

DataWeave: An Intelligent Architecture for Context-Aware Predictive Learning

Dharaniya Raja*

Citation: Raja D. Data Weave: An Intelligent Architecture for Context-Aware Predictive Learning. *J Artif Intell Mach Learn & Data Sci* 2026 9(4), 3534-3539. DOI: doi.org/10.51219/JAIMLD/dharaniya-raja/694

Received: 09 September, 2026; **Accepted:** 02 October, 2026; **Published:** 05 October, 2026

*Corresponding author: Dharaniya Raja, India

Copyright: © 2026 Raja D., This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

ABSTRACT

The rapid growth of digital technologies has resulted in the generation of large volumes of structured and unstructured data from diverse sources, including sensors, mobile applications, healthcare systems, business platforms, social networks and Internet of Things (IoT) devices. Extracting meaningful information from such heterogeneous data has become an important requirement for intelligent decision-making and predictive analytics^{1,2}. This paper proposes DataWeave, an intelligent architecture for context-aware predictive learning that integrates data acquisition, preprocessing, feature engineering, contextual integration, machine learning, model evaluation, model selection and prediction within a unified framework. The architecture incorporates contextual characteristics such as temporal, spatial, environmental, behavioural and operational information into the predictive-learning process^{3,4}. DataWeave provides a flexible pipeline through which different machine-learning algorithms can be trained, compared, evaluated and selected for prediction. The proposed framework is designed to improve the representation of real-world observations by combining conventional predictive features with contextual information. The architecture can potentially be applied to healthcare, agriculture, finance, education, transportation, environmental monitoring, business analytics and IoT-based applications. Experimental evaluation using suitable benchmark or real-world datasets is required to quantitatively determine the contribution of contextual integration to predictive performance.

Keywords: DataWeave, Artificial Intelligence, Machine learning, Predictive analytics, Context-aware learning, Feature engineering, Intelligent systems, Data science, Contextual integration

1. Introduction

The increasing adoption of digital technologies has resulted in the continuous generation of large quantities of data from sensors, mobile applications, cloud platforms, healthcare systems, financial services, social networks and connected devices. These data sources provide significant opportunities for discovering hidden patterns and supporting data-driven decision-making. Data mining and knowledge-discovery techniques provide systematic approaches for transforming raw data into useful knowledge^{1,5}.

Machine-learning techniques have become an important component of modern intelligent systems because they can learn relationships from historical observations and use learned patterns to make predictions for new observations^{6,7}. However, the development of reliable predictive systems requires more than selecting an appropriate algorithm. Data quality, preprocessing, feature representation, feature engineering, model selection and evaluation all influence predictive performance^{8,9}.

In conventional predictive-learning systems, observations are frequently represented using a fixed collection of attributes.

However, real-world observations are often affected by the circumstances in which they are generated. Temporal information, geographical location, environmental conditions, user behavior and operational states may influence the relationship between input variables and prediction targets^{3,4}.

Context-aware computing provides a mechanism for incorporating information about the circumstances surrounding an observation or user³. Context-aware predictive learning extends this concept by incorporating contextual information into the machine-learning process.

Existing approaches such as Knowledge Discovery in Databases (KDD) and CRISP-DM provide structured workflows for data analysis and machine-learning projects^{1,2}. However, there remains an opportunity to develop a unified architecture that connects heterogeneous data acquisition, preprocessing, feature engineering, contextual integration, machine learning, model evaluation, model selection and prediction.

To address this requirement, this paper proposes DataWeave, an intelligent architecture for context-aware predictive learning. The architecture is designed to weave multiple dimensions of information into a unified predictive-learning pipeline.

1.1. Main contributions

The major contributions of this work are:

- A unified architecture for context-aware predictive learning.
- Integration of contextual information with conventional predictive features.
- A modular pipeline supporting multiple machine-learning algorithms.
- Integration of model evaluation and model selection.
- A framework that can be adapted to different application domains.
- A basis for experimentally investigating the contribution of contextual information to predictive performance.

2. Related Work

2.1. Machine learning and predictive analytics

Machine learning provides computational methods for learning patterns from historical data and applying learned relationships to new observations^{6,7}. Classification, regression, clustering and ensemble-learning methods are widely used in predictive analytics.

Scikit-learn provides a comprehensive collection of machine-learning algorithms and tools for preprocessing, model selection and evaluation⁸.

2.2. Data mining and knowledge discovery

The KDD framework provides a structured process for transforming raw data into useful knowledge. It includes stages such as data selection, preprocessing, transformation, data mining and interpretation¹.

CRISP-DM provides another structured approach consisting of business understanding, data understanding, data preparation, modeling, evaluation and deployment².

These approaches demonstrate that predictive analytics is an end-to-end process rather than simply the selection of a machine-learning algorithm.

2.3. Context-aware computing

Context-aware systems consider information surrounding an observation or user, including time, location, environmental conditions, behavioral information and operational states³.

Context-aware systems can use contextual information to improve the relevance and interpretation of intelligent-system outputs⁴.

2.4. Ensemble machine learning

Algorithms such as Random Forest, XGBoost and LightGBM are widely used for predictive tasks because they can model complex relationships within data¹⁰⁻¹².

Random Forest combines multiple decision trees to produce an ensemble prediction¹⁰. XGBoost provides a scalable tree-boosting framework¹¹, while LightGBM is designed for efficient gradient-boosting-based learning¹².

DataWeave allows different algorithms to be evaluated within the same pipeline rather than depending on a single predefined model.

Explainable artificial intelligence

Explainable AI aims to make machine-learning predictions easier to understand¹³. SHAP provides a unified framework for interpreting the contribution of individual features to model predictions¹⁴.

Explainability can therefore be incorporated into DataWeave as an additional mechanism for understanding model behavior.

2.6. Research gap

Existing methodologies provide important capabilities for data preparation, predictive modeling, context-aware computing and explainability^{1-4,13,14}. However, the proposed DataWeave architecture focuses on connecting these components into a unified context-aware predictive-learning workflow.

3. Problem Statement

Real-world datasets commonly contain:

- Missing values
- Noisy observations
- Duplicate records
- Inconsistent formats
- Irrelevant variables
- Heterogeneous data
- Contextual information

Traditional predictive systems may not adequately represent the circumstances under which individual observations are generated.

For example, the same measurement may have different meanings depending on:

- Time
- Location
- Environmental conditions
- User behavior
- Operational state

Context-aware computing emphasizes the importance of identifying and using contextual information associated with an observation^{3,4}.

Therefore, the research problem addressed in this study is:

How can heterogeneous data and contextual information be systematically integrated into a unified machine-learning architecture to support reliable and context-aware predictive learning?

4. Proposed Methodology

The proposed DataWeave methodology consists of nine major stages:

Data Sources → Data Acquisition → Preprocessing → Feature Engineering → Contextual

Integration → Machine Learning → Model Evaluation → Model Selection → Prediction

4.1. Data sources

Data may originate from public datasets, databases, sensors, mobile applications, IoT devices, application logs, enterprise systems and other relevant sources.

The integration of heterogeneous data sources can provide a broader representation of the target phenomenon^{5,15}.

4.2. Data acquisition

The data acquisition stage collects raw information from the identified sources. The collected information may include structured and unstructured data depending on the application.

4.3. Data preprocessing

Preprocessing is performed to improve data quality.

The process may include:

- Missing-value handling
- Duplicate removal
- Noise reduction
- Data normalization
- Data transformation
- Format standardization

Data preparation is a fundamental component of established data-mining methodologies^{1,2}.

4.4. Feature engineering

Feature engineering transforms processed data into representations suitable for machine learning algorithms.

This stage includes:

- Feature selection
- Feature extraction
- Feature transformation
- Derived-feature generation
- Removal of irrelevant features

Feature representation has an important influence on the ability of machine-learning models to learn useful patterns^{6,7}.

4.5. Contextual integration

This is the central component of DataWeave.

Contextual information is integrated with conventional predictive features.

Possible contextual features include:

- Temporal context
- Spatial context
- Environmental context
- Behavioral context
- Operational context
- Domain-specific context

The resulting representation can be expressed conceptually as:

Integrated Representation = Conventional Features + Contextual Features

Context modeling has been identified as an important component of context-aware intelligent systems^{3,4}.

4.6. Machine learning

The integrated dataset is provided to different machine-learning algorithms.

Possible algorithms include:

- Decision Tree
- Random Forest¹⁰
- Support Vector Machine¹⁶
- K-Nearest Neighbors
- Logistic Regression
- Linear Regression
- Gradient Boosting
- XGBoost¹¹
- LightGBM¹²

4.7. Model evaluation

Models are evaluated using appropriate metrics.

For classification:

- Accuracy
- Precision
- Recall
- F1-score
- ROC-AUC

For regression:

- MAE
- MSE
- RMSE
- R²

Cross-validation can also be used to improve the reliability of performance estimates^{8,17}.

4.8. Model selection

The best-performing model is selected based on:

- Predictive performance

- Computational requirements
- Interpretability
- Application requirements

Model selection and evaluation are important components of reliable machine-learning workflows^{8,17}.

4.9. Prediction

The selected model is finally applied to new observations to generate predictions.

5. DataWeave System Architecture

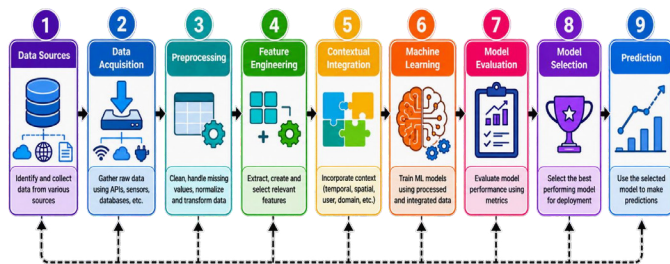


Figure 1: Proposed DataWeave Architecture for Context-Aware Predictive Learning.

Data Sources → Data Acquisition → Preprocessing → Feature Engineering → Contextual

Integration → Machine Learning → Model Evaluation → Model Selection → Prediction

As shown in (Figure 1), the proposed DataWeave architecture follows an end-to-end predictive learning workflow that integrates heterogeneous data sources with contextual information to improve prediction accuracy and reliability. The process begins with data sources and data acquisition, where structured and unstructured data are collected from databases, sensors, APIs and other digital platforms. The acquired data are then preprocessed to handle missing values, remove inconsistencies, normalize the data and prepare it for analysis. Feature engineering is performed to extract and select the most relevant features. Contextual integration incorporates temporal, spatial, user and domain-specific information to provide a more comprehensive representation of the data. The integrated dataset is subsequently used to train machine learning models, whose performance is evaluated using appropriate evaluation metrics. Based on the evaluation results, the best-performing model is selected and applied to generate predictions. This workflow enables DataWeave to provide context-aware and reliable predictive outcomes.

6. Experimental Setup

6.1. Dataset

A suitable real-world or benchmark dataset should be selected for experimental validation.

Parameter	Description
Dataset Name	Actual dataset
Data Source	Actual source
Number of Samples	Actual number
Number of Features	Actual number
Context Features	Actual context
Target Variable	Actual target
Task	Classification/Regression

6.2. Data preprocessing

The experimental preprocessing procedure should describe:

- Missing-value treatment
- Duplicate removal
- Encoding
- Normalization/scaling
- Feature selection
- Outlier handling

6.3. Training and testing

The dataset should be divided into training and testing subsets using the actual experimental configuration.

Cross-validation can additionally be applied during model development^{8,17}.

6.4. Machine-learning models

The experimental study can compare multiple algorithms.

Model	Type
Decision Tree	Tree-based
Random Forest	Ensemble
SVM	Kernel-based
XGBoost	Boosting
LightGBM	Boosting

Only models actually used in the experiment should be reported.

6.5. Evaluation metrics

The selected evaluation metrics should correspond to the prediction task.

7. Results and Discussion

This section should contain actual experimental results.

7.1. Model performance

Table 1: Performance Comparison.

Model	Without Context	With Context	Improvement
Decision Tree	Actual	Actual	Actual
Random Forest	Actual	Actual	Actual
SVM	Actual	Actual	Actual
XGBoost	Actual	Actual	Actual
LightGBM	Actual	Actual	Actual

The comparison should determine whether contextual integration produces measurable improvements.

7.2. Contextual integration analysis

The primary comparison should evaluate:

Without Context

Features → Machine Learning → Prediction versus

With Context

Features + Context → Machine Learning → Prediction

The observed difference should be reported using actual experimental measurements.

7.3. Ablation study

An ablation study can evaluate the contribution of individual contextual components (**Table 2**).

Table 2: Contextual Ablation Study.

Configuration	Context	Performance
A	None	Actual
B	Temporal	Actual
C	Temporal + Spatial	Actual
D	All Selected Context	Actual

This experiment can provide evidence regarding which contextual components contribute most to predictive performance.

7.4. Model comparison

The best-performing algorithm should be identified based on the predefined evaluation metric.

The discussion should explain:

- Which model performed best
- Whether context improved performance
- Which contextual features contributed most
- Whether improvements were consistent across models
- Computational trade-offs

8. Applications

8.1. Healthcare

Context such as patient history and temporal information can support predictive healthcare applications

8.2. Agriculture

Crop, soil, weather and environmental information can be integrated for agricultural prediction.

8.3. Finance

Transaction characteristics and temporal patterns can support predictive risk analysis.

8.4. Education

Learner activity and learning context can support student-performance prediction.

8.5. Transportation

Traffic conditions, location, time and environmental conditions can support congestion and demand prediction.

8.6. Internet of things

Sensor measurements combined with environmental and operational context can support predictive maintenance and intelligent monitoring.

9. Limitations

The effectiveness of DataWeave depends on the quality and availability of input data.

Contextual information may be:

- Missing
- Noisy

- Incomplete
- Redundant
- Difficult to obtain

The integration of a large number of contextual variables may also increase computational complexity.

Performance can vary depending on:

- Dataset
- Application domain
- Algorithm
- Contextual variables
- Experimental configuration

Therefore, domain-specific validation is required.

10. Future Work

Future research can extend DataWeave toward:

- Real-time prediction
- Adaptive learning
- Online learning
- Deep learning
- Federated learning
- Edge computing
- Concept-drift detection
- Automated contextual-feature discovery
- Dynamic model adaptation
- Explainable AI

Federated learning and distributed learning have received significant research attention for scenarios involving decentralized data¹⁸.

Large-scale evaluation using multiple datasets and application domains can also be performed to investigate generalizability.

11. Conclusion

This paper presented DataWeave, an intelligent architecture for context-aware predictive learning. The architecture integrates heterogeneous data acquisition, preprocessing, feature engineering, contextual integration, machine-learning model development, evaluation, model selection and prediction into a unified workflow.

The central contribution of DataWeave is the integration of contextual information with conventional predictive features. By considering temporal, spatial, environmental, behavioral and operational context, the architecture provides a structured approach for developing predictive systems that better represent real-world conditions.

The modular design allows different preprocessing methods, feature-engineering techniques, machine-learning algorithms and evaluation methods to be incorporated according to application requirements.

However, the effectiveness of the architecture must be demonstrated through actual experimental evaluation. Comparative experiments, contextual ablation studies, multiple datasets and appropriate statistical analysis can provide stronger

evidence regarding the contribution and generalizability of DataWeave.

12. References

1. Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases,” *AI Magazine*, 1996;17(3): 37-54.
2. Shearer C. The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing*, 2000;5(4): 13-22.
3. Dey AK. Understanding and Using Context. *Personal and Ubiquitous Computing*, 2001;5: 4-7.
4. Adomavicius G, Mobasher B, Ricci F, et al. Context-Aware Recommender Systems. *AI Magazine*, 2011;32(3): 67-80.
5. Han J, Pei J, Tong H. *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.
6. Mitchell TM. *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.
7. Bishop CM. *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
8. Pedregosa F. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011;12: 2825-2830.
9. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed. Springer, 2009.
10. Breiman L. Random Forests. *Machine Learning*, 2001;45: 5-32.
11. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 785-794.
12. Ke G, Meng Q, Finley T, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, 2017;30: 3146-3154.
13. Lipton ZC. The Mythos of Model Interpretability. *Communications of the ACM*, 2018;61(10): 36-43.
14. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 2017;30.
15. Witten IH, Frank E, Hall MA, et al. *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Morgan Kaufmann, 2016.
16. Cortes C, Vapnik V. Support-Vector Networks. *Machine Learning*, 1995;20: 273-297.
17. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 14th International Joint Conference on Artificial Intelligence, 1995: 1137-1143.
18. Kairouz P, McMahan HB. Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning*, 2021;14(1-2): 1-210.
19. Freund Y, Schapire RE. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 1997;55(1): 119-139.
20. Vapnik V. *The Nature of Statistical Learning Theory*, 2nd ed. Springer, 2000.