Faculty of Mathematics and Information Science

WARSAW UNIVERSITY OF TECHNOLOGY

Advenced Machine Learning

Project 1

Jaremek Łukasz, Szymanek Paulina, Wysocka Patrycja

April 3, 2024

Contents

List of Figures

List of Tables

1 Metodology

1.1 Datasets

We have included all data sets in Table 1 along with a description of their dimensions and the transformations that have been performed on them. They were taken from the following websites:

- `https://www.openml.org/`
- `https://www.kaggle.com/`
- `https://archive.ics.uci.edu`

1.2 Stopping rule

Our stopping rule monitors the change in **log likelihood** during optimization, which is denoted as $\mathcal{L}$ and calculated using predicted probabilities *predictions* and actual labels y for binary classification. The formula for log likelihood is:

$\mathcal{L} = \sum_{i=1}^{n} \left(y_i \log(\text{predictions}_i) + (1 - y_i) \log(1 - \text{predictions}_i) \right)$

The stopping criterion compares the difference between the current log likelihood and the previous log likelihood. If this difference falls below a predefined *threshold*, typically denoted as ϵ, convergence is assumed, in our case as a default value is $\epsilon = 1e-5$. The convergence condition is expressed as:

$|\mathcal{L}_{\text{current}} - \mathcal{L}_{\text{previous}}| < \textit{threshold}$

When this condition is met, the optimization process is deemed to have converged, and the iterative optimization loop is terminated. This method ensures that the optimization stops when further iterations yield marginal improvements in log likelihood, signifying convergence of the model training process.

1.3 Performance measure

As perfomance tests, we performed 10 iterations of each model on each dataset. Each time we generated a new train-test split, where the test size was 0.2. We also used StandardScaler from the scikit-learn library. As the evaluation method we used **balanced accuracy** which is a metric that combines sensitivity and specificity, providing a fair assessment of classification performance, especially in imbalanced datasets. It ensures that the model's accuracy isn't biased towards the majority class and accounts for correct predictions across all classes. The formula is shown below:

$\text{Balanced Accuracy} = \frac{1}{2} \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP} \right)$

Name	Shape	Transformations
Banana quality	7x8000	Quality labels are mapped from strings to integers and then converted to floats.
Biodegradable	31x1055	Class labels are mapped from byte-strings to integers and then converted to floats, and specific columns are dropped due to string correlation with others.
Climate	17x540	Class labels are mapped from byte-strings to integers and then converted to floats, and specific columns ('V1', 'V2', 'V4') are dropped, because they were strongly correlated with others.
Diabetes	8x768	Class labels are mapped from byte-strings to integers and then converted to floats.
Heart attack analysis	13x303	The last column converted to floats.
Ionosphere	31x351	First column (labels) converted to floats and the remaining columns to floats
Japanese Vowels	14x9961	Binary labels are mapped from byte-strings to integers and then converted to floats.
Plates	23x1941	Class labels are mapped from byte-strings to integers and then converted to floats, and specific columns are dropped due to strong correlation with others.
Water Quality	10x3276	Columns with missing values were filled with mean value. Feature columns are converted to floats and the last column with label also to floats.

Table 1: Summary of Datasets: This table presents a detailed overview of nine distinct data sets, highlighting their respective shapes (in terms of rows and columns) and the specific preprocessing transformations applied, such as label encoding, type conversion, and selective feature exclusion.

2 Convergence analysis

Convergence analysis was done for all of the 9 datasets. It was performed on the train data, which was 80% of the whole datasets. For the analysis, the value of log-likelihood function mentioned in 1.2 was check over the number of iterations in which the algorithm stopped.

The results of the analysis for chosen datasets are shown in 1.

In case of ADAM algorithm, the stopping rule was not met for all the datasets and it did not converge in 500 iterations. The change in log-likelihood function is almost linear.

For the IWLS algorithm, convergence was achieved in the least number of iterations, under 50, amongst the implemented algorithms and the stopping rule was called in all datasets. It can be observed that the IWLS did not perform well in case of the small datasets.

The most varied results can be seen in the results of SGD algorithm. Depending on the dataset, the algorithm converges in under 100 iterations, triggers the stopping rule, but may as well converge in few hundred iterations. For both SGD and IWLS, the biggest rise in lok-likelihood value can be observed in early iterations and after that the difference is much smaller.

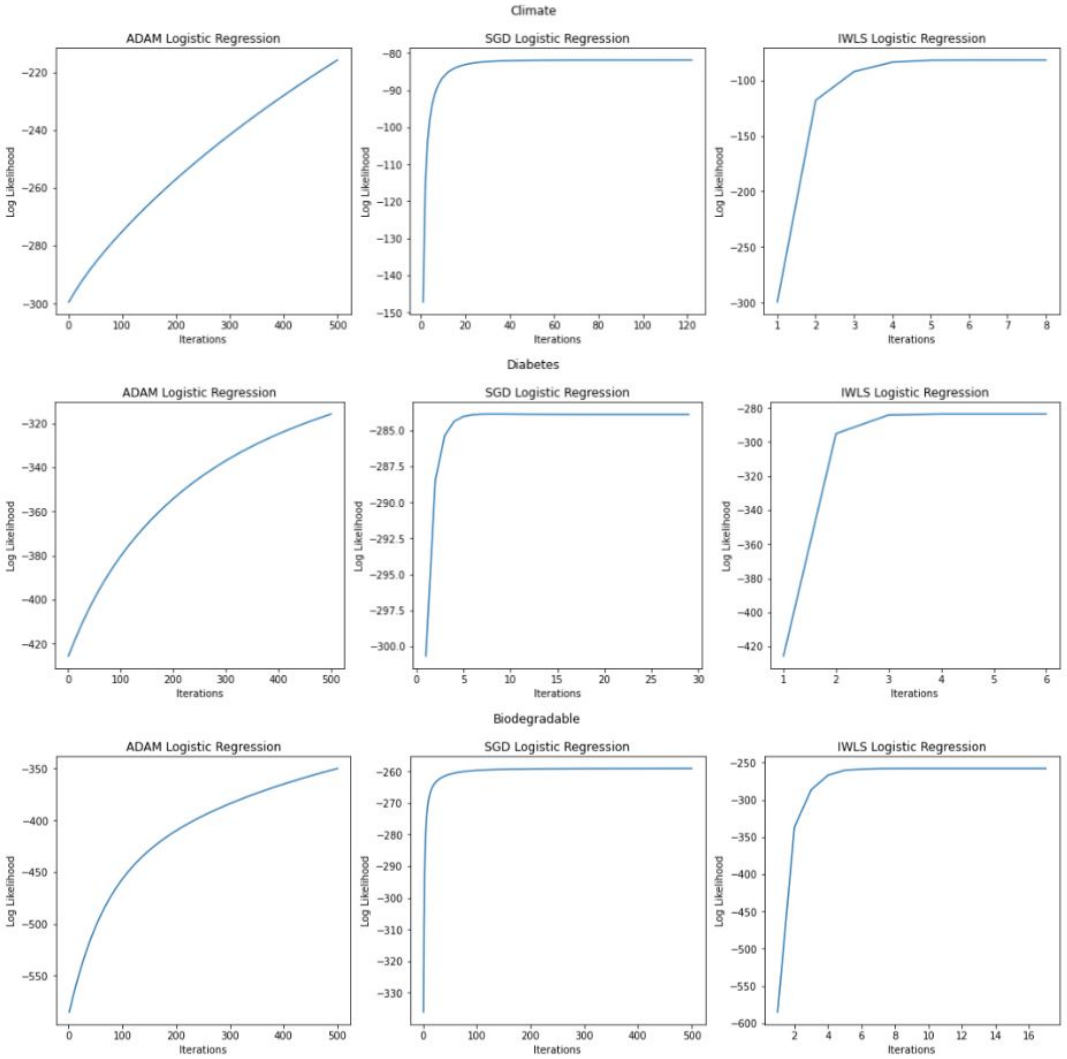

Figure 1: Log-likelihood function value over the number of iterations.

3 Comparison of classification performance

The performance of the three implemented logistic regression methods (IWLS, ADAM, SGD) was compared to four popular classification models: Linear Discriminant analysis), Quadratic Discriminant Analysis, Decision tree and Random Forest. Available implementations from scikit-learn Python library were used. The comparison was done using balanced accuracy and methodology mentioned in 1.3.

The results of the conducted research vary depending on the data set and have no common denominator. The only observation that can be drawn from the analysis of all datasets is that the QDA, Decision Tree and Random Forest methods perform better or comparable to all the rest.

For example, the performance on the Banana Quality dataset (Figure 2) shows that the QDA, Decision Tree and Random Forest methods are clearly better than the others, while on the Hear Attack dataset (Figure 3) they are worse or comparable.

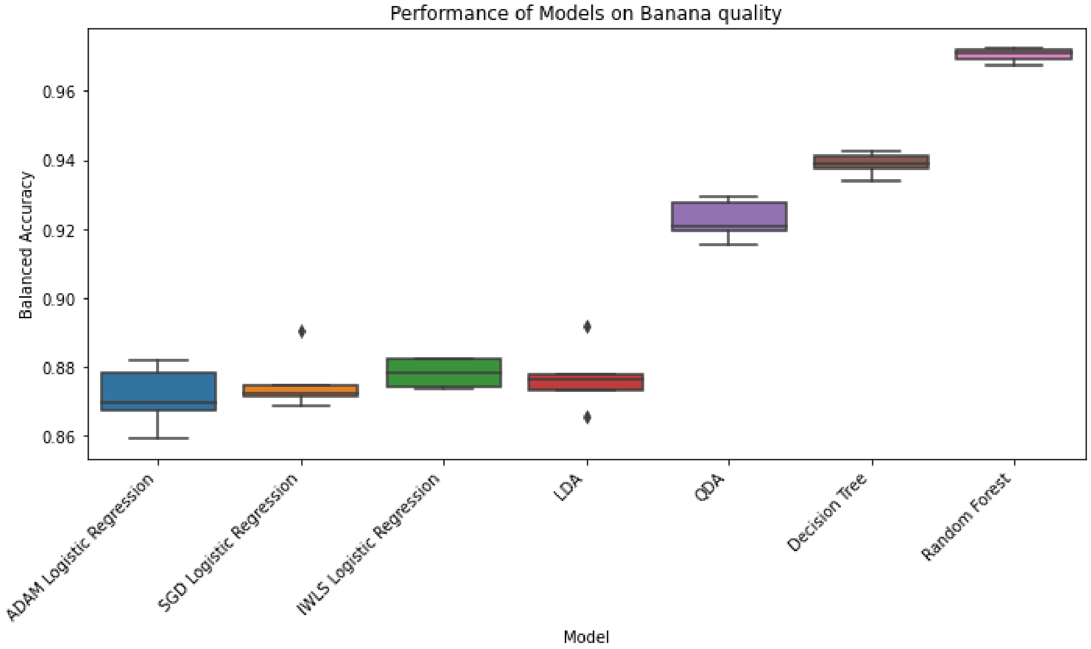

Figure 2: Model performance with other methods on Banana dataset.

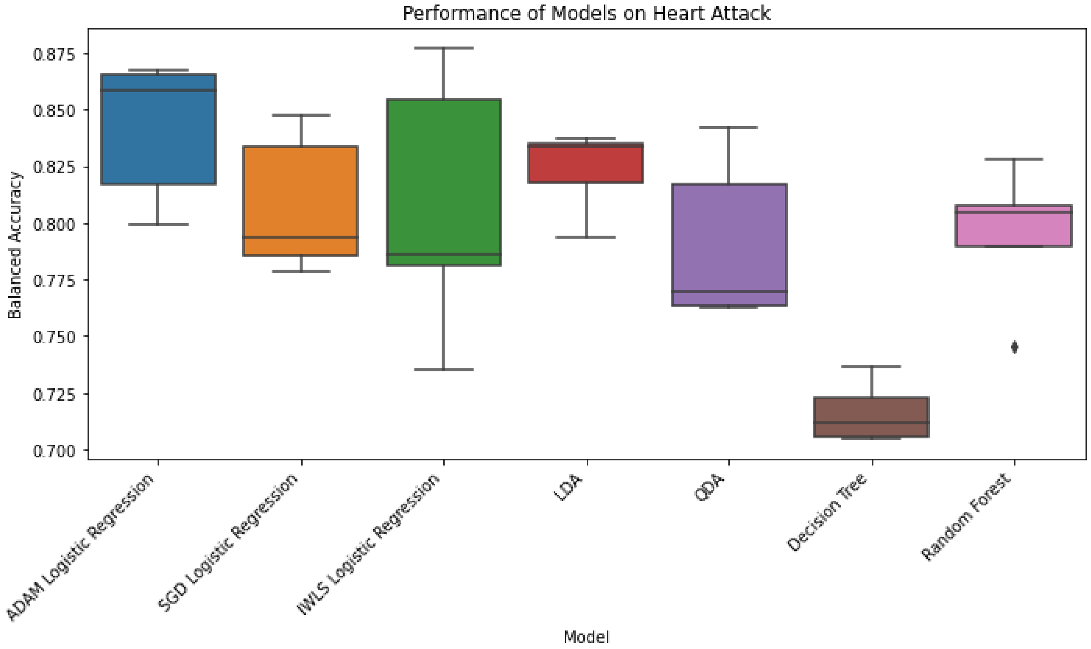

Figure 3: Model performance with other methods on Heart Attack dataset.

4 Comparison of classification performance of models with and without interactions

In case of small datasets, model performance was compared on original dataset and a version of the dataset with interactions between variables. The interaction is defined as a product of two variables. The achieved results are shown in 4, 5 and 6. In case of Water Quality and Banana Quality datasets, adding the interactions yielded higher balanced accuracy values. For both datasets, the difference between the accuracy values is similar for all models. As for the Diabetes dataset, better results can be seen only in case of IWLS algorithm and even then, the difference is not significant. For the ADAM and SGD, accuracy was lower for the dataset with interactions, although the difference is small. The conclusion can be that the results of adding interactions between variables to the datasets varies between datasets, but mostly it will either improve the results or will not have a significant effect on the final score.

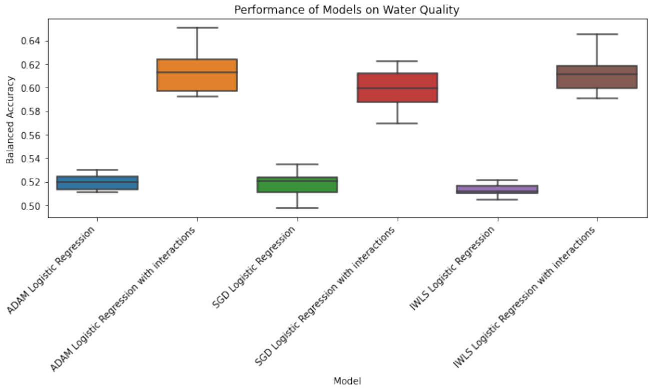

Figure 4: Model performance with and without interactions on Water dataset.

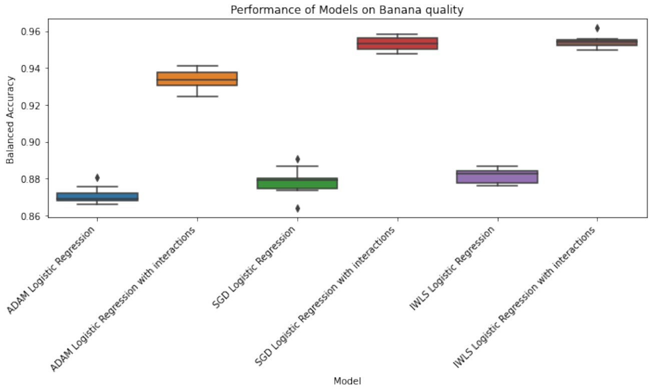

Figure 5: Model performance with and without interactions on Banana dataset.

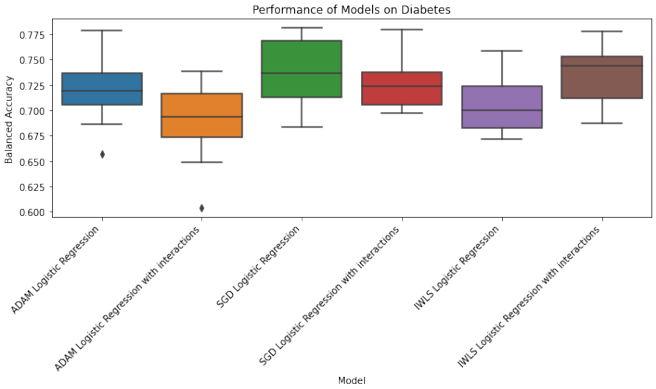

Figure 6: Model performance with and without interactions on Plates dataset.