**London School of Economics and Political Science**

**Department of Statistics**

**MSc Capstone Project 2022-2023**

*NB: Please write separate feedback and mark for the report (80% of the final grade) and provide a final (agreed) mark. Also, provide separate feedback BUT consolidated (agreed) marks for individual contributions (20% of the final grade). Add your names and date at the end of the document.*

**Title of project:** Using Remote Sensing to Improve Efficiency of Water Pollution Monitoring

**Student candidate numbers:** 44326, 44384, 52208

**Part one: Comments and feedback – Report**

**First examiner:**

The report presents results on the performance of predictive models for predicting sewage pollution from ocean satellite data along the UK shoreline. Several different prediction methods are examined, including logistic regression, multi-layer perceptron, random forest, and convolution neural network.

The executive summary well summarises the results of the study. The specific features such as BBP, CDM, KD490, and CHL could have been briefly defined in the executive summary. The executive summary could have summarised the best achieved prediction accuracies and their implications for applications.

The datasets used in the study are well described in Section 2.

The methodology is overall well explained in Section 3. The predictive models include standard machine learning classifiers such as logistic regression, multi-layer perceptron, and random forests, and convolutional neural networks which are commonly used for image recognition tasks. The likelihood function in Equation (3.2) is not correctly defined – missing labels. The concept of a stochastic average gradient descent is mentioned in Section 3.3.2 without being properly defined. The multi-layer perceptron could have been more clearly defined. For example, it is stated that a multi-layer perceptron has three layers of nodes, which is not true in general. A multi-layer perceptron has multiple layers, including input, output and one or more hidden layers. Figure 14 shows a specific multi-layer perceptron with two hidden layers. The loss minimisation process described in Section 3.3.4 is not specific to a multi-layer perceptron but it applies more generally to supervised learning – this part is redundant.

The numerical results provide a comparative analysis of different models studied using standard metrics for evaluation of classifiers, such as F1, precision, recall, AP and ROC-AUC. Random forest outperformed all other models with respect to all metrics except for recall. For recall, multi-layer perceptron achieved best results, while convolutional neural network outperformed random forest.

Overall, the classification task turned out to be challenging with even the best models achieving large false positive rates. This is admitted in in both executive summary and conclusion, which is good. The report could have provided a summary of results with some concrete numbers that can be easily interpreted by practitioners, besides evaluation metrics reported in the numerical results section.

The engineered features consisted of different aggregations of various features over the spatial domain, which are fed to classifiers. This is different from the use of convolutional neural networks which use original input without aggregation. The results obtained are such that the latter approach (for the given data and models considered) does not yield prediction accuracy gains over prediction methods that use engineered aggregate features.

Overall, this is a good report that studies an interesting and well defined problem. Some results are obtained that may prove useful to practitioners.

**Mark awarded for the report:** 70

**Second examiner:**

The focus of the project is on the exploration of low-cost and scalable solutions (in contrast to even duration monitors and manual observations) to predict sewage pollution in the UK, which is a fundamental question in the area of environment protection. The authors consider using remote-sensing data (satellite images from the Copernicus Marine Services, with sewage pollution labels from the Environment Agency) to tackle the task, relying on two satellites (Sentinel2 =: S2, Sentinel3 =: S3) providing data of complementary nature (S2 with higher spatial resolution and lower temporal frequency, and S3 with the other way around). The problem was phrased as a binary classification task (presence / absence of pollution) and the efficiency of 4 classification schemes (logistic regression, random forests, multi-layer perceptron and convolutional neural network) were investigated, with statistics of various proxy variables, varying window sizes, using multiple strategies for missing data substitution. The major challenges tackled were the large amount of missing data (for instance due to cloudy days), and the imbalanced nature of the dataset (pollution happens rarely).

After the presentation of the datasets used, exploratory data analyis, and feature engineering, the authors describe in detail their predictive models. Careful cross-validation is carried out to select the best-performing hyperparameters of the chosen prediction techniques, alongside with the summary of the best performing models, illustration across various categories (such as years, month, and regions), comparison of raw and engineered features, performance comparison on the S2 vs the S3 data, and study of the impact of the window size. The results are thoroughly explained, with practical guide to the practitioner (which is always a plus). I also enjoyed the discussion on the limitations of the methods (which is appreciated even at the largest ML venues) and the future research ideas.

Congratulations to the team on the nice work!

A few questions which come up in me while reading the report:

1) A found the introduction on related works somewhat implicit, it was not clear what the existing / related techniques n the domain are, and how the proposed approach compares against them (at least conceptually).

2) Have the authors considered dependence measures capable of capturing nonlinear dependencies instead of correlation? Representative examples include for instance Shannon mutual information (which can be estimated consistently using pairwise distances of the points, with k-nearest neighbor computation), or kernel methods (which again give rise to consistent estimators via simple linear algebra).

3) The authors talked about computational limitations. As a reader, I would have been curious to see what the computational-accuracy tradeoffs were, how much time training the models took?

4) Some of the optimization techniques seemed to be non-deterministic. How robust the results are w.r.t. multiple runs of the optimizers? (It would have been nice to see for instance \pm std in the tables.)

**Mark awarded for the report:** 70

**Final (agreed) mark for the report:** 70

[0-19 represents a bad fail, 20-49 - fail, 50-59 - pass, 60-69 - merit, 70+ - distinction]

**Part Two: Comments and feedback – Individual contributions**

**Candidate number:** 44326

**Comments (first examiner):** No individual report was uploaded to GitHub. This candidate contributed some exploratory data analysis, predictions by using a convolutional neural network, and some key contributions to the report writing. This candidate had the lead role in communicating with supervisors in both email correspondence and team meetings. The other candidates had more passive role in this regard. This was according to an agreed division of roles among the team members.

**Comments (second examiner):** No individual report was uploaded to GitHub.

**Individual contribution mark (agreed):** 80

**Candidate Number:** 44384

**Comments (first examiner):** This candidate contributed to feature engineering, computing evaluation metrics, principle component analysis, clustering, prediction results by using logistic regression and multi-layer perceptron. For personal development, the candidate may try being more actively engaged in team meetings.

**Comments (second examiner):** This candidate contributed actively to the project by multiple functions (evaluating performance), dimensionality reduction (PCA), clustering (k-means), logistic regression, multi-layer perceptron, and hyperparameter tuning of random forests.

**Individual contribution mark (agreed):** 70

**Candidate number:** 52208

**Comments (first examiner):** No individual report was uploaded to GitHub. This candidate contributed to data acquisition, and applying feature engineering, exploratory data analysis. For personal development, the candidate may try being more actively engaged in team meetings.

**Comments (second examiner):** No individual report was uploaded.to GitHub.

**Individual contribution mark (agreed):** 70

**Date:** 30/08/2023