Universitat Pompeu Fabra Barcelona

UPF – Gestió de projectes

GP2526 Project Work Plan

AIGÜES DE BARCELONA - INCIDÈNCIES EN COMPTADORS INTEL·LIGENTS

Document Control Information

Settings	Value
Document Title:	Challenge Final Report
Project Title:	Water and Urban Resilience: Impact of Heatwaves and Intense Rainfall in Barcelona
Team Number:	102F
Project Core Team:	Arnau Rodon Comas, Asul Garcia Pereira, Jordi Esteve Claramunt, Albert Jané Lardiés and Nahia Anaut Adell
Project Manager:	Arnau Rodon Comas

Revision	Date	Created by	Short Description of Changes
v1.0	30/11/25	PCT	First complete version of the Final Report consolidating project context, data preparation, modeling architecture, evaluation, and value narrative into a single client-ready document.

Table of Contents

1. Executive Summary

This report documents the outcomes of project **“Incidències en Comptadors Intel·ligents”**, conducted within the GP2526 course for Aigües de Barcelona. The project develops an end-to-end analytical pipeline to detect and forecast anomalies in daily water consumption recorded by smart meters. Starting from a large-scale operational dataset with approximately 5 million daily records and enriched with weather information, the team delivers a validated unified dataset, a set of statistical analyses, a two-layer anomaly analytics architecture, and an evaluation and visualization pack aimed at supporting predictive maintenance decisions.

The solution is built around a **two-stage modeling strategy**. An unsupervised k-means component characterises brand-specific “normal” consumption patterns and flags anomalous days as those in the tail of the distance distribution to cluster centroids. On top of these proxy anomaly labels, a supervised LSTM model operates on 90-day sliding windows to forecast the probability that an anomaly will occur in the next day, week or month, using the full engineered feature space including temporal encodings, rolling statistics, hardware descriptors and weather context. This architecture enables both the generation of interpretable anomaly indicators and the estimation of short-term anomaly risk at multiple horizons.

The project has been executed in three iterations aligned with the Work Breakdown Structure: data readiness and EDA (Iteration 1), modeling and evaluation (Iteration 2), and synthesis and storytelling (Iteration 3). Iteration 1 delivers the unified dataset and validation report; Iteration 2 delivers statistical analysis, model baselines; Iteration 3 focuses on this final report, the visualization pack and the final presentation. At the time of reporting, all modeling work is complete and the project is on track with respect to scope and schedule, with remaining effort centred on packaging and communication.

Results indicate that the combined k-means + LSTM pipeline can identify meters and periods with elevated anomaly risk while preserving operational control over false-positive rates through threshold selection. The absence of authoritative anomaly labels, however, means that evaluation relies on proxy thresholds and policy assumptions, which limits the confidence with which absolute values of recall and false-positive rate can be interpreted. Despite this limitation, the project demonstrates that a robust, reproducible and explainable anomaly analytics system can be built over the current data assets, and it provides a foundation for future operationalisation and refinement with business-approved anomaly policies.

2. Background and Motivation

Smart water meters deployed by Aigües de Barcelona generate large volumes of daily consumption data at customer level. This data is a key enabler for **early detection of meter malfunctions and anomalous consumption patterns**, which in turn can reduce non-revenue water, avoid service disruptions, and support more sustainable management of urban water networks. However, the raw data is high-dimensional, zero-inflated and subject to data quality issues, and the organisation does not yet operate with a fully specified and approved anomaly policy. These factors make it challenging to design, implement and evaluate anomaly detection systems that are both technically sound and operationally credible.

Within this context, the GP2526 project defines a ten-week analytical effort structured in three iterations. The objectives are to (i) build a unified, validated dataset and reproducible cleaning pipeline; (ii) conduct statistical and correlation analyses to characterise consumption behaviour and anomaly patterns; (iii) design and train model baselines for anomaly detection and short-term forecasting; and (iv) package the results into an evaluation pack, a visualization pack and a final report that together address the product requirements Rq1–Rq7. The project is governed through a Project Steering Committee (PSC) under a standard work-plan and status-reporting framework, with Aigües de Barcelona as conceptual client and the GP2526 challenge jury as the ultimate evaluator of outcomes and impact.

The work is constrained by the academic calendar, limited computation resources, null access to the client, and a lag in the delivery of the full dataset. These constraints shaped the WBS and required a re-baselining of Iteration 2 around the actual data delivery date, as well as a focus on analytical prototypes rather than production-grade deployment and real-time alerting. Despite these limitations, the project aims to provide Aigües de Barcelona with a concrete demonstration of how statistical analysis and machine learning can be combined to generate actionable anomaly insights from existing data assets, and to produce documentation and artefacts that would allow a future implementation team to continue the work with minimal re-engineering.

3. Product Deliverables

This section summarises the main product deliverables corresponding to the analytical scope of the project: the unified dataset and data-cleaning pipeline, the validation and statistical analysis artefacts, the anomaly detection and forecasting models, and the evaluation and visualization packs. These deliverables map directly to WBS elements 1.1.x and 1.2.x, and collectively support requirements Rq1–Rq4 and Rq7 as defined in the Project Work Plan.

3.1. Unified Dataset and Data Cleaning Pipeline

The **unified dataset v1** consolidates smart-meter consumption records from Aigües de Barcelona into a single, analysis-ready table containing approximately 17 million daily observations. Each record corresponds to a unique combination of supply contract (POLIZA_SUMINISTRO) and date (FECHA), with associated attributes describing consumption, customer location, hardware characteristics and installation metadata. The dataset is complemented with joined weather variables (temperature and precipitation) at municipality level, enabling contextual analysis of consumption patterns.

The data-cleaning pipeline, implemented as reproducible scripts under WP 1.1.1.a, performs a sequence of validation and transformation steps. During loading, the pipeline verifies that all expected columns are present and correctly named, and converts date fields (FECHA and DATA_INST_COMP) from generic object types to datetime64 to allow temporal filtering and integration. Categorical identifiers such as POLIZA_SUMINISTRO and MARCA_COMP are checked for structural consistency, while numeric fields that conceptually encode categories, such as CODI_MODEL and DIAM_COMP, are cast to categorical types to reflect their discrete nature.

Missingness analysis reveals that consumption and core identifiers are fully populated, whereas several hardware and location fields exhibit substantial missingness. This pattern is consistent with records corresponding to pre-installation periods. The pipeline therefore enforces a temporal integrity rule: only observations with FECHA greater than or equal to DATA_INST_COMP are retained. This filter removes most structurally incomplete records and yields a dataset representing meters in operational state. Additional checks ensure uniqueness of the (POLIZA_SUMINISTRO, FECHA) key; redundant duplicates with identical attributes are removed, and the data is verified to contain no negative values in CONSUMO_REAL. The resulting unified dataset, along with the cleaning scripts, satisfies requirement Rq1 on data consistency and integrity and serves as the single source of truth for subsequent analyses and modeling.

3.2 Data Validation, EDA and Statistical Analysis

Data validation is documented in the **Validation Report (1.1.1.b)**, which profiles schema conformity, missingness, key uniqueness and the success of the weather integration. The report confirms that the cleaned dataset exhibits a consistent structure with minimised missingness, correct data types and unique contract-date combinations after deduplication, and that the merge with weather data via keys FECHA and NUM_MUN_SGAB preserves temporal and key integrity. This validation underpins the decision to treat the unified dataset as reference input for all later work packages.

Building on this foundation, the **Statistical Analysis Report (1.2.1)** explores distributions, temporal patterns and relationships among key variables. Initial inspection of CONSUMO_REAL reveals a heavily right-skewed, zero-inflated distribution dominated by numerous zero readings, particularly for one widely deployed meter brand. A log(x+1) transformation yields a more symmetric distribution and exposes within-group variability, motivating the use of robust metrics such as median absolute deviation in later anomaly characterisation. Analysis by brand and diameter shows that meter diameter is the primary driver of consumption scale, with 20–30 mm meters displaying higher and more volatile consumption consistent with industrial or multi-user contexts, while 15 mm meters exhibit lower, more stable usage typical of residential customers.

Temporal analysis identifies clear weekly and seasonal patterns, including higher average consumption in summer months, which supports the inclusion of temporal encodings and rolling statistics as core features in the modeling pipeline. Correlation analysis between consumption and weather variables indicates negligible linear relationships at daily resolution, suggesting that weather acts mainly through longer-term seasonality rather than as a direct driver of short-term anomalies.

Finally, the report characterises candidate anomaly patterns and highlights that hardware characteristics (brand, diameter) and temporal dynamics are more informative for anomaly detection than raw weather signals alone, findings that feed directly into feature screening and threshold design for WP 1.2.2 Model baselines.

3.3 Anomaly Detection and Forecasting Models

The **Model Baselines deliverable (1.2.2)** implements a two-layer anomaly analytics architecture composed of an unsupervised k-means clustering stage and a supervised LSTM forecasting stage, both driven by a shared feature-engineering backbone. Raw daily observations with identifiers, consumption, hardware and weather variables are transformed into a rich set of time-series features, including log-scaled consumption, per-meter z-scores and robust z-scores, temporal encodings (cyclical month and weekday, season, weekend flags), rolling means and variances at multiple horizons, lagged values, change indicators, and one-hot encodings of meter brand, model and diameter.

In the first layer, **brand-specific k-means models** learn baseline consumption patterns using a carefully selected subset of features representing level, seasonality, volatility and trend. For each brand, the algorithm clusters daily observations, computes distances to centroids and marks the top tail (e.g. 2.5% farthest points) of the distance distribution within each cluster as anomalies. The outputs include the cluster label, distance and an anomaly-intensity score, as well as a binary is_anomaly flag. These signals serve both as proxy labels and as additional contextual features for the second modeling layer.

In the second layer, **per-brand LSTM models** operate on sliding 90-day sequences built per meter. For each window, the input is a feature matrix with one row per day and one column per engineered feature; the targets are three binary indicators denoting whether an anomaly occurs in the next day, in the next seven days, or in the next thirty days, derived by looking ahead in the is_anomaly series. The architecture uses two stacked LSTM layers followed by dense layers with batch normalisation

and dropout, and a three-unit sigmoid output layer producing per-horizon risk probabilities. To avoid information leakage, training and validation splits are performed at meter level rather than at sequence level.

Class imbalance is explicitly addressed by computing sample weights per horizon and brand, assigning higher weight to positive sequences and giving more emphasis to shorter horizons (day > week > month) in the loss function. Training is regularised using early stopping and other standard techniques to reduce overfitting, and multiple configurations are logged as part of the baseline comparison. Together, the k-means and LSTM components implement a scalable, brand-aware anomaly detection and forecasting system that satisfies requirement Rq3 on model performance and contributes to Rq4 and Rq7 through the provision of interpretable anomaly scores and multi-horizon risk outputs.

3.4 Evaluation Pack and Visualization Pack

The Evaluation Pack (1.2.3) is implemented as an interactive Streamlit application that orchestrates the full anomaly-detection pipeline end to end. Users can upload a CSV following the production schema and selectively execute feature engineering, k-means clustering, LSTM training and LSTM prediction, while the app logs each run and exposes a step-aware summary report generated by the backend pipeline. The dashboard provides structured evaluation views, including high-level run metrics (input records, engineered records, total static anomalies, number of LSTM models), per-brand and per-horizon validation tables with precision, recall, F1 and AUC, and a text summary that synthesises key findings and sanity checks for each stage without requiring direct access to notebooks or logs.

Complementing this, the Visualization Pack (1.3.2) is delivered through the same application as a set of interactive views that render the diagnostic plots and monitoring tools produced by the k-means and LSTM pipelines. For each run, the app discovers and displays k-means optimisation curves, cluster-size and anomaly-rate plots, distance distributions and representative time series with highlighted static anomalies, as well as LSTM training curves (loss, accuracy, AUC, precision, recall) and per-horizon validation summaries. The dedicated “Daily Tracker” tab allows users to select a reference date and inspect static anomalies at day, week and month horizons together with the distribution of predicted risk levels, making it possible to analyse the current health of the meter population in a way that is directly actionable for operations.

Together, the evaluation and visualisation views in the Streamlit app form a single, consistent interface between the analytical work and decision makers, satisfying requirement Rq4 on interpretability and providing a practical vehicle for model review, validation and operational uptake.

4. Justification of Design Choices

4.1 Technology Stack and Development Environment

The solution is implemented on a Python-based analytics stack, with data preparation and modeling in notebooks and scripts and an optional Streamlit interface for interactive experimentation. This choice reflects the need for rapid iteration under a ten-week academic schedule, widespread familiarity within the team, and strong ecosystem support for time-series analysis and deep learning. The Streamlit application orchestrates the end-to-end pipeline—feature engineering, k-means clustering, LSTM training and prediction—allowing different pipeline steps to be executed selectively and providing immediate feedback on outputs and metrics.

Environment and dependency management are formalised through WP 1.4.2 “Environment specification & Runbook,” which pins package versions and documents execution steps. This design is driven by requirement Rq5 (Reproducibility): the environment specification must allow a fresh clone of the repository to run the full pipeline end-to-end, and the runbook must enable another team member to reproduce outputs without assistance. GitHub is used for version control of all scripts, notebooks and visualization code, and Google Drive is used for versioned documentation, with documents referencing relevant commit hashes to preserve traceability between code and figures.

This combination ensures alignment with the course methodology while meeting the project’s technical-readiness success factor of automated, reproducible workflows.

4.2 Data Management and Storage Strategy

Data management is centred on the unified dataset v1 defined in WP 1.1.1, which acts as the single source of truth for all subsequent analyses and models. Raw smart-meter data is ingested, cleaned and validated through a scripted pipeline that standardises types, enforces key uniqueness and applies temporal integrity rules, with results documented in the Validation Report (1.1.1.b). This design directly addresses Rq1 (Data consistency & integrity) by ensuring no schema or join-key errors remain and by clearing the data-quality issues log before modeling proceeds.

From a storage standpoint, the project deliberately keeps a limited number of curated data artefacts: the unified dataset, a plotting-friendly df_plot variant, and derived feature tables for k-means and LSTM. This reduces the risk of divergence between multiple “competing” datasets and simplifies configuration management. Weather data is integrated at daily resolution using FECHA and municipality keys, but subsequent analysis shows negligible linear correlation between weather and daily consumption.

As a result, weather is retained for context and potential seasonal aggregation but is not treated as a primary driver of anomalies, which streamlines feature sets and avoids over-fitting to weak signals.

4.3 Anomaly Labeling and Scoring Methodology

Anomaly detection is implemented as a two-stage strategy, combining distribution-based statistical rules and clustering-based distance thresholds. At the statistical level, anomalies are first characterised using standard deviation, IQR and MAD-based modified z-score rules applied per brand–model–diameter group, with a reading marked anomalous if any rule flags it.

This combination is chosen to balance sensitivity (IQR, MAD) and robustness to extreme outliers (standard deviation), and it provides interpretable baselines for both EDA and subsequent design of modeling labels and thresholds.

For model training and prediction, anomalies are defined via brand-specific k-means clustering over a feature subset capturing level, volatility and recent dynamics. Days are considered anomalous if their distance to the closest cluster centroid lies in the top tail of the distance distribution within that cluster. This approach is chosen for three reasons. First, it accommodates the strong heterogeneity across brands and diameters identified in the statistical analysis, allowing each brand to have its own "normal" patterns. Second, it provides a continuous anomaly-intensity signal (distance) that can be exploited as a feature by the LSTM. Third, it avoids reliance on hand-crafted static thresholds, which are difficult to justify without a fully agreed anomaly policy.

Given the absence of authoritative anomaly labels from operations, the project adopts a proxy-label strategy under Rq2 (Anomaly labeling / evaluation protocol). The Evaluation Pack documents the chosen k-means thresholding policy, links it back to EDA findings and anomaly hypotheses, and defines reproducible steps for generating labels and computing metrics. The Status Report explicitly notes that anomaly policies and thresholds remain a pending decision with the PSC and conceptual PO, and that current labels are used provisionally with assumptions documented in the Decision & Assumptions Log. This design choice makes the evaluation transparent and revisable while still enabling quantitative assessment in this academic phase.

4.4. Modeling Strategy and Feature Engineering

The modeling architecture uses per-brand LSTM networks operating on 90-day windows to predict the probability of an anomaly occurring over three horizons (next day, week, month). The decision to build brand-specific models follows directly from the statistical analysis, which shows that meter characteristics (especially diameter and brand) dominate both consumption level and anomaly rates. Distinct models allow the network capacity to focus on homogeneous behaviour, improving learning efficiency and interpretability.

Feature engineering emphasises meter attributes, temporal context and robust statistics, in line with the conclusions of the statistical analysis. The feature set includes log-scaled consumption, rolling means and dispersion measures, percentage of zeros, temporal encodings (month, season, weekday) and encodings of brand, model and diameter, while weather is used in aggregated or secondary form due to its weak short-term predictive power. K-means outputs (cluster label, distance, anomaly-intensity score) are incorporated as additional features, enabling the LSTM to learn patterns such as sustained high anomaly scores before a future event.

Class imbalance is addressed by horizon-specific sample weights that up-weight rare positive sequences and prioritise near-term horizons.

This design aligns with the operational preference for catching imminent anomalies while still providing information about medium-term risk. To mitigate over-fitting, the models use train/validation splits at meter level, standardisation fitted only on training data, recurrent dropout, and early stopping, consistent with the project's risk mitigation strategy for over-fitting.

Alternative baselines (e.g. simpler statistical or tree-based models) were considered within WP 1.2.2, but the LSTM was retained as the primary sequential model because it directly exploits temporal structure and multi-horizon outputs, while simpler models serve as reference points and fall-back options if instability is detected.

4.5 Visualization and Stakeholder Narrative

Visualization and reporting choices are driven by Rq4 (Explainability) and Rq7 (Visualization of anomalies & performance). The Visualization Pack (1.3.2) provides plots that show anomaly flags over time, distributions of anomaly scores, performance curves and threshold trade-offs, alongside a one-page "How to interpret" guide targeted at non-technical stakeholders. These artefacts are designed to be consistent with the metrics in the Evaluation Pack (1.2.3) and to include concrete examples of true and false alarms, which are explicitly required acceptance criteria for Rq4.

The Final Report (1.3.1) and Final Presentation (1.3.3) together satisfy Rq6 (Documentation), ensuring that the executive summary, visualisations and narrative are aligned and can be reviewed and signed off by the PSC acting as conceptual Project Owner. This layered communication strategy—technical documentation, evaluation metrics, and stakeholder-friendly visuals—supports explainability without compromising the technical rigour of the underlying analyses.

5. Requirements Coverage and Acceptance Criteria

The project requirements Rq1–Rq7 are defined in the Project Work Plan and mapped to specific deliverables, verification methods and due dates. The following paragraphs summarise how the final solution addresses each requirement and how acceptance is or can be demonstrated.

Rq1 – Data consistency & integrity. Rq1 is addressed by the Unified Dataset v1 (1.1.1), supported by the Data Cleaning Scripts (1.1.1.a) and Validation Report (1.1.1.b). Acceptance hinges on the Validation Report confirming the absence of schema and key errors, successful join-key consistency checks and resolution of all logged issues. These conditions are met in the current dataset, which has standardised types, enforced uniqueness of contract-date keys and validated weather merges.

Rq2 – Anomaly labeling / evaluation protocol. Rq2 is primarily fulfilled by the Evaluation Pack (1.2.3), with inputs from the EDA pack v1 (1.1.2) and the initial hypotheses on anomaly patterns. The evaluation protocol defines how proxy labels are derived from k-means and statistical rules, documents the rationale based on EDA, and specifies reproducible evaluation steps. Acceptance requires PSC approval of the protocol document and evidence that sample labels or proxy events are defined and reproducible. While the protocol and associated metrics are in place, the Status Report notes that formal approval of the anomaly policy remains an open change request, which affects confidence in metrics but not the ability to deliver the artefacts.

Rq3 – Model performance (≤ X% FPR, ≥ Y% recall). Rq3 is covered by the Model Baselines (1.2.2) and the Evaluation Pack (1.2.3), particularly the Metrics Report with confidence intervals and the error-analysis component. The Metrics Report computes performance measures per brand and horizon under the current proxy labels, and the error analysis summarises strengths and weaknesses across representative cases. Acceptance is defined as achieving the agreed false-positive and recall thresholds on a pilot set, documented in the metrics with CIs and confirmed in PSC minutes. Given that thresholds X and Y are placeholders in the academic context and that anomaly policy approval is pending, the project interprets Rq3 as demonstrating a plausible operating region and clearly documenting the dependency of metrics on labeling assumptions.

Rq4 – Explainability (stakeholder-readable). Rq4 is satisfied through the Visualization Pack (1.3.2), supported by model-baseline feature analyses and evaluation case studies. The pack includes anomaly and performance plots, threshold trade-off views and example time series with labelled anomalies, accompanied by a concise “How to interpret” document. Acceptance requires a PSC review of this one-pager, inclusion of examples of true and false alarms, and a clarity check with a non-technical reviewer, which are integrated into the planned review process for Iteration 3.

Rq5 – Reproducibility (environment, versioning, runbook). Rq5 is addressed by the Schedule / Monitoring Pack with environment and runbook annex (1.4.2), and is supported by the data-cleaning scripts and the training & tuning runs for the models. The environment specification pins dependencies and versions, while the runbook provides step-by-step instructions to run data preparation, training and evaluation. Acceptance is defined as a fresh clone being able to run the pipeline end-to-end with pinned versions, and as another team member reproducing key outputs without assistance, with times noted. In addition, the Decision & Assumptions Log and

configuration-management practices ensure that decisions and code/document versions remain traceable over time.

Rq6 – Documentation (final report + executive summary). Rq6 is fulfilled by this Final Report (1.3.1) and the Final Presentation (1.3.3), which together provide the consolidated narrative of objectives, data, methods, results, limitations and implications. Acceptance requires PSC sign-off on the report, the existence of a one-page executive summary, and consistency between the written report and the slide deck. These conditions are embedded in the WBS and project-management plan for Iteration 3, and are monitored through project-management deliverables (1.4.x).

Rq7 – Visualization of anomalies & performance. Rq7 is jointly covered by the Visualization Pack (1.3.2) and the Evaluation Pack (1.2.3). The visualisations are designed to show how anomalies are defined and flagged, how thresholds affect precision/recall trade-offs, and how model performance varies across brands and horizons, with plots cross-checked against the metrics report to ensure consistency. Acceptance is defined as plots that clearly explain anomalies, thresholds and trade-offs, matching the reported metrics and validated through PSC review.

6. Assumptions and Limitations

6.1. Data-Related Assumptions and Limitations

The project assumes that the datasets provided by Aigües de Barcelona are representative of typical operations and that the cleaned unified dataset v1 accurately reflects the underlying population of meters and behaviours. Although extensive validation has been performed—checking for schema conformity, duplicate removal, join-key integrity and temporal consistency—residual data-quality issues may remain, particularly in fields with high missingness or legacy coding quirks.

The analysis further assumes that the temporal coverage of the dataset is sufficient to capture relevant seasonal patterns and anomaly behaviour. Statistical analysis confirms weekly and seasonal structures, but some meter combinations, especially rarer models and diameters, may be under-represented, leading to higher uncertainty in their anomaly rates and model performance. Finally, the integration of weather data is limited to daily aggregates, and the weak observed correlation with consumption suggests that environmental effects may be under-captured or require higher-resolution data for precise modeling.

6.2 Modeling and Evaluation Assumptions and Limitations

A central assumption is that anomaly labels derived from statistical rules and k-means distance thresholds are reasonable proxies for true operational anomalies. The project treats these labels as training and evaluation targets for the LSTM models, acknowledging that they do not incorporate human review or cost-based optimisation as a utility company would. The anomaly policy and specific thresholds remain formally unapproved; the Status Report highlights this as an open change request, with assumptions documented and used provisionally in the Evaluation Pack and Decision & Assumptions Log. As a consequence, absolute values of metrics (FPR, recall) should be interpreted with caution, and results are best viewed as scenario-specific rather than definitive performance guarantees.

Modeling assumes that 90-day windows provide a sufficient history to capture patterns that precede anomalies and that per-brand models are adequate to handle heterogeneity across the meter fleet. While supported by statistical evidence on diameter and brand effects, this design may still miss very long-term degradation patterns or interactions across brands and network segments. Furthermore, limited availability of positive anomaly sequences leads to significant class imbalance, which is mitigated but not eliminated by sample weighting and regularisation. This may reduce the models' ability to generalise to rare or novel anomaly types.

The project also assumes that the chosen baselines (statistical rules, k-means + LSTM) are adequate to demonstrate feasibility within the course scope. There was no mandate to exhaustively explore all possible model families (e.g. advanced probabilistic models or graph-based approaches), so future work may find improvements by broadening the modeling toolkit once more robust labels and operational feedback are available.

6.3 Scope and Project-Management Limitations

The project is constrained by the fixed ten-week academic calendar, limited compute resources and intermittent client availability.

These constraints motivated a WBS that prioritises analytical prototypes over production-grade deployment and led to a re-baselining of Iteration 2 around the actual date of full dataset delivery. As a result, out-of-scope items include live system integration, customer-facing dashboards, full MLOps pipelines and real-time alerting; the solution focuses on offline analysis, modeling and visualisation.

Governance limitations also apply: the Project Owner is conceptual in this academic setting, and PSC members emulate PO decisions.

This arrangement ensures methodologically sound oversight but cannot fully replicate the iterative negotiation of anomaly policies, risk tolerances and operational procedures that would occur in a live utility environment. Consequently, some assumptions—such as acceptable false-positive rates, response times and prioritisation of meter segments—are defined by the project team and course methodology rather than by line-of-business stakeholders, and should be revisited before any real-world deployment.

7. Risks, Issues and Challenges

This section summarises the main risks, issues and challenges that have shaped the project and remain relevant for interpreting its outcomes. It follows the structure of the Risk & Issues Register and the Status Reports, focusing on data quality, methodological constraints and project-management aspects.

7.1 Data Gaps and Data Quality Constraints

The project started from early extracts whose schemas and join keys were still stabilising. The top data-related risk (R1) identified in the Status Report points to missing values and inconsistent keys as potential sources of degradation for anomaly detection and credibility of results.

This risk has been mitigated by implementing a systematic data-profiling and cleaning pipeline, maintaining a Data Quality Log, enforcing key uniqueness in the unified dataset and validating the weather merge. Residual risk remains, particularly for fields with high initial missingness or legacy encodings, where business semantics could not be fully verified within the academic setting.

Temporal coverage and representativeness also impose constraints. While the dataset contains sufficient data to identify weekly and seasonal patterns, some meter groups (specific brand–diameter–model combinations) are sparsely represented, which increases uncertainty in their anomaly rates and in the stability of model performance for those sub-populations.

In addition, the integration of daily weather aggregates shows weak linear correlation with consumption, suggesting that any weather-driven effects may be partially masked by temporal aggregation or confounded with seasonality.

7.2 Methodological and Technical Constraints

On the methodological side, the most critical limitation is the lack of an approved anomaly labeling policy. Issue 02 in the Status Report explicitly records that there are no authoritative anomaly labels or agreed business rules, forcing the project to rely on proxy labels and threshold assumptions derived from statistical rules and k-means distance tails. This affects the robustness and interpretability of recall and false-positive metrics, and it is the main reason why the overall project status remains Amber despite modeling and evaluation being technically complete.

A second key constraint is model complexity under data scarcity. Risk R2 highlights the danger of overfitting and poor generalisation due to limited labelled data and the complexity of LSTM-based architectures.

Mitigation measures include per-brand models to reduce heterogeneity, meter-level train/validation splits to avoid leakage, and the use of cross-validation, early stopping, regularisation and comparisons with simpler statistical baselines. Nonetheless, the possibility remains that certain rare or novel anomaly patterns are not well captured, particularly in under-represented meter segments.

Technical constraints also stem from limited computing resources and the need to maintain a reproducible but lightweight environment. The choice of a Python stack with Streamlit-based orchestration balances expressiveness and speed of iteration, but constrains model sizes,

hyperparameter search breadth and the feasibility of more computationally intensive alternatives (for example, large ensembles or complex probabilistic models).

7.3 Project Management and Scheduling Challenges

From a project-management perspective, the main issue has been schedule alignment with dataset availability. Issue 01 in the Status Report documents that the original schedule for Iteration 2 was misaligned with the actual delivery of the full dataset, creating a risk of compressing modeling work into Weeks 7–8.

This issue has been addressed by an approved change request that re-baselined milestones around the real delivery date (28/10) and pulled internal deadlines forward by two to three days per milestone, without reducing scope. The change is reflected in the updated Work Plan and Follow-up Register and is considered closed, with no residual delay on final due dates.

Risk R5 captures the broader schedule pressure due to overlapping academic deadlines, which could have delayed work packages and final deliverables.

Mitigation consists of internal early deadlines, explicit schedule monitoring under WP 1.4.2, and the option to rebalance tasks across the team; these measures, combined with the re-baselining, have kept the project on track in terms of scope and schedule. Finally, governance constraints—most notably the fact that the Project Owner is conceptual and PSC members emulate client decisions—limit the extent to which anomaly policies and acceptance criteria can be negotiated and iterated, which in turn keeps the project at Amber until labeling and evaluation criteria are formally agreed.

8. Value and Next Steps

8.1 Potential Operational Impact for Aigües de Barcelona

The project's primary value proposition lies in demonstrating that smart-meter data, when combined with statistical analysis and machine learning, can support a structured anomaly analytics pipeline for predictive maintenance. The two-layer architecture provides two complementary capabilities: first, an unsupervised k-means layer that produces interpretable anomaly scores per meter and day, and second, a supervised LSTM layer that estimates anomaly risk over the next day, week and month, enabling prioritisation of assets and interventions.

If transferred to an operational setting with validated anomaly policies, this pipeline could help Aigües de Barcelona reduce non-revenue water and service disruptions by identifying meters likely to fail or misbehave before issues become critical. For example, brands and diameters with systematically higher anomaly rates, as identified in the statistical analysis, could be placed under closer monitoring or proactive maintenance programmes. The combination of anomaly scores, multi-horizon risk estimates and stakeholder-friendly visualizations would allow operations teams to balance sensitivity and workload by tuning thresholds and focusing on high-risk segments where interventions yield the greatest expected benefit.

From a broader perspective, the project also generates organisational value by consolidating a single validated dataset, a reusable cleaning pipeline, a documented evaluation protocol and a set of configuration-managed artefacts (runbook, environment specification, decision log). These assets reduce onboarding time for future data-science work on smart meters and provide a template for similar analytical projects in other domains, aligning with the Work Plan's success criteria around data readiness, technical readiness and value narrative.

8.2 Recommendations and Future Improvements

Several steps are recommended to move from this academic prototype to a production-ready anomaly analytics system. First, Aigües de Barcelona should work with domain experts and operations staff to define and approve a formal anomaly policy. This policy would specify what constitutes an actionable anomaly (for example, magnitude and duration thresholds, context-specific rules) and how false positives and false negatives are valued in terms of operational cost. Once agreed, this policy can be used to refine the labeling strategy, recalibrate k-means thresholds and adjust LSTM classification thresholds, leading to more meaningful performance metrics and decision rules.

Second, the models should be validated on additional datasets and, where possible, on cases with known operational outcomes (for example, meters that have been inspected or replaced). This would help assess generalisation beyond the current sample and quantify performance in real-world conditions, addressing the overfitting risk identified in R2. As data accumulates, alternative model families (such as gradient-boosted trees with time-aware features or probabilistic models tailored to count data) could be explored to complement or replace the LSTM, using the current baselines as benchmarks.

Third, the technical stack should be extended towards operationalisation. This includes hardening the pipeline to run on scheduled jobs, integrating with existing monitoring and ticketing systems, and possibly migrating from a notebook- and Streamlit-based prototype to containerised services orchestrated within Aigües de Barcelona's infrastructure. Configuration management and reproducibility practices defined in WP 1.4.2 can be reused and extended to cover deployment environments, thereby maintaining the traceability achieved in this project. Finally, continued collaboration between data scientists, operations and management, following the governance patterns outlined in the Work Plan, will be essential to ensure that model outputs are interpreted correctly and translated into effective maintenance strategies.