**CIND820 XJH - Big Data Analytics Project**

**Title: Context Awareness in Data Mining: Gaining Deeper and More Accurate Data Insights**

**Table of Contents**

[1. Abstract 3](#_Toc192305748)

[2. Literature Review 5](#_Toc192305749)

[2.1. Introduction 5](#_Toc192305750)

[2.2. Context-Awareness in Data Mining 5](#_Toc192305751)

[2.3. Key Concepts in Context-Aware Data Mining 6](#_Toc192305752)

[2.3.1. Feature Engineering with Context 7](#_Toc192305753)

[2.3.2. Context-Aware Clustering 8](#_Toc192305754)

[2.3.3. Context-Aware Classification 9](#_Toc192305755)

[2.4. Summary 10](#_Toc192305756)

[2.5. Challenges & Research Gaps 11](#_Toc192305757)

[2.6. Conclusion 12](#_Toc192305758)

[3. Objectives and Research Questions 12](#_Toc192305759)

[4. Methodology 13](#_Toc192305760)

[5. Data Presentation 14](#_Toc192305761)

[5.1. Data Cleaning, Summary Statistics, Correlations and Anomalies 14](#_Toc192305762)

[5.2. Data Comments 15](#_Toc192305763)

[6. References 17](#_Toc192305764)

1. Abstract

Data mining and knowledge discovery aim to extract valuable information from large datasets. Specifically, data mining focuses on the discovery of patterns, correlations, and anomalies using techniques such as classification, clustering, and regression. These techniques analyze large volumes of current and historical data to identify patterns or establish causal relationships. Traditionally, algorithms focus on identifying patterns within the primary data itself and do not utilize context data. However, integrating contextual data can greatly enhance their accuracy and the quality of discovered patterns.

Context data refers to additional information that provides a deeper understanding of the primary data. In the context of data mining, this could include factors such as temporal data (seasonality, time of day), spatial data (geographic location), environmental data (weather conditions), and socio-economic data (demographics, economic indicators). By incorporating such context data, models can leverage additional variables from the surrounding environment to capture more nuanced trends and patterns, leading to more accurate and relevant predictions.

This project aims to evaluate the introduction of context data into various data mining techniques. More specifically, it will strive to assess the best way to incorporate external contextual data and identify the techniques that stand to benefit the most from context awareness. The project will employ data from the Canada Government Datasets, as well as the Ontario Government Datasets. Specifically, the Canadian Student Tobacco, Alcohol, and Drugs Survey results will be augmented with information from provincial datasets on programs and services that focus on children and youth, such as nutrition programs, mental health services, and social services, as well as provincial laws on substance use (differences in legal age, taxation, and restrictions on advertising) and school/regional policies on substance use prevention.

To address the stated problem, various data mining techniques and tools may be employed. Techniques such as classification and clustering will be used to uncover patterns and relationships within the datasets. Regression analysis will be used to predict trends and assess the impact of different variables. Tools like Python, with libraries such as Pandas for data manipulation, Scikit-learn for machine learning algorithms, and Statsmodels for statistical modeling, will be utilized. Additionally, visualization tools like Matplotlib will aid in presenting the findings in an understandable manner. By combining these techniques and tools, the project aims to create comprehensive and context-aware analyses that provide deeper, more nuanced and more accurate insights into the data.

Link to the working dataset: <https://open.canada.ca/data/dataset/1f15ca45-8bfd-4f9c-9ec6-2c0c440e69c2>

1. Literature Review
   1. Introduction

Data mining and knowledge discovery aim to extract valuable information from large datasets. Specifically, data mining focuses on the discovery of patterns, correlations, and anomalies using techniques such as classification, clustering, and regression. These techniques analyze large volumes of current and historical data to identify patterns or establish causal relationships. In the evolving landscape of data mining, context awareness has emerged as a fundamental approach to enhancing the accuracy, relevance, and adaptability of data mining techniques.

Traditionally, algorithms focus on identifying patterns within the primary data itself and do not utilize context data. However, real-world data is often influenced by external elements such as time, location, environmental conditions, user behavior, and socio-economic elements [‎1]. Integrating these contextual factors into data mining processes can greatly enhance their accuracy and the quality of discovered patterns.

The objective of this literature review is to examine various techniques, applications, and challenges in context-aware data mining, drawing insights from recent research studies. By exploring key concepts such as feature engineering, context-aware clustering, and predictive modeling, this review aims to highlight the benefits of integrating contextual information into the data mining process.

* 1. Context-Awareness in Data Mining

Context-aware data mining improves data analysis by incorporating external factors that influence patterns and decision-making. Context data refers to additional information that provides a deeper understanding of the primary data. In the context of data mining, context can be classified into several types [‎2, ‎3], including:

* **Temporal context** which identifies time-based trends, such as seasonality or time of day.
* **Spatial context** whichintegrates geographical elements such as location data.
* **Environmental context** which considers external factors such as weather conditions or device information.
* **Socio-economic context** which considers external factors such as demographics or economic indicators**.**
* **User-specific context** whichleverages personal preferences, browsing history, and behavioral data.

By incorporating such context data, models can leverage additional variables from the surrounding environment to capture more nuanced trends and patterns, leading to more accurate and relevant predictions [‎1-3].

* 1. Key Concepts in Context-Aware Data Mining

As previously noted, context-aware data mining enhances traditional methods by integrating external factors such as time, location, user behavior, and environmental variables. This enables models to better capture complex patterns and dependencies, leading to improved accuracy and adaptability. Two key areas where context-awareness has been applied are clustering and classification. This literature review will focus on context-aware classification and clustering, and several related key concepts.

* + 1. Feature Engineering with Context

Feature engineering with context involves identifying, selecting, and transforming raw data into meaningful features that enhance model performance by incorporating relevant contextual information [2, 3]. In context-aware data mining, feature engineering plays a crucial role in bridging the gap between raw data and its real-world influences, ensuring that models account for external factors that impact decision-making.

Feature engineering facilitates data integration, which involves merging data from multiple sources into a unified representation, requiring explicit mappings to transform source data into the target format. To enable this process, researchers have explored techniques such as manual feature extraction, entity linking rules and AI-driven schema matching, as well as ontology alignment. Additionally, advancements in machine learning (ML) and natural language processing (NLP) have enabled the extraction of meaningful insights from unstructured text to support data integration [‎1].

Currently available feature engineering techniques include [1]:

* **Manual Feature Extraction:** Definition and extraction of features by experts based on domain knowledge.
* **Automated Feature Learning:** Automatic extraction of relevant features from data by ML algorithms, including deep learning and autoencoders.
* **Contextual Attribute Selection with Dimensionality Reduction Techniques:** Methods like Principal Component Analysis (PCA) help identify and integrate the most significant contextual features while eliminating redundancy [4].
* **Ontology-Based Feature Engineering:** Ontology frameworks define relationships between contextual attributes to enhance adaptability and interpretation [4].

Feature selection and extraction methods enhance data relevance by incorporating contextual variables. The relevance of feature engineering in context-aware data mining lies in its ability to improve predictive accuracy, reduce noise, and enhance interpretability. Without proper feature engineering, models may fail to capture the underlying contextual dependencies, leading to suboptimal performance [2-4].

* + 1. Context-Aware Clustering

Clustering is the process of grouping similar data points based on inherent patterns. While traditional clustering algorithms (e.g., k-means, DBSCAN) rely solely on intrinsic data attributes, context-aware clustering techniques benefit from integrating contextual data to improve pattern recognition. Scheele et al. [5] demonstrated the effectiveness of geographic context in clustering social media messages for situational awareness.

Key techniques in context-aware clustering include:

* **Spatial-Temporal Clustering:** This technique integrates both spatial (location-based) and temporal (time-based) data to refine clusters [6]. It is frequently applied in geographic information systems (GIS), traffic prediction, and disease outbreak monitoring. For example, Avram et al. [7] explored the role of spatial-temporal context in supply chain finance, while Vajirkar et al. [4] applied similar approaches to medical data analysis.
* **User-Centric Clustering:** This technique uses personalized behavioral data to refine clusters. It is frequently applied in e-commerce, social media analysis, and personalized marketing. For example, a graph-based temporal clustering approach fuses short-term and long-term user preferences for recommendations [8].
* **Ontology-Based Clustering:** This technique uses domain-specific ontologies to define relationships between concepts, thereby improving cluster accuracy. It is frequently applied in healthcare, bioinformatics, and semantic web applications. For example, a context-aware framework that clusters wireless medical data by integrating medical ontologies [4].
* **Context-Based Anomaly Detection:** This technique improves the identification of irregular patterns by integrating additional environmental factors. It is frequently applied in cybersecurity, finance and fraud detection. For example, Batarseh & Kulkarni [9] introduced the Context-driven Data Science Lifecycle (C-DSL) to address bias and improve anomaly detection in cybersecurity applications.
  + 1. Context-Aware Classification

Incorporating contextual data into classification models can lead to more robust predictions. Classification assigns labels to data points based on predefined categories. Context-aware classification improves predictive accuracy by integrating additional contextual information into feature selection, model training, and decision-making.

Key techniques in context-aware classification include:

* **Context-Driven Feature Selection:** This technique selects relevant contextual attributes (e.g., weather, user demographics) to enhance classification performance. It is frequently applied in fraud detection, medical diagnosis, and financial risk assessment. For example, a context-aware feature selection method was developed for classifying financial transactions by integrating user and temporal context [10].
* **Adaptive Decision Boundaries:** In this method, classification models adjust decision boundaries dynamically based on real-time contextual changes. It is frequently applied in disease prediction, personalized medicine, and real-time event classification. For example, the CASP-DM process extends decision trees to include adaptive context-awareness [11].
* **Context-Aware Deep Learning:** This technique uses embedding to incorporate contextual data into deep learning models. It is frequently applied in NLP, image recognition, and cybersecurity. For example, BERT (Bidirectional Encoder Representations from Transformers) captures linguistic context for more accurate text classification [12].
  1. Summary

Context-aware clustering and classification techniques significantly enhance traditional data mining by integrating spatial, temporal, user, and environmental factors. They have a multitude of applications:

* **Healthcare:** Enhancing predictive analytics for patient diagnosis and personalized treatment by incorporating environmental and behavioral context as demonstrated by Vajirkar et al. [4].
* **E-Commerce:** Improving recommendation systems by integrating user preferences, browsing patterns, and contextual factors such as time of browsing and purchase history as demonstrated by Lore et al. [13].
* **Smart Cities:** Optimizing urban planning and traffic management using real-time contextual data from sensors and IoT devices as demonstrated by Scheele et al. [5].
* **Cybersecurity:** Strengthening adaptive intrusion detection systems by incorporating contextual network traffic analysis and behavioral patterns as demonstrated by Batarseh et al. [9].
* **Finance & Fraud Detection:** Enhancing risk assessment models by factoring in transaction history, user behavior, and external economic conditions as demonstrated by Martínez-Plumed et al. [11].

It is important to note that while these methods improve model accuracy and adaptability, they also introduce challenges related to computational cost, data complexity, and interpretability.

* 1. Challenges & Research Gaps

Despite advancements, context-aware data mining faces several challenges:

* **Data Collection & Quality Issues:** Inconsistent, noisy, or biased contextual data can reduce model reliability. Also, contextual features may be missing or sparsely distributed leading to data imbalance.
* **Computational Complexity:** Processing large-scale contextual datasets increases complexity and requires significant computational resources and optimization techniques since integrating diverse contextual features is non-trivial.
* **Privacy & Security Concerns:** Handling sensitive contextual data introduces ethical and legal considerations.
* **Scalability & Generalization:** Context-aware models often struggle to adapt across diverse domains and applications, limiting their effectiveness in real-world scenarios. Furthermore, real-world contexts continuously evolve, requiring adaptive clustering techniques.
* **Explainability**: Understanding how contextual factors influence classification decisions is crucial but can be quite difficult given the diverse contextual features and dataset complexity.

Future research should focus on scalable, real-time context-aware models, ensuring ethical and explainable AI solutions.

* 1. Conclusion

Context-aware data mining enhances the accuracy and adaptability of data-driven models by integrating relevant external factors. Applications in healthcare, cybersecurity, finance, and smart cities demonstrate its potential in real-world scenarios. However, challenges such as data quality, computational costs, and privacy concerns must be addressed to maximize its benefits. Future research should focus on developing scalable, ethical, and privacy-preserving context-aware models for broader applications.

1. Objectives and Research Questions

Research has demonstrated that including contextual factors in traditional processes can improve the overall results achieved through various data mining methodologies [1-4, 7]. This project aims to evaluate the introduction of context data into two data mining techniques: classification and clustering. More specifically, the project will:

* Analyze the different context-aware feature engineering methods outlined in the literature to determine the best approach for incorporating external contextual data.
* Compare improvements in classification and clustering results when integrating contextual data, identifying which technique benefits the most.
* Identify and discuss factors that influence data mining behavior and evaluate how they impact performance across different contextual settings.
* Examine system behavior under varying contextual conditions to determine how similar input sets produce different outcomes based on contextual changes.

In doing so, the project will address the following research questions:

* How can external contextual data be effectively integrated into feature engineering for classification and clustering techniques?
* Which of the two data mining techniques - classification or clustering - benefits the most from context-awareness?
* How does the system’s mining behavior change under varying contextual conditions when applied to the same input data?

1. Methodology

The project will employ data from the Canada Government Datasets. Specifically, the Canadian Student Tobacco, Alcohol, and Drugs Survey (CSTADS) results will be augmented with information from provincial datasets on programs and services that focus on children and youth, such as nutrition programs, mental health services, and social services, as well as provincial laws on substance use and school/regional policies on substance use prevention.

More specifically, provincial-level data will be collected for each province in Canada from government sources (e.g., Health Canada, Statistics Canada, CanLII, provincial health departments). For example, for provincial laws on substance use, key variables to collect will be legal age for purchasing tobacco, alcohol, e-cigarettes and cannabis, taxation rates, and advertising and sales restrictions (e.g., bans on flavored products, point-of-sale promotions, etc.). The collected data will be merged with CSTADS data by province (since the CSTADS dataset includes province-level identifiers for students). Subsequently, contextual features will be created and added to CSTADS. This process will be repeated for various contextual elements. Subsequently, the created dataset will then be analyzed to address the research questions.

Various data mining techniques and tools will be employed. Classification and clustering will be used to uncover patterns and relationships within the datasets. Regression analysis will be used to predict trends and assess the impact of different variables. Tools like Python, with libraries such as Pandas for data manipulation, Scikit-learn for machine learning algorithms, and Statsmodels for statistical modeling, will be utilized. Additionally, visualization tools like Matplotlib will aid in presenting the findings in an understandable manner. By combining these techniques and tools, the project aims to create comprehensive and context-aware analyses that provide deeper, more nuanced and more accurate insights into the data.

1. Data Presentation

Link to the working dataset: <https://open.canada.ca/data/dataset/1f15ca45-8bfd-4f9c-9ec6-2c0c440e69c2>

* 1. Data Cleaning, Summary Statistics, Correlations and Anomalies

A Google Collaboratory notebook which contains the information is attached as Appendix 1.

* 1. Data Comments

The dataset contains 168 attributes comprised of a combination of attributes that describe the location and characteristics of the respondents (inherent contextual data), as well as attributes that represent survey responses. In addition, the survey has derived variables, which can be computed through a combination of several attributes.

Of note, this dataset incorporates several methods of denoting missing information:

* 96 and 996: Valid skip (based on skip patterns)
* 98: Prefer not to answer
* 99 and 999: Not stated (i.e., no response, invalid/un-codable, or suppressed)

All of the above will be treated as missing values. Attributes which contain over 75% missing values will be eliminated from the dataset, leaving a total of 138 attributes.

The attributes are all coded numerically however many of them represent categorical variables. Categorical variables will be standardized in accordance with the CSTADS user guide. Attributes are coded as follows:

|  |  |
| --- | --- |
| **Codes** | **Themes** |
| SEQID, PROVID, GRADE, DVGENDER, DVURBAN, DVRES, DVORIENT, DVDESCRIBE, WTPUMF, GH\* | Respondent data and demographic information. |
| SS\* / WP\* / SC\* / CA\* / TP\* | Questions regarding cigarettes. |
| ELC\* / VAP\* / CI\* | Questions regarding e-cigarettes and vaping. |
| ALC\* | Questions regarding alcohol. |
| NRG\* | Questions regarding energy drinks. |
| CAN\* | Questions regarding cannabis. |
| UND\* / MET\* / XTC\* / HAL\* / HER\* / COC\* / SYN\* / BZP\* / TNB\* / TRP\* / GLU\* / SAL\* / SLP\* / STI\* / DEX\* / GRV\* / SED\* / PR\* / | Questions regarding other substances. |
| POLY\* | Questions regarding simultaneous use of multiple substances. |
| PH\* | Questions regarding perceptions of harm. |
| DR\* | Questions regarding driving under the influence. |
| BEH\* | Questions regarding parental attitudes, |
| BUL\* | Questions regarding bullying. |
| DVTY1ST, DVTY2ST, DVLAST30, DVAMTSMK, DVCIGWK, DVNDSMK, DVAVCIGD | Derived variables regarding smoking. |

The project will be focusing on smoking, alcohol and cannabis use, as such columns associated with the use of other substances will be eliminated.

1. References
2. Debruyne, C. (2021). Introducing Context and Context-awareness in Data Integration: Identifying the Problem and a Preliminary Case Study on Informed Consent. In *Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services (iiWAS '20)*, 178–183. https://doi.org/10.1145/3428757.3429116
3. Dey, A. K., Abowd, G. D., & Salber, D. (2001). A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications. *Human–Computer Interaction*, *16*(2–4), 97–166.   
   https://doi.org/10.1207/S15327051HCI16234\_02
4. Abowd, G.D., Dey, A.K., Brown, P.J., Davies, N., Smith, M., Steggles, P. (1999). Towards a Better Understanding of Context and Context-Awareness. In: Gellersen, HW. (eds) Handheld and Ubiquitous Computing. HUC 1999. Lecture Notes in Computer Science, vol 1707. Springer, Berlin, Heidelberg.   
   https://doi.org/10.1007/3-540-48157-5\_29
5. Vajirkar, P., Singh, S., Lee, Y. (2003). Context-Aware Data Mining Framework for Wireless Medical Application. In: Mařík, V., Retschitzegger, W., Štěpánková, O. (eds) Database and Expert Systems Applications. DEXA 2003. Lecture Notes in Computer Science, vol 2736. Springer, Berlin, Heidelberg.   
   https://doi.org/10.1007/978-3-540-45227-0\_38
6. Scheele, C., Yu, M., & Huang, Q. (2021). Geographic context-aware text mining: Enhancing social media message classification for situational awareness. *International Journal of Digital Earth*.   
   https://doi.org/10.1080/17538947.2021.1936321
7. Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial-temporal data. *Data & Knowledge Engineering*, 60 (1), 208-221. https://doi.org/10.1016/j.datak.2006.01.005
8. Avram, A., Matei, O., Pintea, C., & Anton, C. (2020). Innovative platform for designing hybrid collaborative & context-aware data mining scenarios. *Mathematics*, 8(11), 1867. https://doi.org/10.3390/math8111867
9. Xiang, L., Yuan, Q., Zhao, S., et al. (2010). Temporal recommendation on graphs via long- and short-term preference fusion. *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 723-732. https://doi.org/10.1145/1835804.1835896
10. Batarseh, F. A., & Kulkarni, A. (2019). Context-driven data mining through bias removal and data incompleteness mitigation. *arXiv preprint*. https://doi.org/10.48550/arXiv.1909.07956.
11. Liu, W., Liu, N., & Huang, J. (2012). Context-aware feature selection and classification. *IEEE Transactions on Knowledge and Data Engineering*, 24 (7), 1320-1332. https://doi.org/10.1109/TKDE.2011.142
12. Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., et al. (2017). CASP-DM: Context-aware standard process for data mining. *arXiv preprint*. https://arxiv.org/abs/1709.01974
13. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171-4186.   
    https://doi.org/10.48550/arXiv.1810.04805
14. Lore, K. G., & Reddy, K. K. (2019). "Implicit Context-Aware Learning and Discovery for Streaming Data Analytics." arXiv preprint.