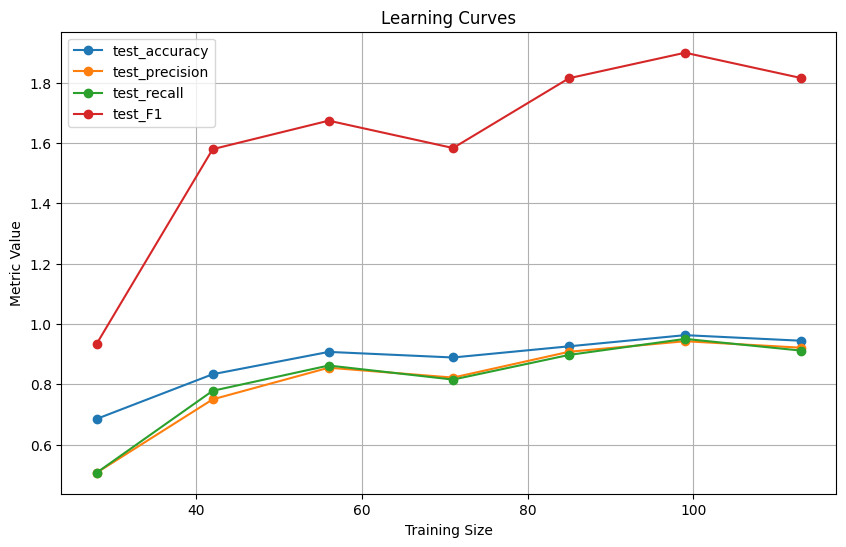
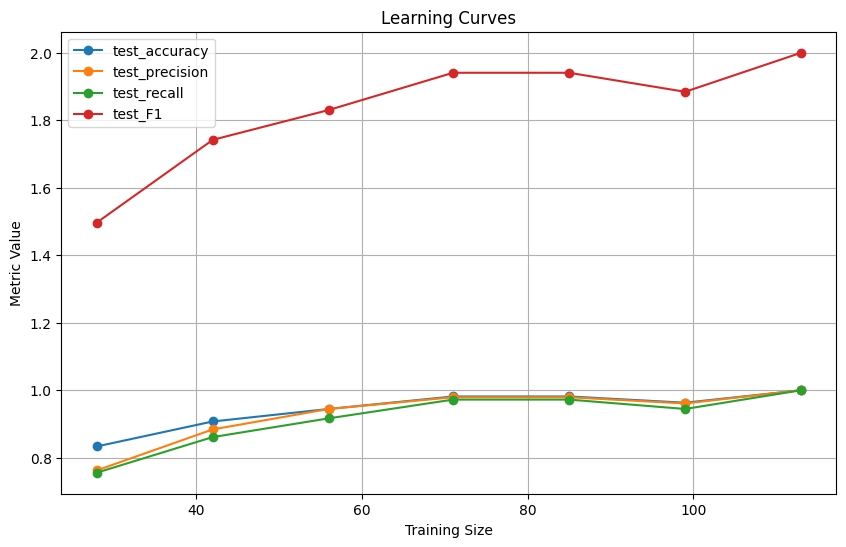


Figure B3.1: Graph of test metrics against differing training sizes for GD (left) and mini-batch GD (right).

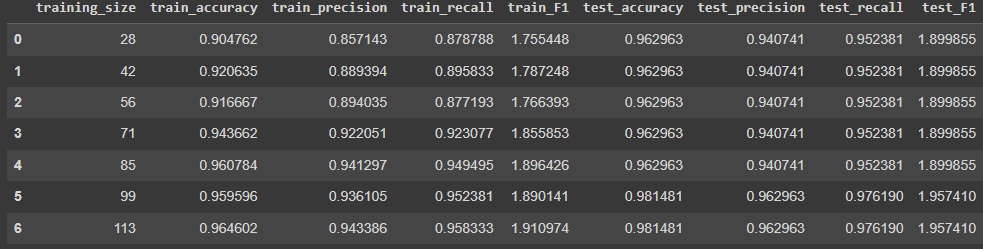


Figure B3.2: Table of metrics for the GD plot above.

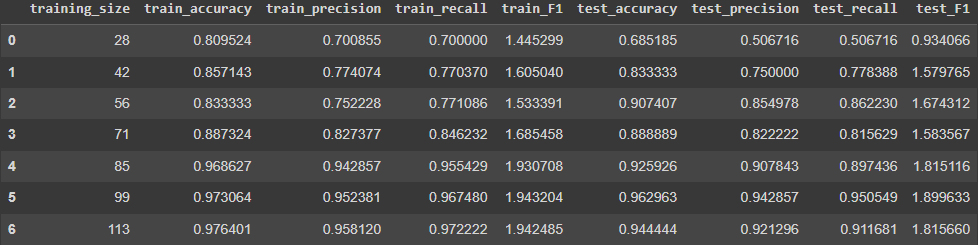


Figure B3.3: Table of metrics for the Mini-batch GD plot above.

B4. Varying Mini-Batch Sizes

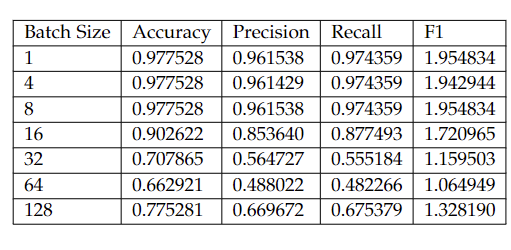


Table B4.1: Performance metrics against mini-batch size for MB gradient descent. We observe a general downward trend.

B5. Varying Learning Rates

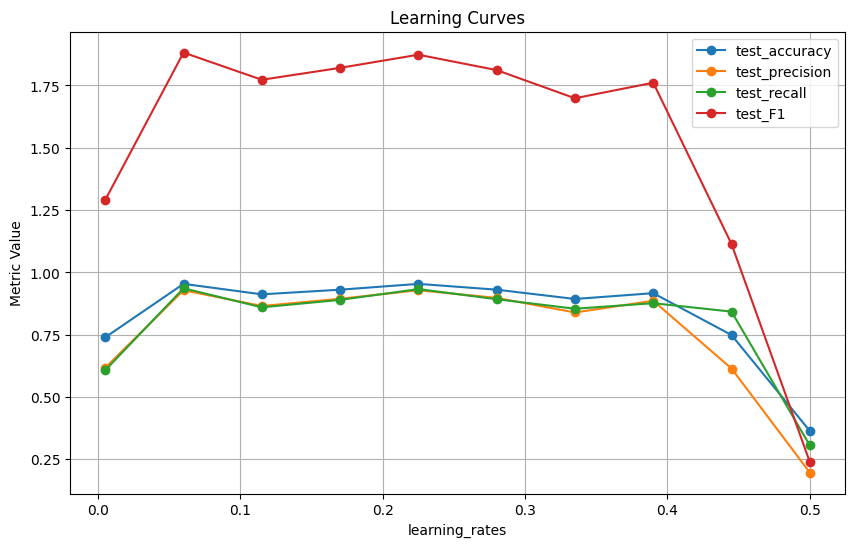
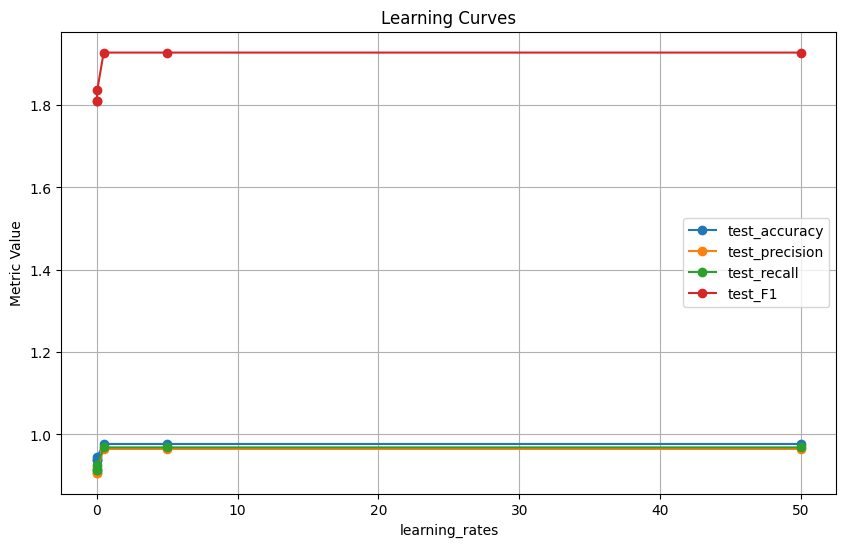
****

Figure B5.1: Plots of learning rates vs. performance metrics. (Mini-batch size = 8) GD on the left and Mini-batch on the right.

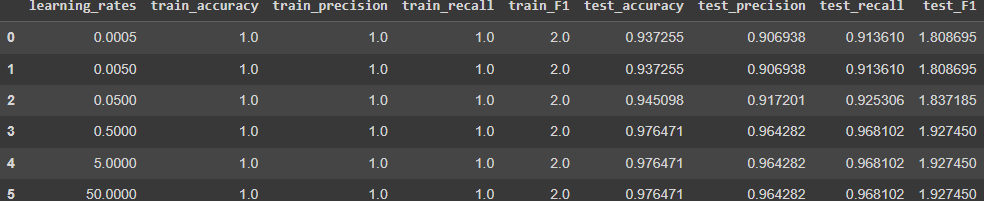


Figure B5.2: Table of metrics from the above GD plot.

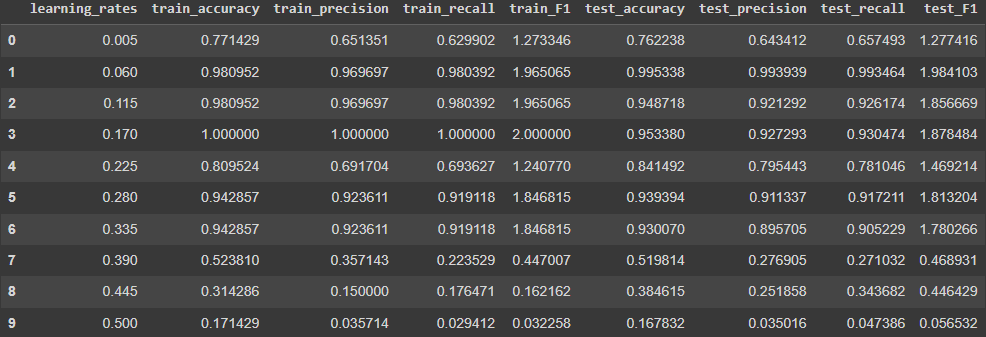


Figure B5.3: Table of metrics from the above Mini-batch plot. A range of [0.06, 0.17] seems most adequate for a batch-size of 8.

B6. Finding optimal parameters

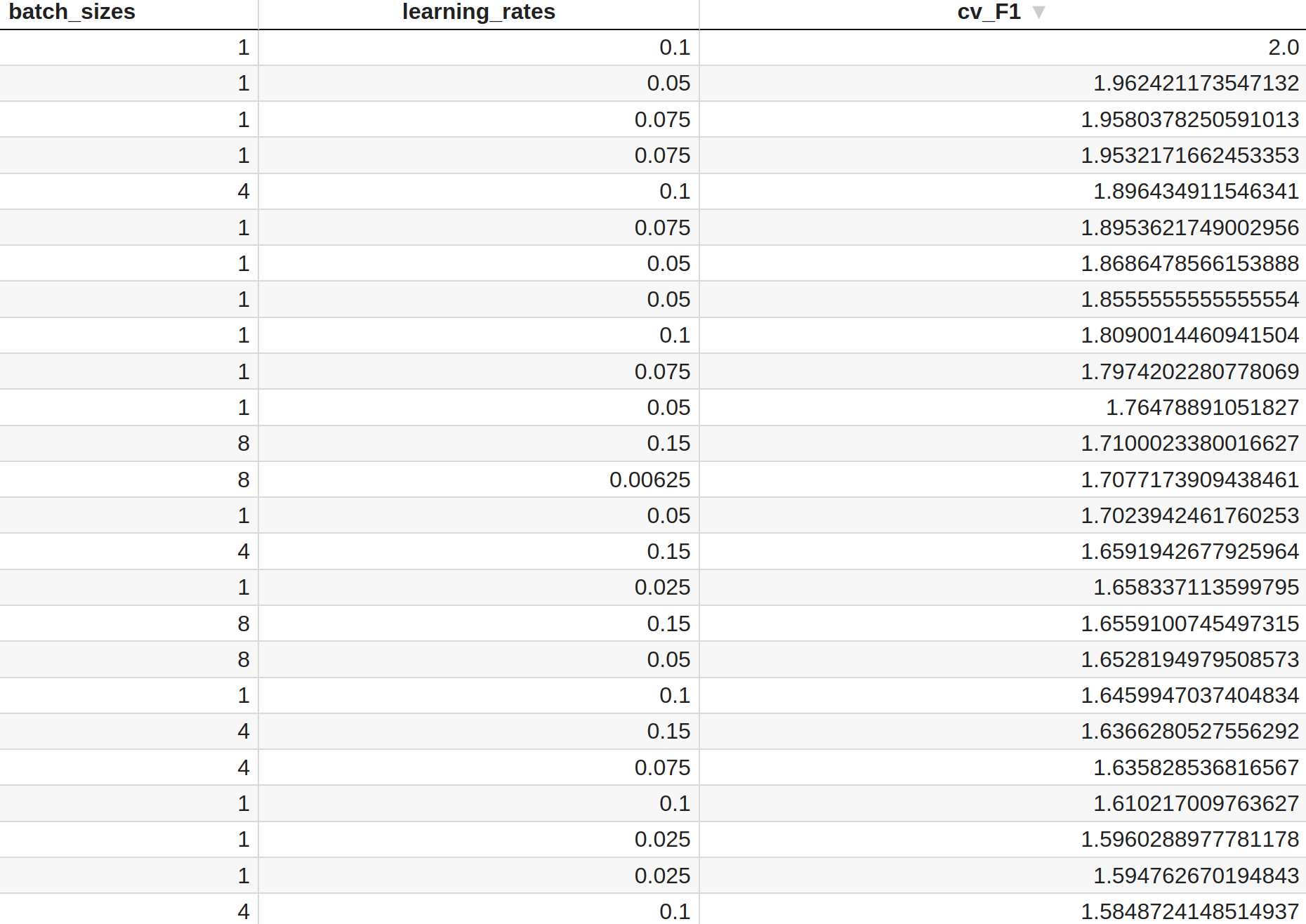


Figure B6.1: F1 benchmark table with different batch\_sizes and learning\_rates

B.9 Additional Tweaks

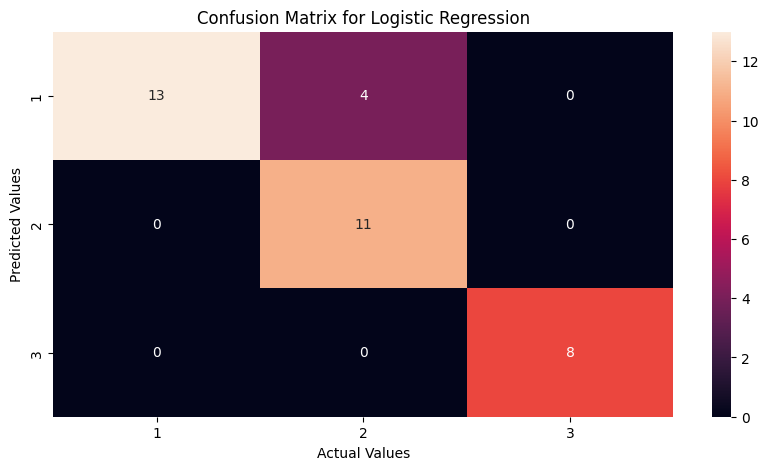


Figure B9.1 Resulting correlation matrix from L1-regularized and ‘ASH’-less dataset. With metrics avg\_accuracy: 0.925926, avg\_precision: 0.921569, avg\_recall: 0.911111, avg\_F1: 1.788159.