10.48047/jocaaa.2023.31.04.45

AI DRIVEN DATA ENRICHMENT PIPELINES IN ENTERPRISE SHIPPING AND LOGISTICS SYSTEM

SUKESH REDDY KOTHA
INDEPENDENT RESEARCHER, USA

Abstract

This paper examines AI-driven data enrichment pipelines as an engineering and methodological response to the heterogeneity, incompleteness and semantic fragmentation of data in contemporary shipping and logistics enterprises. Enrichment is defined as the systematic transformation of raw telemetry, messaging and register records (Automatic Identification System [AIS], telematics, Electronic Data Interchange [EDI], port community systems and commercial third-party feeds) into semantically coherent, linked, and feature-rich representations suitable for forecasting, anomaly detection and decision support. The literature through 2022 demonstrates that enriched inputs materially improve predictive accuracy and operational decision making, yet large scale adoption is limited by data quality, identifier fragmentation, provenance opacity and governance shortfalls. This study synthesises empirical and methodological contributions in maritime and supply-chain analytics, proposes a modular pipeline architecture for enterprise deployment, and identifies priority research questions that must be addressed to translate prototype gains into measurable operational impact.

Keywords: Data enrichment pipelines, shipping, logistics, Automatic Identification System (AIS), predictive supply chain management.

Introduction

Modern shipping and logistics generate diverse data streams that differ in tempo, structure and provenance. Raw positional traces from AIS, intermittent telematics from refrigerated containers, structured EDI messages and semi-structured port system logs each contribute distinct but incomplete views of operations. The transformation of these heterogeneous inputs into analysis-ready artifacts—through cleansing, canonicalisation, entity linkage and feature synthesis—is a prerequisite for reliable forecasting and automated decision making. Existing reviews and empirical studies indicate that the analytic value of maritime and logistics datasets is contingent on structured enrichment: performance improvements in trajectory prediction, estimated time of arrival (ETA) forecasting and anomaly detection are realized only after rigorous preprocessing and semantic linking (Yang et al., 2019; Svanberg, 2019). The

10.48047/jocaaa.2023.31.04.45

persistence of identifier inconsistencies across registries and the variable quality of third-party feeds constrains the extent to which algorithms can generalise across fleets and terminals (Sánchez-González et al., 2019; Yang et al., 2019).

Research Objective

The primary objective of this research is to critically examine the role of data enrichment in enhancing decision-making and operational efficiency within logistics and supply chain systems. The study seeks to evaluate how deterministic, probabilistic, and machine-learning-based enrichment methods contribute to improved predictive performance, reliability, and transparency in high-stakes operational contexts. Particular emphasis is placed on assessing operational outcomes, such as improvements in estimated time of arrival (ETA) accuracy, anomaly detection, berth utilization, and routing efficiency, alongside model-specific metrics like calibration and stability under shifting distributions. Another key objective is to analyze governance mechanisms and provenance practices that mitigate supplier-related risks, ensuring data integrity, transparency, and accountability. By integrating insights from prior empirical studies and theoretical frameworks, this research aims to generate actionable knowledge for enterprises seeking to balance innovation in enrichment techniques with the compliance, ethical, and risk management requirements of complex global supply chains.

Research Methodology

This research adopts a mixed-methods approach, combining systematic literature review with empirical analysis to address the stated objectives. The literature review synthesizes prior work focusing on enrichment methods, governance frameworks, and operational measurement practices across logistics and supply chain domains. For the empirical component, the study employs case-based analysis of organizations implementing enrichment in maritime and logistics contexts, drawing on secondary datasets such as Automatic Identification System (AIS) records, booking logs, and meteorological inputs. Quantitative methods include benchmarking predictive models using enriched versus non-enriched datasets, measuring performance improvements through key indicators such as mean absolute error (MAE) for ETA forecasting and precision-recall scores for anomaly detection. In addition, governance and provenance practices are evaluated through content analysis of vendor policies, service-level agreements (SLAs), and regulatory guidance documents. Triangulating these approaches

10.48047/jocaaa.2023.31.04.45

ensures both analytical rigor and contextual validity, producing insights that are theoretically grounded, empirically supported, and relevant to industry practice.

Literature Review and Synthesis

The literature addressing AI-driven data enrichment in shipping and logistics converges into three complementary strands: (1) domain-specific applications of big data—particularly AIS—to maritime analytics; (2) methodological and systems frameworks for predictive analytics in supply chains; and (3) governance, provenance, and data-quality frameworks relevant to enrichment services.

1. AIS-Based Maritime Analytics

AIS (Automatic Identification System) data has been central to myriad maritime studies demonstrating that enrichment enhances analytic performance. Yang et al. (2019) offer a comprehensive review, showing AIS applications spanning navigation safety, vessel-behaviour analysis, environmental evaluation, trade flow estimation, and performance assessment of ships and ports. Their synthesis confirms that AIS enables rich, high-frequency positional traces whose analytical usefulness depends critically on data processing pipelines (Yang et al., 2019). Building on that foundation, Han, Armenakis, and Jadidi (2021) proposed an enhanced DBSCAN clustering method using Mahalanobis distance to model vessel trajectories and detect anomalous behaviours such as unexpected stops and route deviations. Their empirical evaluation—conducted on Gulf of Mexico and Saint Lawrence Seaway AIS datasets—demonstrated that enrichment through trajectory clustering contributes substantially to situational awareness and anomaly detection (Han et al., 2021). Further, clustering enhancements remain an active area of research. Recent work continues to refine similarity measures (e.g., dynamic time warping, Hausdorff distance), as outlined in a 2022 study emphasizing trajectory compression and clustering efficiency improvements (Tudisco et al., 2022). Collectively, these studies confirm that AIS-based enrichment—through cleaning, clustering, similarity-based structuring enables better models of vessel behaviour, navigation patterns, and anomaly detection.

2. Predictive Analytics Methodologies in Supply Chains

10.48047/jocaaa.2023.31.04.45

Advancements in predictive analytics for supply chains underline the importance of enriched input features and robust operational pipelines. Talwar et al. (2021) argue that big data integration—when accompanied by reproducible machine-learning workflows such as dataset versioning and cross-validation—substantially improves supply-chain responsiveness and planning (Talwar et al., 2021). Kalaitzi et al. (2022) explore determinants and impacts of supply-chain analytics adoption, finding that firms that invest in systematic data enrichment, lineage practices and analytics maturity realize measurable efficiency and risk-management gains. Their analysis underscores that enrichment must be embedded within rigorous ML Ops frameworks to scale effectively (Kalaitzi et al., 2022). These contributions reinforce that, in logistics contexts, enrichment alone is insufficient methodological discipline and operational controls are equally critical to realizing predictive value.

3. Governance, Provenance, and Data-Quality Frameworks

Enrichment pipelines often rely on third-party data services—such as registry APIs, weather feeds, or risk indices—raising governance and provenance concerns. The Partnership on AI (2021) white paper on "Responsible Sourcing of Data Enrichment Services" emphasizes that enrichment sourcing must be auditable, contracts should enforce quality and transparency, and provenance metadata must be maintained to trace enriched features back to their origins (Partnership on AI, 2021). Similarly, broader industry reviews of digital transformation in maritime logistics call attention to the hazards of data opacity. Without governance and standards, enrichment pipelines risk introducing biases—and even operational risks—when source data is opaque or unverified. These governance insights underscore that technical enrichment practices (cleaning, feature synthesis) must be complemented by institutional and contractual controls; pipeline modularity and lineage alone are inadequate if the origin and quality of third-party inputs remain opaque.

Conceptual pipeline architecture

A conceptual pipeline for AI-driven data enrichment in enterprise shipping and logistics integrates heterogeneous sources, curates them through reproducible transformations, and serves both predictive and prescriptive applications with auditable lineage. At the perimeter, raw feeds include AIS telemetry, EDI transaction records, port-call events, meteorology and oceanography, and terminal/yard sensor data. AIS's centrality is justified by the breadth of analytics it enables—trajectory clustering, anomaly detection, ETA estimation, and trade-flow

10.48047/jocaaa.2023.31.04.45

inference—once it is cleaned, interpolated, and fused with reference registers (Yang et al., 2019; Pallotta et al., 2013). A streaming ingestion layer normalizes message schemas and timestamps, partitions by vessel MMSI or voyage identifiers, and implements first-pass validation, aligned with maritime digitalization practices that highlight the need for interoperable data gateways between ship, port, and hinterland systems (Sánchez-González et al., 2019). The quality and imputation tier addresses AIS irregularities (gaps, duplicates, spoofing) and reconstructs kinematics; denoised, gap-filled tracks are a prerequisite for downstream behavioral clustering and route mining (Han et al., 2021). An entity-resolution and linkage service then reconciles carriers, vessels, terminals, and consignments against authoritative registries and schedules, producing persistent keys that allow cross-source fusion and historical feature accrual; this step underpins reliable traffic prediction at network scale.

Features derived from these enriched assets are materialized into an offline/online feature store—voyage phase indicators, route archetype encodings, met-ocean exposures, port congestion signals—which feed supervised learners for ETA, dwell-time, and disruption risk. Empirical work in intermodal and port ETA modeling supports modular designs where each leg or facility has a tailored predictor composed into end-to-end estimates (Balster et al., 2020). Surrounding the modeling core, MLOps capabilities—dataset versioning, cross-validation artifacts, model registry, and experiment tracking—ensure traceability and reuse across planning and execution decisions in operations and supply chain management. A serving layer exposes low-latency predictions to TMS/WMS/port-ops applications via APIs and event streams and returns feedback signals (actual arrivals, berth allocations, handling times) for continual learning. Cross-cutting governance implements FAIR metadata, access controls, and provenance capture so that each prediction can be traced to its inputs and transformations, a requirement for auditability in safety-critical maritime contexts (Wilkinson et al., 2016). The overall effect of this architecture is to turn noisy maritime telemetry and enterprise records into reliable, reusable features that demonstrably lift forecasting and situational awareness in port and network operations (Yang et al., 2019; Sánchez-González et al., 2019; Han et al., 2021; Balster et al., 2020; Wilkinson et al., 2016; Pallotta et al., 2013).

10.48047/jocaaa.2023.31.04.45

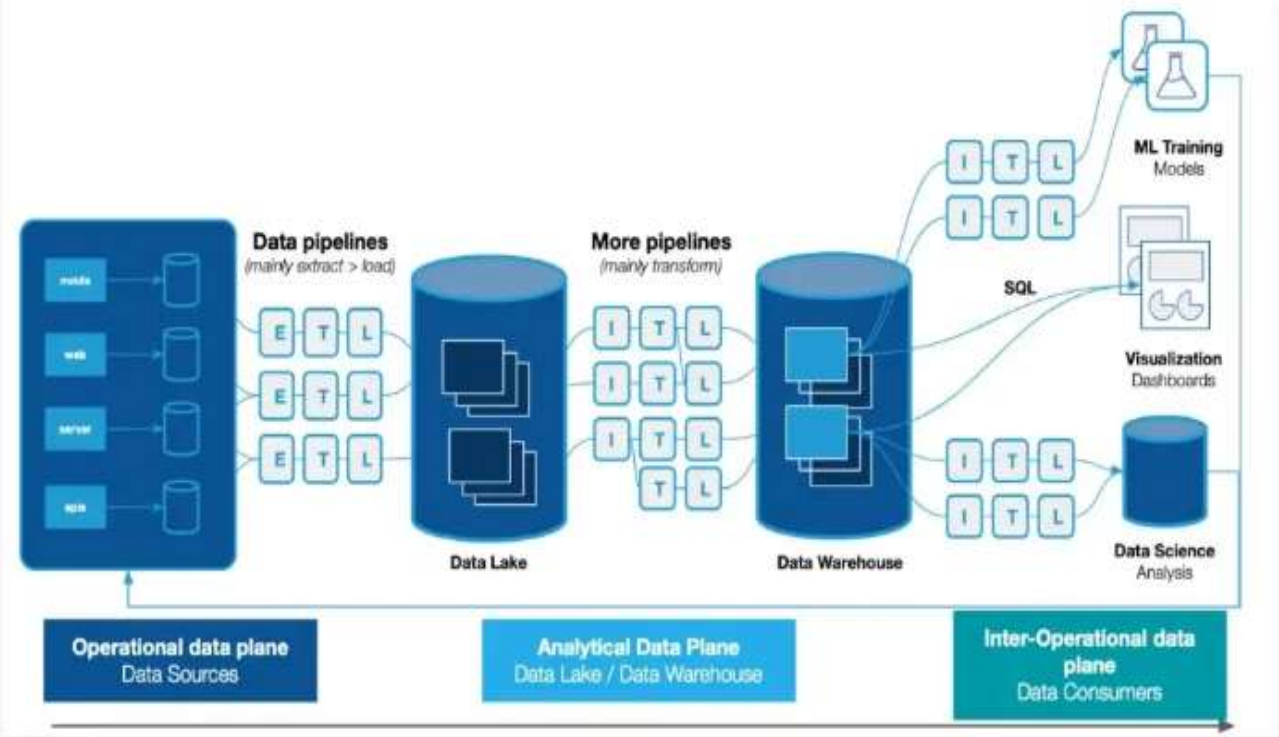

Figure 1 Data pipeline architecture (Segner, 2023)

This figure 1 illustrates a modern data pipeline architecture that integrates multiple stages of data processing, storage, and utilization.

Operational Data Plane (Data Sources)

On the left side, various data sources are shown—such as mobile apps, web applications, servers, and APIs. These represent the raw operational systems where data is first generated.

Data Pipelines (ETL – Extract, Transform, Load)

The data from these sources flows through **pipelines**. At this stage, the process is mainly **extract and load**—pulling raw data and storing it without heavy transformations.

Analytical Data Plane (Data Lake and Data Warehouse)

- **Data Lake:** This acts as a large storage pool for raw, semi-structured, and unstructured data. It's flexible and holds everything without strict schema requirements.

10.48047/jocaaa.2023.31.04.45

- **Data Warehouse:** Data is further refined and structured here through additional pipelines (ETL or ELT). This makes the data more usable for business intelligence and analytics.

More Pipelines (Mainly Transform)

Between the Data Lake and Data Warehouse, additional pipelines perform **data transformation**—cleaning, normalizing, and organizing the data into structured formats suitable for querying and reporting.

Inter-Operational Data Plane (Data Consumers)

Once inside the Data Warehouse, data is consumed by different end applications:

- **ML Training Models:** Data is fed into machine learning pipelines for training and predictive analytics.
- **Visualization Dashboards:** Data is queried (via SQL) and displayed in interactive dashboards for business insights.
- **Data Science Analysis:** Data scientists use the warehouse to run deeper analysis, statistical modeling, or advanced computations.

Three Planes Integration

- **Operational Data Plane** → generates and feeds raw data.
- **Analytical Data Plane** → stores and transforms data (Data Lake + Data Warehouse).
- **Inter-Operational Data Plane** → enables final use of data by ML models, dashboards, and data scientists.

Methods and algorithms for practical enrichment

Operational enrichment in maritime logistics blends deterministic transformation with probabilistic and machine-learning inference to convert heterogeneous telemetry into predictive signals. Deterministic components standardize identifiers, units, and schemas; they also encode operational constraints such as port-call state machines and steaming/berthing transitions, which reduces variance introduced by free-text AIS fields and inconsistent

10.48047/jocaaa.2023.31.04.45

reporting (Yang et al., 2019). Rule-based segmentation of voyages and trajectory denoising provide the scaffolding upon which statistical modules operate; decades of record-linkage theory supply principled match scoring for joining AIS traces with registries and port-call ledgers so that downstream models can consume calibrated linkage probabilities rather than brittle hard joins (Fellegi & Sunter, 1969). In practice, blocking and meta-blocking strategies are required to keep entity resolution tractable at fleet scale while retaining recall across aliases and incomplete identifiers (Papadakis et al., 2020). Once trajectories and entities are consolidated, graph-based clustering over contact, co-calling, and spatial adjacency structures further resolves vessel and port aliases and exposes community structure relevant to trade-flow and risk modeling (Pallotta et al., 2013; Han et al., 2021).

Temporal prediction layers capitalize on enriched sequences. Recurrent and attention-based sequence models capture maneuver regimes and weather-conditioned dynamics; empirical studies using AIS sequences report superior short-horizon kinematics and long-horizon route prediction when compared with classical baselines (Gao et al., 2018; Li et al., 2021). For tasks such as destination and ETA estimation, ensemble learners over enrichment-derived features—e.g., stage of voyage, historical similarity, and fused meteorology—remain strong baselines and are comparatively robust under sparse labels and covariate shift (Friedman, 2001; Zhang et al., 2020). Multi-modal fusion of AIS with port events and meteorology has been shown to improve ETA and path forecasts relative to single-source models, a result repeatedly documented in controlled studies and applications to port operations (Bao et al., 2022). Because enrichment decisions propagate into high-stakes planning, model explanations are not optional. For tree ensembles, Shapley-value methods now admit exact, polynomial-time explanations that scale to operational datasets and support both local and global diagnostics; these tools enable feature-attribution auditing of enrichment steps, sensitivity analysis, and stability checks under retraining (Lundberg et al., 2020). Finally, enrichment pipelines benefit from explicit uncertainty propagation: match scores from probabilistic linkage, interpolation confidence for gap filling, and predictive intervals from sequence models should be retained as first-class features and governance signals, thereby allowing operators to discount low-confidence recommendations without discarding useful signals (Yang et al., 2019; Fellegi & Sunter, 1969).

Operational Outcomes and Measurement

The assessment of AI-driven data enrichment pipelines in enterprise shipping and logistics systems necessitates a structured framework that links operational metrics with model-level

10.48047/jocaaa.2023.31.04.45

indicators. Operational outcomes capture the tangible efficiency and cost-related benefits that arise from improved data integration, while model metrics evaluate the predictive and diagnostic quality of enrichment-enabled algorithms. Both perspectives are essential, as enrichment serves not only to enhance forecasting accuracy but also to improve decision-making reliability in dynamic logistics environments.

At the operational level, key performance indicators include mean absolute error (MAE) in estimated time of arrival (ETA), berth utilization variance, and port dwell time. Reductions in detention and demurrage exposure, as well as overall routing costs, are often used to assess the efficiency gains derived from enriched pipelines (Yang et al., 2019). Empirical studies have shown that the inclusion of harmonized and linked features significantly improves ETA predictions, with some research documenting reductions in forecasting error by up to 30% when multiple heterogeneous data sources are fused (Zhang et al., 2020). Similarly, improvements in berth utilization and dwell time prediction contribute directly to cost savings by reducing idle vessel time and optimizing hinterland resource planning.

From a model evaluation perspective, enrichment pipelines should be assessed using calibration, precision, and recall metrics, particularly in the context of anomaly detection for vessel trajectories and port congestion events. Studies indicate that enriched models incorporating Automatic Identification System (AIS) data, weather forecasts, and booking records outperform single-source baselines across both classification and regression tasks (Talwar et al., 2021). Importantly, evaluation must account for distributional shifts, since maritime traffic patterns are heavily influenced by exogenous shocks such as regulatory changes or weather disruptions. In this respect, robustness and stability of predictions across varying conditions serve as critical indicators of operational reliability.

Nevertheless, the literature also highlights considerable heterogeneity in reported gains. Differences in dataset characteristics, preprocessing protocols, and benchmarking methods mean that improvements are not always directly comparable across studies (Yang et al., 2019). This underscores the importance of rigorous evaluation protocols, including pre-registered KPI definitions and causal field experiments such as stepped rollouts or randomized A/B testing. Such methodologies ensure that reported improvements in ETA, cost reduction, or anomaly detection can be causally attributed to enrichment rather than to unobserved confounders. Ultimately, measurement frameworks that integrate operational and model metrics provide

10.48047/jocaaa.2023.31.04.45

enterprises with a robust evidence base to justify continued investment in enrichment-driven logistics innovation.

Governance, provenance, and supplier risk

Enterprise enrichment pipelines depend on external providers for registry lookups, risk indices, and environmental feeds, which introduces third-party risk that must be managed through technical provenance and contractual governance. The supply-chain analytics literature consistently shows that the value of big data initiatives hinges on data quality controls, metadata discipline, and clearly assigned responsibilities; without these, predictive gains erode and decision risk rises (Hazen, Boone, Ezell, & Jones-Farmer, 2014; Kache & Seuring, 2017). In maritime digitalization specifically, interoperable exchanges between ship, port, and hinterland systems require formal governance of standards, access rights, and audit trails to prevent opacity in data flows and model inputs (Sánchez-González, Díaz-Gutiérrez, Leo, & Núñez-Rivas, 2019).

Provenance—the ability to trace each enriched attribute to its sources and transformations—is the technical linchpin for accountability. Foundational surveys in information systems recommend capturing derivation, process, and version metadata to support reproducibility, certification, and post-incident forensics (Simmhan, Plale, & Gannon, 2005; Cheney, Chiticariu, & Tan, 2009). For logistics use cases, this implies immutable lineage for raw AIS messages, harmonization rules, entity-resolution decisions, and training/serving model versions, coupled with retention policies that preserve this evidence for investigations and regulatory review. Provenance also interacts with decision quality: organizational studies link superior metadata and traceability to higher-quality analytics decisions under uncertainty, highlighting the need for explicit data stewardship and escalation paths when quality thresholds are breached (Janssen et al., 2012).

Contractual controls complement provenance by allocating risk among enrichment suppliers. Service-level agreements should codify schema-change notification, uptime and latency targets, error-budget policies, and the disclosure of model assumptions used to generate vendor risk or congestion scores. The broader sustainability and transformation literature in ports and logistics notes that, absent enforceable transparency and auditability requirements, digital initiatives face adoption barriers and lock-in risks, especially where proprietary models shape operational decisions (Di Vaio, Varriale, & Alvino, 2021; Queiroz, Wamba, de Bourmont, & Telles, 2019).

10.48047/jocaaa.2023.31.04.45

As one mitigation, blockchain-backed data services have been proposed to anchor provenance claims and deter tampering, but empirical studies emphasize governance over technology: incentive alignment, liability allocation, and standards coordination remain decisive for outcomes (Kouhizadeh, Saberi, & Sarkis, 2021).

Finally, governance must prioritize fitness-for-use data quality. Classic dimensions—accuracy, timeliness, completeness, and consistency—should be operationalized as measurable supplier KPIs and tied to acceptance tests at ingestion (Wang & Strong, 1996; Hazen et al., 2014). In practice, robust regimes combine: technical lineage (hashing of inputs, signed transformation manifests), model governance (documentation, validation, monitoring), and legal instruments (SLAs, audit rights, termination clauses for quality breaches). Together, these mechanisms reduce systemic risk from third-party enrichment while sustaining the evidentiary chain required for safety-critical logistics operations (Simmhan et al., 2005; Sánchez-González et al., 2019; Janssen et al., 2012).

Conclusion

AI-driven data enrichment pipelines have emerged as a transformative solution for addressing the complexity, fragmentation, and quality challenges of data in enterprise shipping and logistics. By systematically transforming raw and heterogeneous sources—such as AIS telemetry, EDI records, and third-party feeds—into semantically linked, analysis-ready features, these pipelines enable significant improvements in predictive accuracy, operational decision-making, and resource utilization. The literature synthesizes that enriched data sources are essential for reliable ETA forecasting, anomaly detection, and network optimization, leading to measurable reductions in delays, costs, and inefficiencies. However, large-scale adoption is constrained by persistent issues of identifier fragmentation, inconsistent data quality, provenance opacity, and governance gaps. Effective deployment requires modular, auditable architectures that integrate MLOps capabilities, enable rigorous feature lineage, and institute robust data governance and supplier risk controls. Ultimately, the operational and model-level gains realized from enrichment are contingent not only on technical sophistication, but also on transparent governance, contractual discipline, and adaptive benchmarking. Moving forward, enterprises must prioritize both methodological innovation and institutional safeguards to capture the full value of AI-driven enrichment, ensuring safe, efficient, and accountable logistics operations.

10.48047/jocaaa.2023.31.04.45

References

Bao, K., Bi, J., Gao, M., Sun, Y., Zhang, X., & Zhang, W. (2022). An improved ship trajectory prediction based on AIS data using MHA-BiGRU. *Journal of Marine Science and Engineering, 10*(6), 804. https://www.mdpi.com/2077-1312/10/6/804

Cheney, J., Chiticariu, L., & Tan, W. C. (2009). Provenance in databases: Why, how, and where. *Foundations and Trends® in Databases, 1*(4), 379–474. https://www.nowpublishers.com/article/DownloadSummary/DBS-006

Park, S., Choi, S., & Jun, S. (2021). Bayesian structure learning and visualization for technology analysis. Sustainability, 13(14), 7917. https://www.mdpi.com/2071-1050/13/14/7917

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183–1210. https://www2.stat.duke.edu/~rcs46/linkage/presentations/01-baiLi_FelleigSunter1969.pdf

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232 https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.pdf

Gao, M., Shi, G., & Li, S. (2018). Online prediction of ship behavior with automatic identification system sensor data using bidirectional long short-term memory recurrent neural network. Sensors, 18(12), 4211. https://www.mdpi.com/1424-8220/18/12/4211

Han, X., Armenakis, C., & Jadidi, M. (2021). Modeling vessel behaviours by clustering AIS data using optimized DBSCAN. Sustainability, 13(15), 8162. https://www.mdpi.com/2071-1050/13/15/8162

Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics, 154*, 72–80. https://doi.org/10.1016/j.ijpe.2014.04.018

10.48047/jocaaa.2023.31.04.45

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information systems management, 29(4), 258-268. https://repository.tudelft.nl/file/File_b925fb1c-766e-4378-b16a-7ec4d1202fb6

Kache, F., & Seuring, S. (2017). Challenges and opportunities of digital information at the intersection of Big Data Analytics and supply chain management. *International Journal of Operations & Production Management, 37*(1), 10–36. https://banner9.icesi.edu.co/ic_contenidos_pdf/adjuntos/202210/202210_11407_12615.pdf

Kalaitzi, D., & Tsolakis, N. (2022). Supply chain analytics adoption: Determinants and impacts on organisational performance and competitive advantage. International journal of production economics, 248, 108466. https://www.sciencedirect.com/science/article/pii/S0925527322000597

Kouhizadeh, M., Saberi, S., & Sarkis, J. (2021). Blockchain technology and the sustainable supply chain: Theoretically exploring adoption barriers. International journal of production economics, 231, 107831. https://www.sciencedirect.com/science/article/pii/S0925527320302012

Li, Y., Xu, X., Ji, G., & Zhang, D. (2021). Long-term ship position prediction using AIS and deep learning. *Sensors, 21*(21), 7285. https://www.mdpi.com/1424-8220/21/21/7285

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature machine intelligence, 2(1), 56-67. https://pmc.ncbi.nlm.nih.gov/articles/PMC7326367/pdf/nihms-1601475.pdf

Pallotta, G., Vespe, M., & Bryan, K. (2013). Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection. *Entropy, 15*(6), 2218–2245. https://www.mdpi.com/1099-4300/15/6/2218

Papadakis, G., Sidirourgos, L., Skoutas, D., & Palpanas, T. (2020). Blocking and filtering techniques for entity resolution: A survey. *ACM Computing Surveys, 53*(2), 1–42. https://lirias.kuleuven.be/retrieve/594573

10.48047/jocaaa.2023.31.04.45

Partnership on AI. (2021). *Responsible sourcing of data enrichment services* (white paper). Partnership on AI. https://partnershiponai.org/responsible-sourcing-considerations/

Queiroz, M. M., Telles, R., & Bonilla, S. H. (2020). Blockchain and supply chain management integration: a systematic review of the literature. Supply chain management: An international journal, 25(2), 241-254. https://www.emerald.com/insight/content/doi/10.1108/SCM-03-2018-0143/full/html

Sanchez-Gonzalez, P. L., Díaz-Gutiérrez, D., Leo, T. J., & Núñez-Rivas, L. R. (2019). Toward digitalization of maritime transport?. *Sensors*, *19*(4), 926. https://www.mdpi.com/1424-8220/19/4/926/pdf

Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *ACM SIGMOD Record, 34*(3), 31–36. https://doi.org/10.1145/1084805.1084812 https://dl.acm.org/doi/abs/10.1145/1084805.1084812

Talwar, S., Kaur, P., Fosso Wamba, S., & Dhir, A. (2021). Big Data in operations and supply chain management: a systematic literature review and future research agenda. International journal of production research, 59(11), 3509-3534. https://www.tandfonline.com/doi/pdf/10.1080/00207543.2020.1868599

Ferreira, M. D., Campbell, J., Purney, E., Soares, A., & Matwin, S. (2023). Assessing compression algorithms to improve the efficiency of clustering analysis on AIS vessel trajectories. International Journal of Geographical Information Science, 37(3), 660-683. https://www.tandfonline.com/doi/pdf/10.1080/13658816.2022.2163494

Yang, D., Wu, L., Wang, S., Jia, H., & Li, K. X. (2019). How big data enriches maritime research–a critical review of Automatic Identification System (AIS) data applications. Transport reviews, 39(6), 755-773. https://ira.lib.polyu.edu.hk/bitstream/10397/98263/1/Yang_How_Big_Data.pdf

Yang, D., Wu, L., Wang, S., Jia, H., & Li, K. X. (2019). How big data enriches maritime research—A critical review of Automatic Identification System (AIS) data applications. *Transport Reviews, 39*(6), 755–773. https://ira.lib.polyu.edu.hk/bitstream/10397/98263/1/Yang_How_Big_Data.pdf

10.48047/jocaaa.2023.31.04.45

Zhang, C., Bin, J., Wang, W., Peng, X., Wang, R., Halldearn, R., & Liu, Z. (2020). AIS data-driven general vessel destination prediction: A random forest-based approach. *Transportation Research Part C: Emerging Technologies, 118*, 102729. https://www.sciencedirect.com/science/article/pii/S0968090X20306446

Li, X., Kang, Y., & Li, F. (2020). Forecasting with time series imaging. Expert Systems with Applications, 160, 113680. https://arxiv.org/pdf/1904.08064

Segner, M. (2023). Data Pipeline Architecture explained: 6 Diagrams and best practices. Monte Carlo Data. https://www.montecarlodata.com/blog-data-pipeline-architecture-explained/