WARSAW UNIVERSITY OF TECHNOLOGY
FACULTY OF MATHEMATICS AND INFORMATION SCIENCE

Project I Report

Advanced Machine Learning

Students:
Karina Tiurina (335943)
Nikita Kozlov (317099)
Róża Klimek(329533)

Supervisor:
mgr Anna Kozak

Warsaw 2024

Contents

1 Methodology

Task 1

Data Acquisition.

In Task 1, we focus on analyzing and modeling datasets with distinct characteristics and sizes. Our dataset selection includes both small and large datasets, catering to different research questions and computational capabilities.

We have chosen the following datasets as a part of our project.

Small datasets with at most 10 variables:

1. Rice Dataset (Cammeo and Osmancik): This dataset comprises 7 features with a binary target variable, distinguishing between two rice varieties, Cammeo and Osmancik, across 3810 instances.
2. Statlog (Shuttle): Featuring 7 features and 7 classes among 58000 instances. Predominantly, 80% of the data pertains to class 1, with the remaining classes recoded to 0.
3. Banknote Authentication: Contains 4 features and 1400 instances. It categorizes banknotes into 2 classes: authentic (1) and fake (0).

Large datasets with more than 10 variables:

1. Online Shoppers Purchasing Intention Dataset: Encompasses 17 features across 12330 instances, aiming to predict the purchasing intentions of online shoppers.
2. EEG Eye State: With 14 features and 15000 instances, this dataset distinguishes between open and closed eye states.
3. Web Page Phishing Dataset: Features 19 attributes over 100000 instances, classifying web pages as phishing (1) or legitimate (0).
4. Airline Satisfaction Dataset: Comprises 24 features with 130000 instances, classifying airline passenger satisfaction into satisfied (1) or neutral/dissatisfied (0).
5. Dataset for Link Phishing Detection: This dataset includes 84 features and 19400 instances, aimed at phishing link detection
6. Optdigits: Contains 65 features over 6500 instances, categorizing digits as valid (1) or invalid (0)

Data Preprocessing.

Our data preprocessing strategy is designed to ensure that the datasets are in the best possible form for analysis and modeling. This strategy involves a series of steps to increase the quality and integrity of the data and fix common issues like missing values, multicollinearity, and variable encoding. The process consists of the following steps:

1. Data Cleaning

The initial step involves data cleaning. This includes the removal of any instances with missing or incomplete information. By doing so, we clean the dataset up so that no instances contain empty values.

2. Applying standard scaler

During our earlier experiments we have noticed that the performance of our models on denormalized datasets is marginally worse than that of when the features are normalized. Therefore, we have applied normalization for every feature of our dataset.

3. Addressing Multicollinearity

Multicollinearity occurs when two or more predictor variables in a dataset are highly correlated. This condition can distort the results of statistical analyses by affecting the stability and interpretation of coefficient estimates. To mitigate this, we assess the degree of correlation among predictors and systematically eliminate those that exhibit a high degree of multicollinearity. The goal is to retain a set of predictors that provide independent and valuable information about the target variable. Our chosen collinearity threshold is 0.8.

4. Variance Inflation Factor (VIF) Analysis

We use Variance Inflation Factor (VIF) analysis to identify and remove variables that contribute to multicollinearity. VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity in the model. Variables with a high VIF are considered problematic, as they may skew the model outcomes. By setting a VIF threshold of 10.0, we systematically exclude features that exceed this

limit, ensuring that the remaining variables contribute meaningfully and independently to the prediction models.

5. Encoding and Final Dataset Structuring

For datasets with categorical variables, appropriate encoding techniques are applied to transform these variables into a format that can be efficiently processed. This step is crucial for handling non-numeric data, such as the classification of rice varieties or the categorization of web pages as phishing or legitimate. The final dataset is then structured into a matrix of predictor variables and a vector of the target variable.

Task 3.1

Suggested stopping rule is the following:

1. If number of iterations is higher than the limit value (by default 500);
2. If mean loss difference across last 10 iterations is lower than the specified value (by default 0.001)

Based on conducted experiments, IWLS is able to converge in just 11 iterations. SGD and ADAM are dependent on the datasets; however, SGD is able to converge faster. ADAM implementation would converge faster with the stopping value 0.01, but for the fair comparison, it takes the same value as the other models. Results of the experiments are available in 'Task3.3_ConvergenceAnalysis.ipynb'

Task 3.2

Balanced accuracy is a performance measurement metric used to evaluate classification models. It considers the distribution of classes in the dataset by calculating the average accuracy across all classes, ensuring that each class contributes equally to the overall score. The results of our implementation performance testing are highly dependent on the dataset. In general, IWLS was able to provide more stable (although not the highest) results on all datasets. Most difficult dataset for all models is '5. EEG Eye State'. 'Task3.2-BalancedAccuracy.ipynb' contains full list of boxplots for each dataset separately.

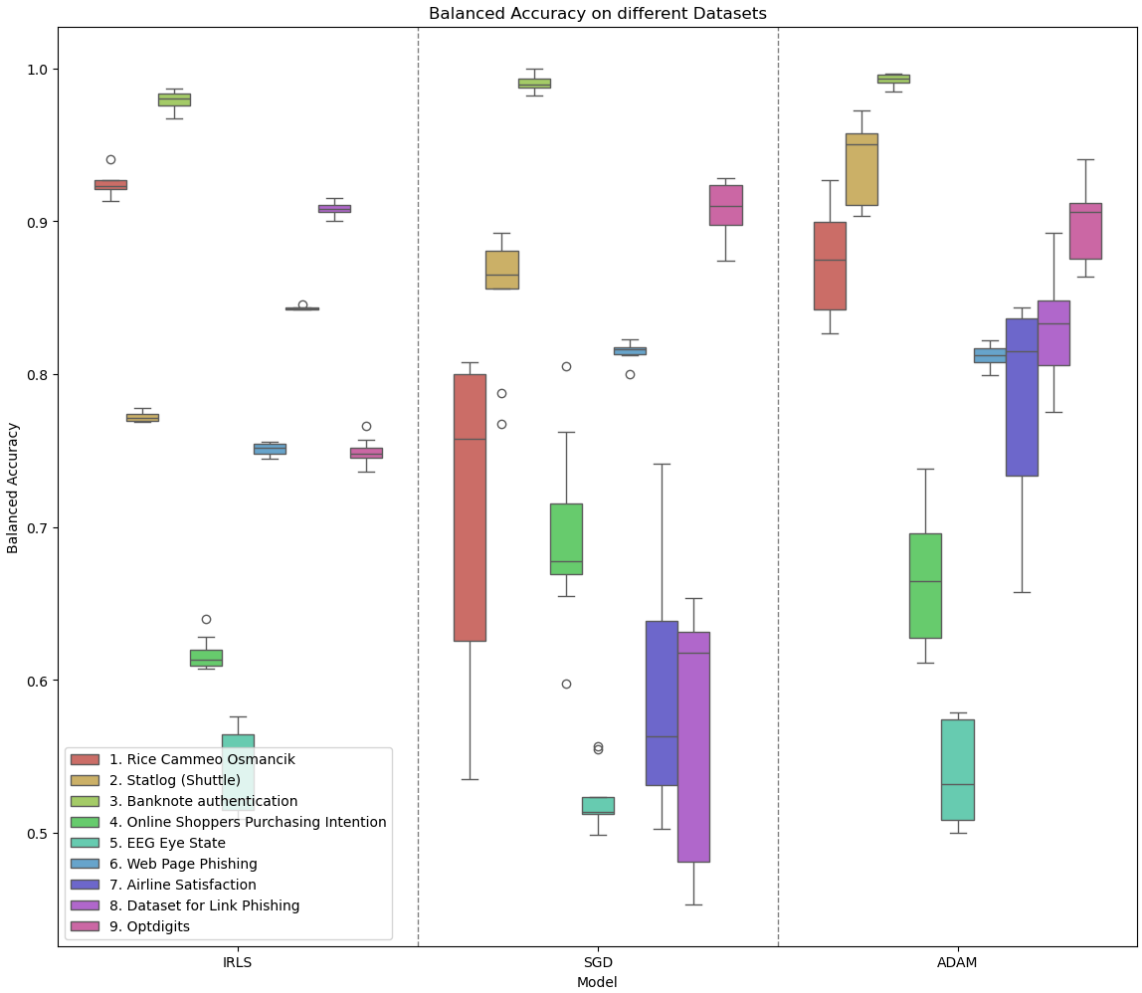

Figure 1.1 Balanced accuracy of IRLS, SGD and ADAM implementations

Source	IRLS	SGD	ADAM
1. Rice Cammeo Osmancik	92.41	70.48	87.59
2. Statlog (Shuttle)	77.19	85.35	94.08
3. Banknote authentication	97.93	99.07	99.22
4. Online Shoppers Purchasing Intention	61.73	69.65	66.78
5. EEG Eye State	53.84	52.10	53.84
6. Web Page Phishing	75.11	81.42	81.19
7. Airline Satisfaction	84.32	59.26	78.02
8. Dataset for Link Phishing	90.82	56.15	83.16
9. Optdigits	74.90	90.83	89.94

Figure 1.2 Mean balanced accuracies in percentage

2 Convergence analysis

Task 3.3

Convergence analysis allows to additionally assess the performance and efficiency of the implemented classification models. Additionally, it provides insights into the training process by visualizing the log likelihood function across iteration. Figures 2.1, 2.2 and 2.3 contain results of the analysis of IWLS, SGD and ADAM models from 'Task3.3_ConvergenceAnalysis.ipynb'.

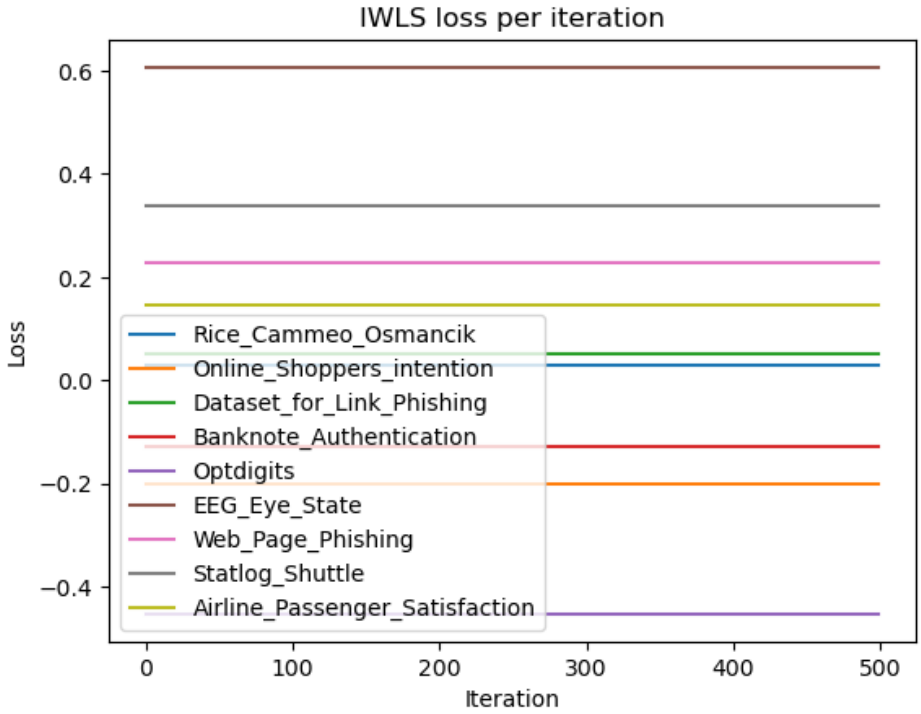

Figure 2.1 IWLS loss per iteration

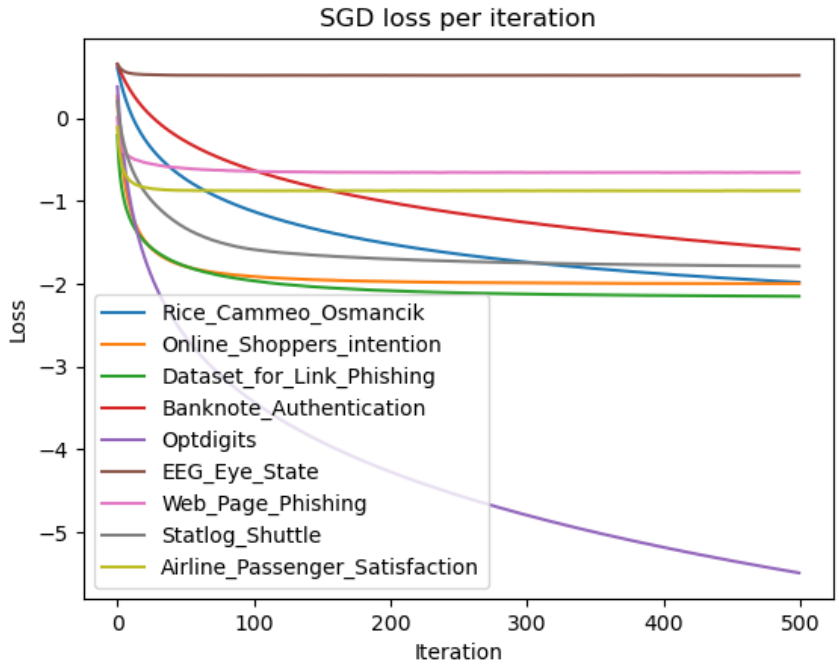

Figure 2.2 SGD loss per iteration

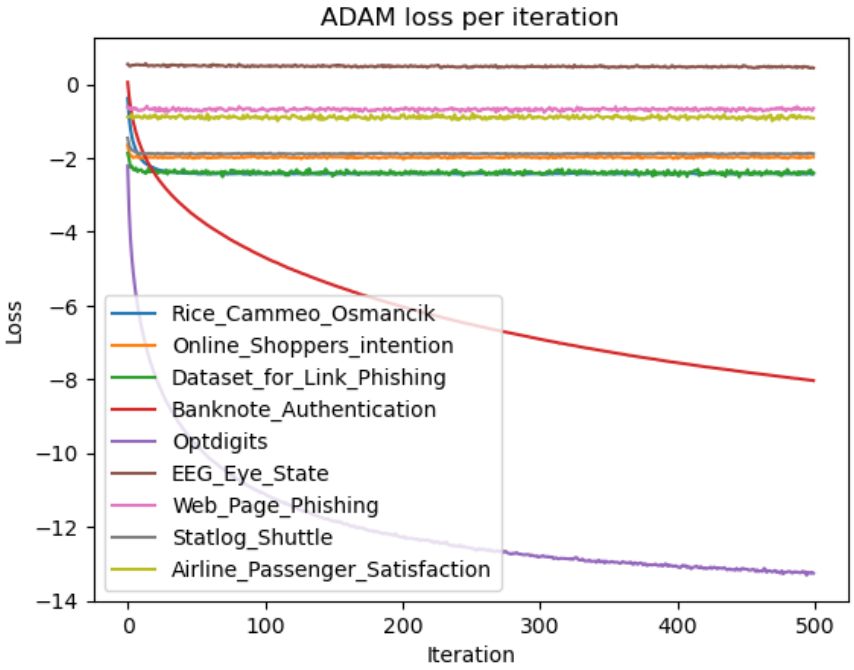

Figure 2.3 ADAM loss per iteration

3 Comparison of classification performance

Task 3.4

Task 3.4 aims to assess the performance of various classification algorithms across different datasets. This chapter focuses on comparing logistic regression (implemented via Iterative Weighted Least Squares (IWLS), Stochastic Gradient Descent (SGD), and Adaptive Moment Estimation (ADAM)) with four popular classification methods: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Trees, and Random Forests. To ensure a fair comparison, we use standardized implementations provided by the scikit-learn library.

The evaluation of each classification method is carried out over multiple iterations to account for variability in train-test splits and to average out random effects. We split the datasets into a 70-30 ratio for training and testing, respectively. The performance metrics include balanced accuracy, F1 score, precision, and recall, which offer a comprehensive view of each classifier's effectiveness.

For logistic regression methods (IWLS, SGD, ADAM), we pay particular attention to their behavior on unbalanced datasets, as the balanced accuracy score will be more indicative of performance compared to simple accuracy.

The results, summarized in several images, show a varied performance of the methods across different datasets.

We present the most notable results in the report. All other boxplots and a table are provided in the jupyter notebook for tasks 3.4 and 3.5.

Generally speaking, the performance of IWLS, SGD and ADAM vary throughout different datasets. We suspect that class imbalance and general dataset diversity may hinder the ability of our models to capture patterns in the datasets.

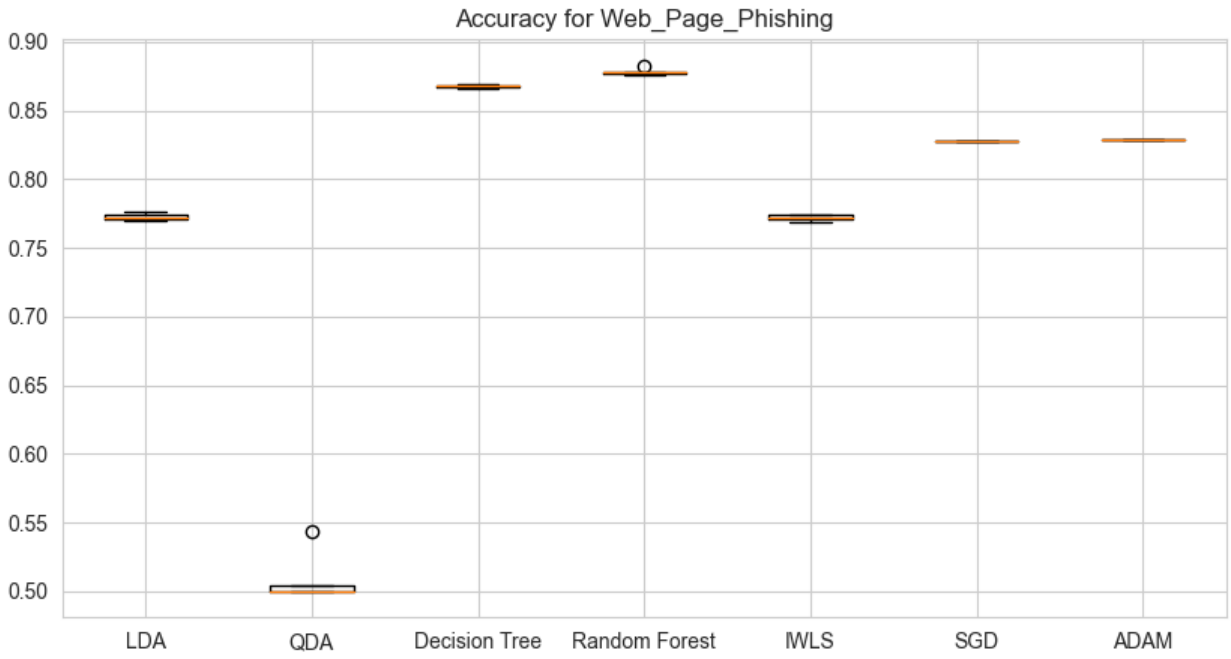

Figure 3.1 Example of a good results for Web Page Phishing dataset

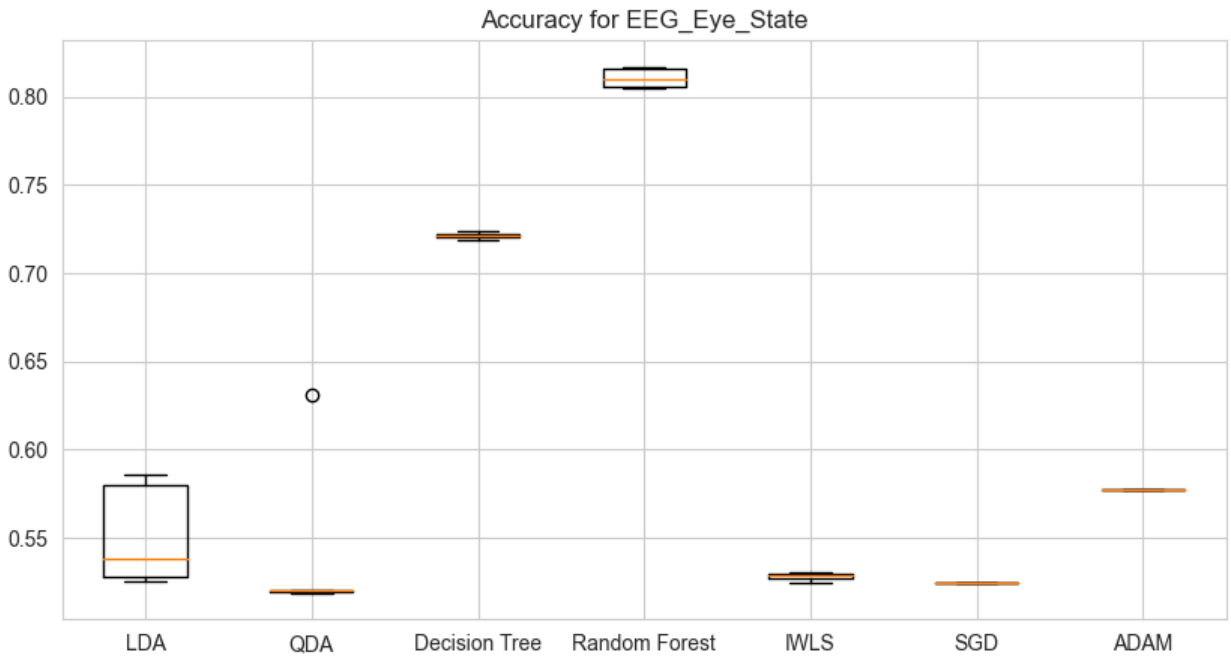

Figure 3.2 Example of results for EEG Eye State dataset

4 Comparison of classification performance of models with and without interactions

Task 3.5

In Task 3.5, our investigation centers on the comparative performance analysis of logistic regression variants. The study encompasses both standard logistic regression and its interaction-augmented counterpart across three algorithms: Iterative Weighted Least Squares (IWLS), Stochastic Gradient Descent (SGD), and Adaptive Moment Estimation (ADAM). Each method was run with and without feature interactions, resulting in six distinct logistic regression variants for analysis on small datasets.

Each logistic regression variant was subjected to 10 iterations of evaluation, using a 70-30 train-test split to measure performance. This approach aimed to provide a reliable performance estimate, considering potential variability. The metrics for comparison included balanced accuracy, F1 score, precision, and recall. This provided an assessment of each model's classification capabilities. We focused on small datasets: Rice (Cammeo and Osmancik), Statlog (Shuttle), and Banknote Authentication, to observe the effects of interaction terms in logistic regression models on datasets with fewer features.

The introduction of interaction terms into logistic regression models was based on the premise that feature interactions can capture complex relationships in the data that single features cannot. This addition is expected to enhance the model's predictive capacity, especially in datasets where features are interdependent.

Results discussion per dataset.

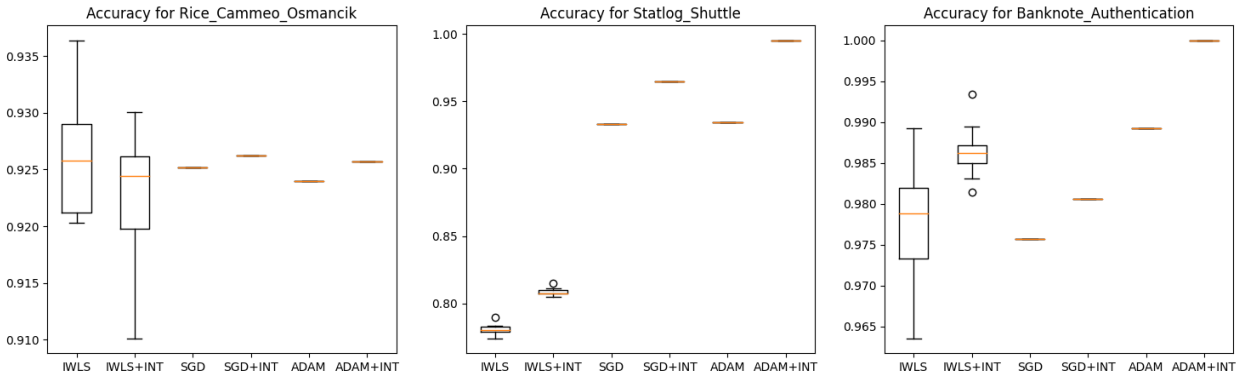

Figure 4.1 Results for Task 3.5

Rice (Cammeo and Osmancik)

- Without Interactions: The standard logistic regression methods managed to capture the dataset's variance to a reasonable extent. The model with the most variance in results is IWLS perhaps due to lack of training.
- With Interactions: The IWLS+INT model did not indicate an improvement in classification precision. For SGD and ADAM the improvement was not dramatic.

Statlog (Shuttle)

- Without Interactions: All methods performed satisfactorily, with the IWLS being the lowest.
- With Interactions: The models with interactions, especially ADAM+INT, showed a notable increase in performance, likely due to the models capturing complex patterns within the multiclass dataset.

Banknote Authentication

- Without Interactions: Given the dataset's simplicity and binary nature, the base regression models yielded high scores across all metrics.
- With Interactions: The introduction of interaction terms did not significantly alter the performance, suggesting that the main effects were already sufficient in distinguishing between authentic and fake banknotes.

References

[1] UCI Machine Learning Repository. Rice Dataset (Cammeo and Osmancik). Available at: https://archive.ics.uci.edu/dataset/545/rice+cammeo+and+osmancik.

[2] UCI Machine Learning Repository. Statlog (Shuttle) Dataset. Available at: https://archive.ics.uci.edu/ml/datasets/Statlog+%28Shuttle%29.

[3] UCI Machine Learning Repository. Banknote Authentication Dataset. Available at: https://archive.ics.uci.edu/ml/datasets/banknote+authentication.

[4] UCI Machine Learning Repository. Online Shoppers Purchasing Intention Dataset. Available at: https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset.

[5] UCI Machine Learning Repository. EEG Eye State Dataset. Available at: https://archive.ics.uci.edu/dataset/264/eeg+eye+state.

[6] Kaggle. Web Page Phishing Dataset. Available at: https://www.kaggle.com/datasets/danielfernandon/web-page-phishing-dataset.

[7] Kaggle. Airline Passenger Satisfaction Dataset. Available at: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction.

[8] Kaggle. Dataset for Link Phishing Detection. Available at: https://www.kaggle.com/datasets/winson13/dataset-for-link-phishing-detection.

[9] OpenML. Optdigits Dataset. Available at: https://www.openml.org/search?type=data&sort=qualities.NumberOfNumericFeatures&status=active&order=desc&qualities.NumberOfFeatures=between_10_100&qualities.NumberOfClasses=%3D_2&id=980.