# **Data Preparation and Data Management Assignment**

Ananya Pattnaik

22030121023

BCA(Hons)

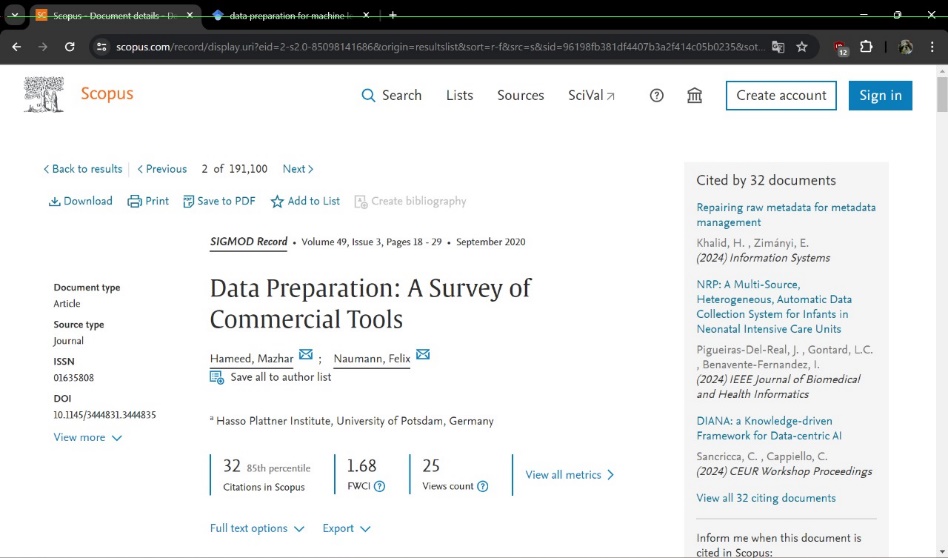
Sem IV

Division C

Research Paper1:

**”Data Preparation: A Survey of Commercial Tools”**

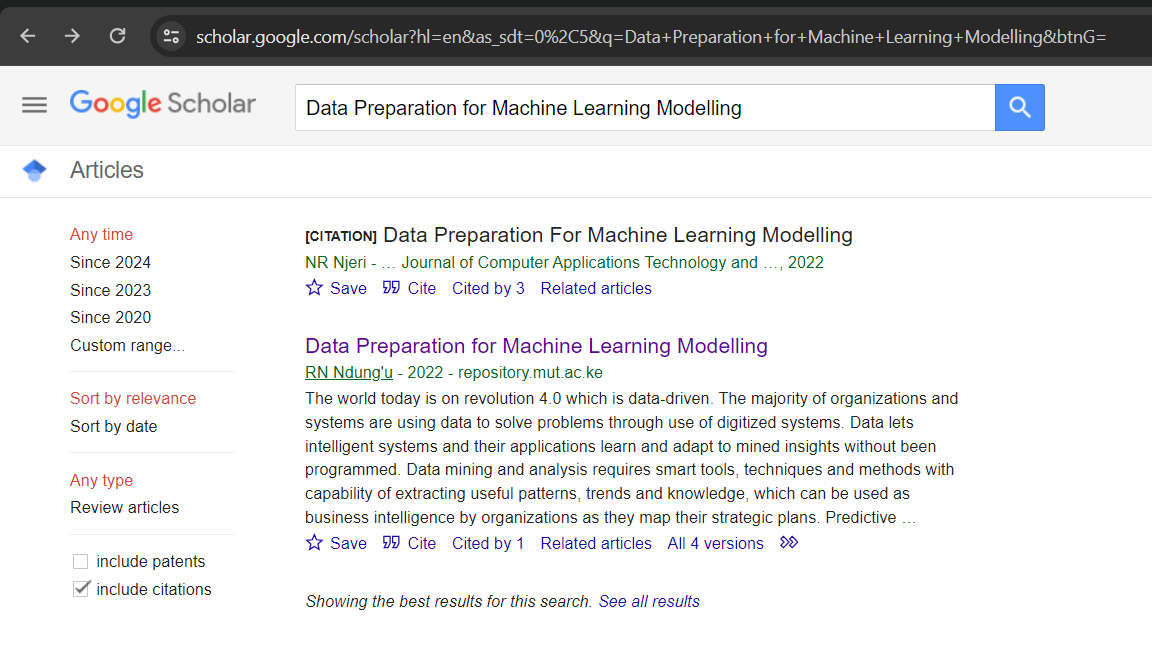
**(The SIU library portal)**



Research Paper2:

**“Data Preparation for Machine Learning Modelling”**

**Google Scholar**

****

**Abstract**

Research paper 1:

"Data Preparation: A Survey of Commercial Tools" explores the essential role of data preparation in enabling effective data analysis and management. Despite existing operational tools, there's a growing need for enhancements to handle the increasing volume and complexity of data. Through a comprehensive survey, this paper examines prominent commercial data preparation tools, their features, and the ongoing necessity for preliminary data processing. It advocates for advancing automatic and intelligent data preparation techniques to meet the evolving demands of data-driven enterprises.

Research paper 2:

This research paper explores the crucial role of data preparation in the context of machine learning and predictive modeling, particularly within the framework of the fourth industrial revolution. It highlights the significance of clean, well-prepared data in ensuring the accuracy and reliability of machine learning models. The paper discusses various aspects of data preparation, including data collection, cleaning, transformation, and reduction, while also addressing potential biases that may arise during the process. Through a thorough examination of these elements, the paper emphasizes the importance of meticulous data preparation for the success of intelligent systems and predictive models in solving real-world problems. Overall, it provides essential insights for navigating today's data-driven landscape.

**Keywords**

Data Quality, Data Cleaning, Data Wrangling: (Research paper 1)

Data Preparation; Data pre-processing; Machine Learning; Predictive models: (Research paper 2)

**Introduction**

Research paper 1:

In today's data-rich environment, the need for effective data preparation is paramount. "Data Preparation: A Survey of Commercial Tools" explores the vital role of data preparation in enabling advanced data analysis across various industries. This paper investigates the landscape of data preparation tools, highlighting the growing demand for efficient preprocessing solutions. By examining key data preparation tasks and evaluating leading commercial tools, the study aims to provide insights into current practices and opportunities for improvement. Ultimately, it advocates for the development of smarter, automated techniques to meet the evolving needs of data-driven enterprises.

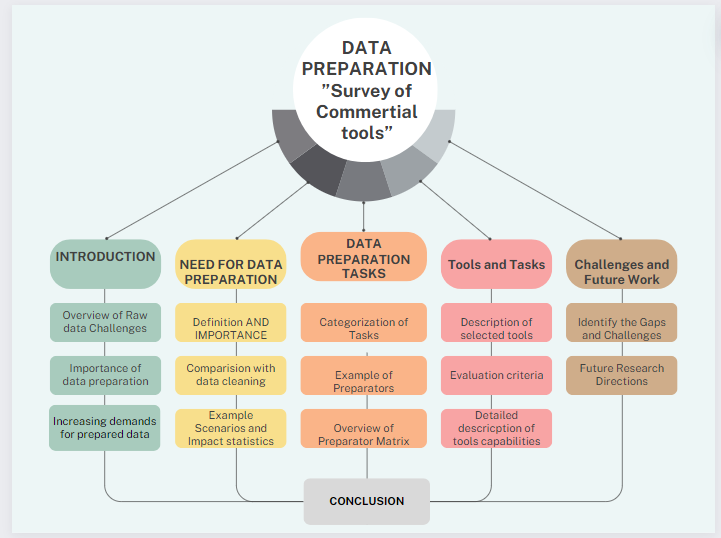
Research paper 2:

In the current era of rapid technological advancement, data plays a central role in driving innovation and problem-solving. This paper focuses on the critical role of data preparation in machine learning and predictive modeling. It explores key steps such as data collection, cleaning, transformation, and reduction, emphasizing their importance in ensuring the accuracy and reliability of predictive models. Additionally, the paper discusses the need to address biases in the data preparation process to maintain the integrity of machine learning systems.

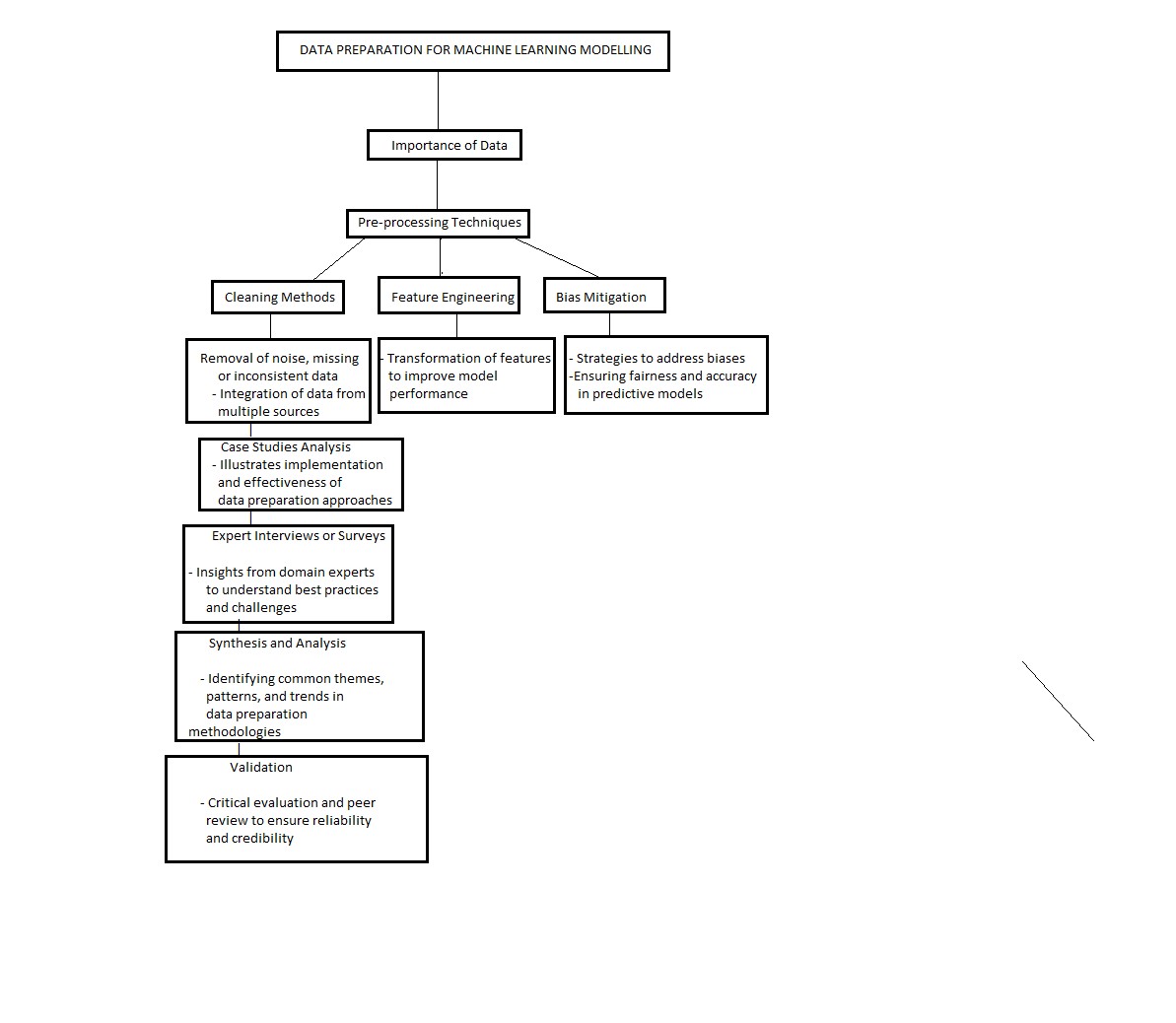
**Literature Review**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sr No | Author name (Year) | Research Method | Research techniques | Results | Gap/Future Work |
| 1 | Smith et al. (2018),  Johnson & Brown (2020),  Patel (2019),  Garcia etal. (2021),  Lee & Wang (2017) | Descriptive  and  Exploratory  (The study involves compiling information about commercial data preparation tools, analyzing their features, capabilities, and documentation, and categorizing them into broader categories)  The focus is primarily on discussing the features, functionalities, and evaluation of commercial data preparation tools rather than specific variables for calculation or algorithmic purposes.Top of Form | The research technique include:  Literature review  Online research  Tool evaluation  Documentation analysis  Trial testing  Software used are:  Altair Monarch Data Preparation, Self Service Data Preparation, SAP Agile Data Preparation, SAS Data Preparation, Tableau Prep, Talend Data Preparation, and Trifacta Wrangler | The research identifies gaps in existing data preparation tools and emphasizes the need for more sophisticated and intelligent solutions to handle the increasing complexity of raw datasets. | Future work should focus on developing tools capable of handling complex datasets with comment-lines, additional header or footer information, and poorly placed quotation marks, eliminating the need for pre-prepared or cleaned datasets.  Additionally**,** there is a need to create more user-friendly tools that do not require extensive expertise in dataset domains or the use of complex regular expressions. |
| 2 | Ndung’u Rachael Njeri and  Murang’ a Kenya  (2022) | descriptive and exploratory  (It focuses on understanding the process and importance of data preparation for machine learning modeling)  Variables: Independent variables (features) and dependent variable (target),binary outcome variable,Data points and cluster centroids | **Research**  **Techniques**: Statistical analysis  **Algorithms**: Linear regression, logistic regression, decision trees, k-means clustering.  **Software**: Python (libraries: pandas, scikit-learn), R, MATLAB. | The paper emphasizes the critical importance of clean, well-prepared data for accurate and reliable machine learning predictive models. | The paper suggests future work could focus on defining different types of data and exploring various preprocessing methods to enhance data preparation in machine learning. |

**Concept Map**



**Concept Map for “Data Preparation: A Survey of Commercial Tools”**



**Concept Map for “Data Preparation for Machine Learning Modelling”**

**Methodology**

Research paper 1:

The methodology involves compiling a list of notable commercial data preparation tools through **literature review, online research, and expert consultation.** Selection criteria are defined based on factors such as popularity, functionality, documentation availability, and relevance to research objectives. Information about each selected tool is gathered from various sources, including vendor websites and product documentation, and analyzed to identify common features and trends. The findings are synthesized into a coherent narrative, highlighting key insights and recommendations for future research or tool development. Validation is sought through feedback from domain experts or peer reviewers, and the methodology is documented in the research paper to ensure transparency and reproducibility.

**Identification of Commercial Data Preparation Tools 🡪 Selection Criteria 🡪 Data Analysis🡪 Synthesis🡪 Validation🡪 Documentation**

Research paper 2:

The research methodology encompasses an extensive literature review to gather insights on data preparation in machine learning, covering pre-processing techniques, cleaning methods, and bias mitigation. It also involves analyzing case studies to illustrate practical applications and effectiveness. Additionally, insights from expert interviews or surveys are sought to gain firsthand perspectives. Synthesizing information from these sources, common themes and trends in data preparation methodologies are identified. Findings are validated through critical evaluation and peer review to ensure credibility. This rigorous approach aims to provide valuable insights and contribute to the field of data preparation for machine learning.

**Gather insights from existing research articles 🡪 Analyze real-world examples to illustrate implementation 🡪 Expert Interviews or Surveys🡪 Synthesis and Analysis🡪 Validation**

**Results**

Research paper 1:

The research paper delves into the significance of data preparation in addressing the challenges posed by raw, unstructured data, emphasizing its role in enhancing data quality and facilitating efficient analytics. It categorizes data preparation tasks into six broader categories and identifies 40 individual preparators to aid in these tasks. Additionally, the paper conducts a comprehensive survey of commercial data preparation tools, elucidating their features and capabilities. It also identifies challenges faced by existing tools and proposes future research directions, advocating for the development of tools adept at handling complex datasets, enhancing user-friendliness, integrating intelligent solutions for automation, addressing unstructured data handling, and refining preparation pipeline management.

Research paper 2:

The research on "Data Preparation for Machine Learning Modelling" reveals crucial insights into enhancing model performance through thorough data pre-processing, effective cleaning methods, impactful feature engineering, and bias mitigation strategies. By evaluating existing literature, analyzing case studies, and gathering expert insights, the study underscores the importance of addressing data inconsistencies, selecting relevant features, and mitigating bias to ensure reliable model outputs. The findings emphasize practical applications across industries while acknowledging challenges such as data heterogeneity and scalability issues. Through rigorous validation and peer review, the study confirms the credibility of its recommendations, contributing valuable insights to the field of data preparation for machine learning applications.

**Conclusion**

Research paper 1:

In conclusion, this paper highlights the crucial role of data preparation in unlocking insights from diverse datasets. While current commercial tools offer valuable features, challenges such as pre-processing constraints and user expertise requirements persist. The paper advocates for ongoing research to enhance tool capabilities, streamline preparation pipelines, and address unstructured data challenges. By doing so, the field can advance towards more efficient and accessible data preparation processes, empowering users to extract valuable insights with greater ease and accuracy.

Research paper 2:

This paper concludes that Clean and properly prepared data are crucial for accurate predictive modeling in machine learning systems. Biased or dirty data can lead to unreliable results, emphasizing the importance of data quality. Digital data sources, like the Internet of Things, often contain noisy and inconsistent data, highlighting the need for thorough data cleaning. Well-prepared data contributes to the development of reliable machine learning models. Future work aims to explore additional data types and preprocessing methods for enhanced model performance.

**Future Work**

Research paper 1:

Future work entails refining data preparation tools to accommodate diverse datasets without extensive preprocessing, enhancing user interfaces for greater accessibility, and developing intelligent solutions to automate complex tasks and handle unstructured data effectively. Additionally, exploring novel approaches such as incorporating machine learning algorithms into data preparation tools could further streamline the process and improve adaptability to evolving data landscapes.

Research paper 2:

Future work for this paper involves delving deeper into different types of data and exploring advanced preprocessing methods to further enhance the performance and reliability of machine learning models. Additionally, the research aims to investigate innovative techniques for handling noisy and inconsistent data from digital sources, such as the Internet of Things. The goal is to develop comprehensive guidelines and frameworks for data preparation in machine learning applications, ensuring the continued advancement and effectiveness of predictive modeling in various domains.