Configuration Manual

MSc Research Project
Data Analytics

Alina Mathew
Student ID: 23412399

School of Computing
National College of Ireland

Supervisor: Dr. Abdul Razzaq

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Alina Mathew
Student ID:	23412399
Programme:	Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Dr. Abdul Razzaq
Submission Due Date:	28/01/2026
Project Title:	Configuration Manual
Word Count:	659
Page Count:	5

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	ALINA MATHEW
Date:	28th January 2026

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Alina Mathew
23412399

1 Configuration Manual

This configuration manual contains the required software setting and critical settings that are needed to replicate the experiments conducted in this research project.

1.1 Execution Environment

All the experiments were performed in a Google Colab environment.

- **Platform:** Google Colab
- **Runtime:** Python 3
- **Hardware accelerator:** GPU enabled (NVIDIA T4 or similar), when available
- **RAM:** $\approx$ 12–15 GB (standard Colab runtime)

All years of the survey (2022, 2024, 2025) were done with the same configuration.

1.2 Python Packages

Python libraries that were utilized in the project were:

2 Configuration Manual

All the experiments were run in a generic Google Colab Python 3 GPU session. The code used in the study was based on the following versions of the packages and is adequate to recreate the entire preprocessing, modelling, and analysis pipeline:

- `python` 3.10 (Colab default)
- `numpy` 1.26.4
- `pandas` 2.1.4
- `scikit-learn` 1.3.2
- `imbalanced-learn` 0.12.2 (SMOTE)
- `matplotlib` 3.7.1

- `seaborn` 0.12.2
- `shap` 0.45.0
- `lightgbm` 4.1.0 (FLO feature selection)
- `torch` 2.1.0+cu121 (Residual MLP joint model)
- `autogluon.tabular` 1.1.0 (AutoML baseline)
- `google-colab` utilities (file upload/download)

All the codes were interactively run within Google Colab notebooks. No external or local cluster architecture was needed. Tthis applies to all survey years data(2022, 2024, 2025).

2.1 Data Inputs

For each survey year (2022, 2024, 2025), a preprocessed CSV file was used as the modelling dataset. Each file were consistently encoded for comparability:

- All predictors encoded as numeric or ordinal integer-coded features.
- For example in 2024, three DevX targets:
  - `Frustration` (ordinal, 3 levels)
  - `JobSat` (ordinal, 3 levels)
  - `JobPerspectiveClass` (ordinal, 3 levels)
- Feature subset defined by FLO (union of top 5 features per target).

Colab had a standard file upload widget through which data files were uploaded.

2.2 Preprocessing and Feature Selection

The configuration used in all years was as follows:

- Missing values imputed using MICE via `IterativeImputer` (BayesianRidge base estimator, maximum 20 iterations, random state 42).
- Categorical/ordinal predictors restored to discrete scales by rounding/clipping after imputation.
- FLO feature selection using LightGBM classifiers:
  - `n_estimators = 400`, `learning_rate = 0.05`, `subsample = 0.9`.
  - Scoring metric: balanced accuracy.
  - Top 5 features per DevX target; final feature set = union of all selected features.

2.3 Joint Label and Sampling Configuration

- Joint label constructed by factorising the tuple `(Frustration, JobSat, JobPerspectiveClass)` into a single multiclass label (up to 27 classes).
- For the neural model, SMOTE oversampling (`k_neighbors = 3`, random state 42) applied *only* to the training split, stratified on the joint label.

2.4 Train–Test Splitting and Scaling

- **Joint MLP:** 80% train, 10% validation, 10% test (stratified on the joint label).
- **AutoML baseline:** 80% train, 20% test (stratified on the joint label).
- For the neural model, features standardised using `StandardScaler` fitted on the training data and applied to validation and test sets.

2.5 Model Configurations

Residual MLP (Joint DevX Model)

- Input: FLO-selected numeric features.
- Architecture: feed-forward Residual MLP with:
  - Hidden dimension: 256
  - Depth: 5 residual blocks
  - Activation: GELU
  - Dropout: 0.3 (stem), 0.2 (blocks)
  - Output: softmax over joint classes
- Optimiser: AdamW (`learning_rate = 3e-4`, `weight_decay = 1e-4`).
- Batch size: 1024, maximum 80 epochs.
- Early stopping: based on validation macro-F1, patience = 12 epochs.
- Learning-rate schedule: cosine annealing with warm restarts.

AutoML Baseline (AutoGluon Tabular)

- Label: joint class ID.
- Problem type: multiclass classification.
- Presets: `best_quality`.
- Evaluation metric: balanced accuracy.
- Time limit: 3600 seconds per year.

2.6 Explainability and Clustering

- Calibrated Random Forests trained per DevX target (200 trees, isotonic calibration).
- SHAP values computed using `TreeExplainer` from `shap`.
- Gaussian Mixture Models for Latent Profile Analysis:
  - Covariance type: full
  - Random state: 42
  - Candidate components: $k = 2 \dots 7$; selection by minimum BIC.
  - Stability assessed via bootstrap Adjusted Rand Index (ARI).

2.7 Reproducibility Settings

- Random seed: 42 for NumPy, scikit-learn, and PyTorch where applicable.
- All trained models, prediction outputs, and clustering labels saved as CSV or model checkpoints within the Colab environment.

This structure is adequate to replicate the preprocessing, feature selection, supervised modelling, explainability and latent profile analysis adopted in the study.

References

Baldassarre, M. T., Caivano, D., Fucci, D., Romano, S. and Scanniello, G. (2022). Affective reactions and test-driven development: Results from three experiments and a survey, *Journal of Systems and Software* **185**(22).

Beecham, S., Baddoo, N., Hall, T., Robinson, H. and Sharp, H. (2008). Motivation in software engineering: A systematic literature review, *Information and Software Technology* **50**: 860–878.

Butler, J. and Jaffe, S. (2021). Challenges and gratitude: A diary study of software engineers working from home during COVID-19 pandemic, *Proceedings of the 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP '21)*, IEEE, pp. 362–363.

Fagerholm, F. and Jürgen, M. (2012). Developer experience: Concept and definition, *2012 International Conference on Software and System Process (ICSSP)*, pp. 73–77.

Forsgren, N., Kalliamvakou, E., Noda, A., Greiler, M., Houck, B. and Storey, M.-A. (2024). Devex in action: A study of its tangible impacts, *Queue* **21**(6): 47–77.

França, C. A., Silva, F. and Sharp, H. (2018). Motivation and satisfaction of software engineers: A large-scale study, *IEEE Transactions on Software Engineering* **46**(2): 118–140.

Girardi, D., Lanubile, F., Novielli, N. and Serebrenik, A. (2022). Emotions and perceived productivity of software developers at the workplace, *IEEE Transactions on Software Engineering* **48**(9): 3326–3341.

Godliauskas, P. and Šmite, D. (2024). The well-being of software engineers: a systematic literature review and a theory, *Empirical Software Engineering* **30**(1): 35.

Gonçales, L., Farias, K., da Silva, B. and Fessler, J. (2019). Measuring the cognitive load of software developers: A systematic mapping study, *International Conference on Program Comprehension*, pp. 42–52.

Graziotin, D., Fagerholm, F., Wang, X. and Abrahamsson, P. (2018). What happens when software developers are (un)happy, *Journal of Systems and Software* **140**: 32–47.

Greiler, M., Storey, M.-A. and Noda, A. (2023). An actionable framework for understanding and improving developer experience, *IEEE Transactions on Software Engineering* **49**(4): 1411–1425.

Guthardt, T., Kosiol, J. and Hohlfeld, O. (2024). Low-code vs. the developer: An empirical study on the developer experience and efficiency of a no-code platform, *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems*, MODELS Companion '24, Association for Computing Machinery, New York, NY, USA, pp. 856–865.

Meyer, A. N., Barr, E. T., Bird, C. and Zimmermann, T. (2021). Today was a good day: The daily life of software developers, *IEEE Transactions on Software Engineering* **47**(5): 863–880.

Meyer, A., Zimmermann, T. and Fritz, T. (2017). Characterizing software developers by perceptions of productivity, *International Symposium on Empirical Software Engineering and Measurement*, pp. 105–110.

Novielli, N., Calefato, F. and Lanubile, F. (2014). Towards discovering the role of emotions in stack overflow, *Proceedings of the 6th International Workshop on Social Software Engineering*, pp. 33–36.

Patnaik, A. and Padhy, N. (2024). A hybrid approach to enhance software quality by sentiment analysis of developer, *Computing, Communication and Learning*, Springer Nature Switzerland, Cham, pp. 113–125.

Razzaq, A., Buckley, J., Lai, Q., Yu, T. and Botterweck, G. (2025). A systematic literature review on the influence of enhanced developer experience on developers' productivity: Factors, practices, and recommendations, *ACM Computing Surveys* **57**(1): 1–46.

Sampson, M. (2025). Enhancing developer experience: Burnout, resilience, and the digital environment in software engineering, *Proceedings of the 38th International BCS Human-Computer Interaction Conference*, BCS HCI '25, BCS Learning & Development Ltd, Swindon, GBR, pp. 369–374.

Wright, T. and Cropanzano, R. (2000). Psychological well-being and job satisfaction as predictors of job performance, *Journal of Occupational Health Psychology* **5**(1): 84–94.