MSID – Report (Part II)

Author: Kamil Borusiak – 280447

1. Data Split

The dataset was divided into three parts:

- Training set: 70%
- Test set: 20%
- Validation set: 10%

2. Evaluation Metrics

The following metrics were used to evaluate model performance:

- Accuracy: The ratio of correctly predicted observations to the total observations.
- Precision: The ratio of correctly predicted positive observations to the total predicted positives.
- Recall: The ratio of correctly predicted positives to all actual positives.
- F1-score: Harmonic mean of precision and recall.
- Support: The number of actual occurrences of each class in the dataset.
- AUC: Area under the Receiver Operating Characteristic curve, measuring the model's ability to distinguish between classes.
- MSE (Mean Squared Error): The average of squared differences between actual and predicted values
- R^2: Proportion of variance in the dependent variable explained by the model (ranges from 0 to 1).

3. Classification Model Performance

Logistic Regression

Set	Class	Precision	Recall	F1-score	Support
Test	-1.0	0.92	0.86	0.89	217
	1.0	0.91	0.95	0.93	328
Accuracy				0.92	545
Validation	-1.0	0.93	0.83	0.87	214
	1.0	0.90	0.96	0.93	330
Accuracy				0.91	544
Train	-1.0	0.94	0.85	0.89	990
	1.0	0.91	0.97	0.94	1551
Accuracy				0.92	2541

Decision Tree

Set	Class	Precision	Recall	F1-score	Support
Test	-1.0	0.85	0.83	0.84	217
	1.0	0.89	0.90	0.90	328
Accuracy				0.87	545
Validation	-1.0	0.82	0.80	0.81	214
	1.0	0.87	0.89	0.88	330
Accuracy				0.85	544
Train	-1.0	1.00	1.00	1.00	990
	1.0	1.00	1.00	1.00	1551
Accuracy				1.00	2541

SVM

Set	Class	Precision	Recall	F1-score	Support
Test	-1.0	0.94	0.86	0.90	217
	1.0	0.91	0.97	0.94	328
Accuracy				0.92	545
Validation	-1.0	0.95	0.82	0.88	214
	1.0	0.89	0.97	0.93	330
Accuracy				0.91	544
Train	-1.0	0.98	0.86	0.92	990
	1.0	0.92	0.99	0.95	1551
Accuracy				0.94	2541

4. Closed-Form Linear Regression

Set	MSE	R^2
Train	0.2272	0.9617
Test	0.2266	0.9657
Validation	0.2396	0.9641

5. Logistic Regression Summary

Custom Logistic Regression

Set	Accuracy	F1-score	AUC
Train	0.921	0.937	0.961
Test	0.919	0.935	0.961
Validation	0.912	0.930	0.960

Scikit-learn Logistic Regression

Set	Accuracy	F1-score	AUC
Train	0.921	0.937	0.964
Test	0.916	0.931	0.962
Validation	0.906	0.925	0.960

6. CPU vs GPU Training Comparison

Training time is longer on GPU due to the small model size and dataset. GPU has a higher overhead from transferring data between CPU and GPU memory, and initializing CUDA kernels, which outweighs the parallel processing benefits for this simple task.

Device	Training Time	Set	Accuracy	F1-score	AUC
CPU	4.81 s	Train	0.916	0.934	0.959
		Test	0.930	0.943	0.960
		Validation	0.914	0.931	0.958
GPU	6.97 s	Train	0.917	0.934	0.959
		Test	0.930	0.943	0.960
		Validation	0.914	0.931	0.958