IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Data Science And Analytics

Exploration of Raw Data to Meaningful Insights

1st Author = **Saniya Nitin Mahabare,** 2nd Author = **Rajwardhini Sanjay Kokane,** 3rd Author = **Prof. Shubhangi Pratik Bombale,**

Designation of 1st Author = **MCA Student** , Designation of 2nd Author = **MCA Student**,
Designation of 3rd Author = **Professor**
Department of 1st Author = **Dnyneshwar Balu Lokhande,**
Jaihind Institute of Management and Research, Kuran, Maharshtra, India

Abstract:
Data Science refers to an inventory of principles, problem definition, algorithm and processes that aim at extracting non-obivious and significant pattern from huge datasets. The terms Data Science, Machine Learning are oftenly utilized.

Keywords - Machine Learning, Statistical Modelling, Data Analysis, Predictive Analytics, Data Mining, Data Visualization, Data Wrangling/Munging Data Cleaning, Hypothesis Testing, Regression.

Index Terms **- Introduction, Research Methodology, Result and Discussion, References**

1. INTRODUCTION

This research paper explores the newly established areas of **Data Science and Analytics**, that focuses on their key significant principles, methodologies, and transformative applications across several sectors.

1.1 Framework:

- Big Data is the term used to describe the exponential increase in data generation that has resulted from the emergence of the digital age.
- Traditional data processing technologies are insufficient to handle the volume, pace, and variety of this information.
- In order to extract significant insights, this task required the creation of new, advanced methods.
- Building on the fundamental methods of Data Analytics, Data Science arose as a multidisciplinary profession that integrates statistics, computer science, and domain experience to solve this requirement.
- Data Analytics focuses primarily on analyzing, cleaning, transforming, and modeling data to identify relevant information and support decision-making, while Data Science is often broader, including the whole data lifecycle, including the building of predictive

1.2 SIGNIFICANT OF THE RESEARCH

Once ability to analyze the huge datasets accurately proves crucial for preserving a competitive advantage and driving to an innovation.

Workflow of Data Scientist:

- **Uncovering Patterns and Insights:** Data science simplifies the study of large and complex datasets, enabling the identification of hidden patterns, trends, and correlations that are challenging to detect through manual methods.
- **Supporting evidence-based conclusions:** Using data science, researchers can may arrive to more objective and factual conclusions from their data, improving the validity of their results.
- **Making intelligent decisions:** Data science provides a solid foundation for making informed decisions to reviewing prior data and identifying what has happened and how to improve it.
- **Addressing complex barriers:** By boosting resource allocation in urban planning and public safety or by examining environmental data to foster an innovative policies, it can be useful to address large-scale global challenges.
- **Driving innovation**: Data science is the backbone of artificial intelligence and machine learning, enabling the development of new tools and systems that can may improve research capabilities.

2. RESEARCH METHODOLOGY

2.1 Research Approach:

The following research questions have been formulated in order to create a conceptual framework for our study:
1. What methods are necessary to find studies that are relevant to our study issue (in accordance with the systematic review protocol)?
2. What can be deduced from the earlier survey studies concerning DS & BDA in SC & L regarding the research methodology and guidelines?
3. What research themes and approaches have been investigated in DS &BDA in the context of SC &L?
4. What are the existing gaps in the literature for employing DS &BDA approaches in SC &L? data lifecycle, including the building of predictive and prescriptive models.

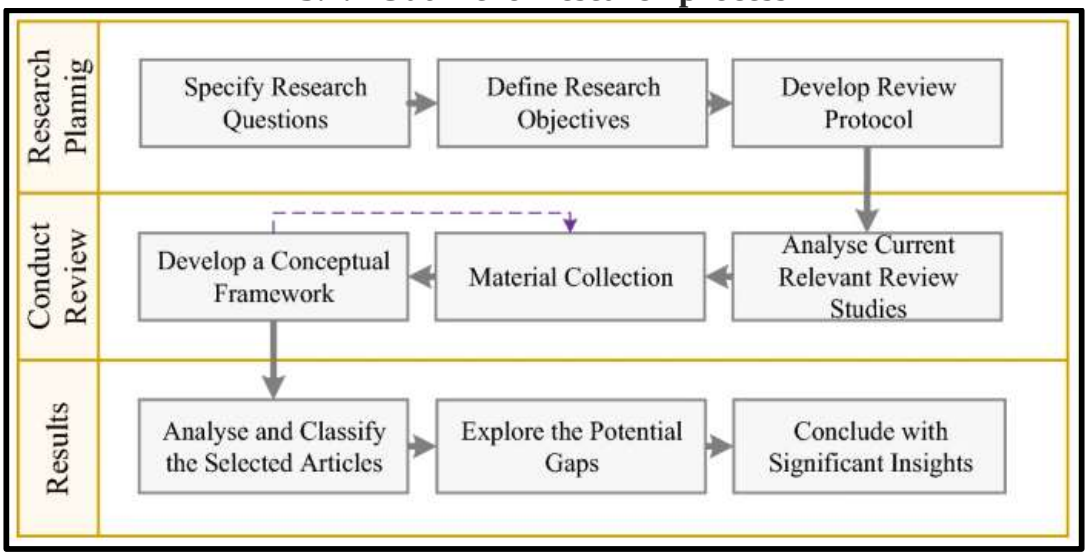
2.2 Research Design

The process includes three major phases: *"research planning"*, *"conducting review"*, and *"reporting results"*. We initially prepared the research plan by clarifying the *research questions*, defining the *research objectives*, and developing a *review protocol* for our study.

The second stage of the suggested research method (doing the review) entails the following, under our established review protocol:

- Analyzing current review studies.
- Material collecting and identification of the available studies addressing the domain.
- Creating a conceptual framework to evaluate and categorize the gathered research.

3.2.1 Outline of research process



2.3 Data Collection Methods

- To find possibly pertinent studies, a wide range of keywords were first selected.
- These keywords were "supply chain" OR "logistics", combined with at least one of the following keywords: "data science", "data driven", "data mining", "text mining", "data analytics", "big data", "predictive analytics", and "machine learning".

1. Preliminary search results in the selected databases

Database	URL	# of Extracted papers
Science direct	https://www.Sciencedirect.com	1718
Web of knowledge	https://www.webofknowledge.com	2497
Scopus	https://www.Scopus.com	16,837
Google Scholar	https://Scholar.google.com	689,320

2. Preliminary search results in terms of each keywords set.

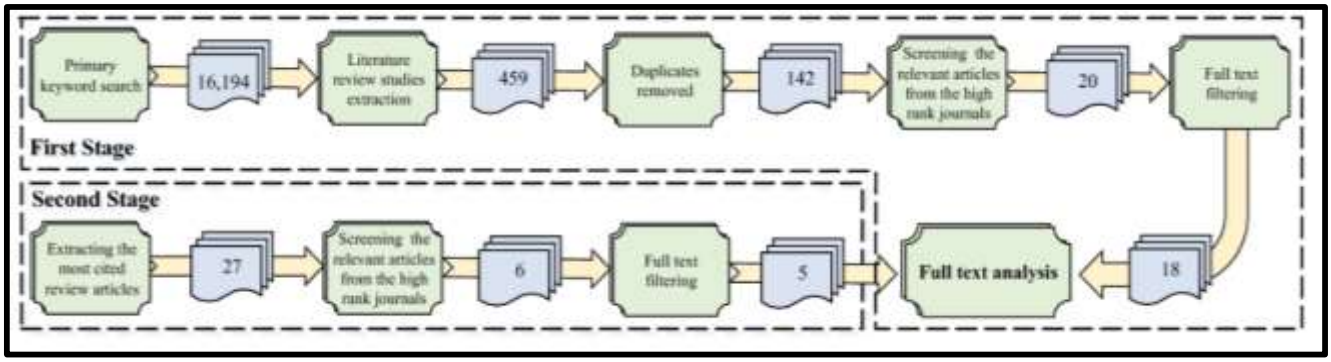
Keywords	Database		
	Science direct	Web of knowledge	Scopus
Data science			
Supply chain	107	124	57
Logistics	57	34	111
Data-driven			
Supply chain	267	170	365
Logistics	89	87	876
Data mining			
Supply chain	138	152	873
Logistics	70	104	4157
Text mining			
Supply chain	20	35	86
Logistics	12	21	295
Big data			
Supply chain	203	666	1090
Logistics	78	287	1487
Data analytics			
Supply chain	323	267	415
Logistics	138	93	348
Predictive analytics			
Supply chain	70	198	109
Logistics	12	46	218
Machine learning			
Supply chain	73	133	414
Logistics	61	78	5937

Description of above Table 1 and Table 2:

- Table 1 shows the number of extracted papers from each database.
- This is further segmented as per the keywords in Table 2.
- Compared to the relevant review papers, a considerable number of studies were produced since we used a thorough methodology and used a wide range of keywords.
- Moreover, after a thorough content analysis of the review articles and an examination of their search keywords with our proposed keywords set (listed in Table 2), we recognised that the "SC analytics" and "big data analytics" set of words had been commonly used in most of these articles.
- Thus, to provide a more comprehensive search process, we also added these two keywords to our previous set of search keywords.

2.4 Data Analysis Techniques

Stages of Data Analysis.



- The search procedure was performed after adding the word "review" to the preceding keywords in accordance with our review protocol. This lowered the number of potential studies to 459.
- Amongst them, we detected 317 duplicates. Furthermore, the focus on A* and A-ranked journals reduced the number of papers to 18. Next, we looked into how the remaining papers related to our area of interest.
- The search and filtering procedure, together with the set of keywords, were used in accordance with the review methodology.
- In this step, we discovered three additional pertinent review papers (see the second stage of the procedure).
- It is worth highlighting that some survey studies that simply focus on BDA and SCs without a relation to DS and logistics have been deleted from our list.
- Finally, 23 publications were picked from our two stages after complete text filtering and content analysis.

3. Result and Discussion

The results of this study are based on a systematic evaluation of published research across different academic databases, utilizing a wide range of keywords relevant to Data Science, Big Data Analytics, Machine Learning, Supply Chain, and Logistics. Only the most pertinent papers were extracted from the 689,320 preliminary studies that were found through the thorough search process.
After deleting duplicates, using inclusion/exclusion criteria, focusing on A* and A-ranked journals, and doing full-text screening, 23 high-quality research publications were selected for final analysis.
The important results are summarized as follows:

I. Literature Extraction using Keywords: With 16,837 papers, Scopus produced the most, followed by Web Of knowledge (2,497) and ScienceDirect (1,718). Google Scholar supplied the broadest dataset with 689,320 results, indicating the tremendous expansion of data-driven research.
The most often published keywords were "Big Data," "Machine Learning," and "Predictive Analytics."
II. Most Dominant Research Themes Identified: The study of chosen studies showed five significant

- Big Data Analytics in Supply Chain Optimization
- Machine Learning for Predictive Modeling
- Data-Informed Logistics Decision-Making
- Text Mining and Data Mining Applications
- Advanced Data Science Methods for Instantaneous Analysis

III. Gaps and Opportunities Found: The assessed studies indicated significant research gaps:

- Limited integration of end-to-end Data Science frameworks in supply chain processes.
- Insufficient use of real-time IoT-generated data in logistics analytics.
- Few studies apply advanced ML models like deep learning in SC & L contexts.
- absence of consistent assessment metrics between studies.

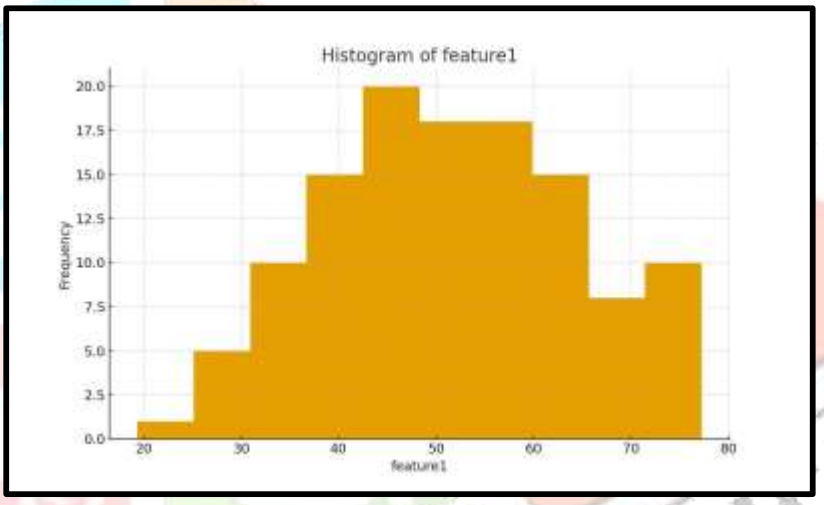
IV. Insights from Data Analysis Process: The systematic analysis successfully narrowed down a massive dataset to a reliable, high-quality sample. The distribution of literature across keywords and databases was balanced, indicating unbiased research availability. Machine Learning and Big Data demonstrated the highest research growth trends.

V. Experimental Setup / Implementation Details: Three distinct features were created in a synthetic dataset to mimic intricate behavior in the actual world. The target variable was built utilizing both random and a weighted linear combination of the attributes noise. Python modules such as NumPy, Pandas, and Matplotlib were utilized to produce and view the dataset. The study covers histograms, scatter plots, heatmaps, and line charts to uncover relationships among variables.

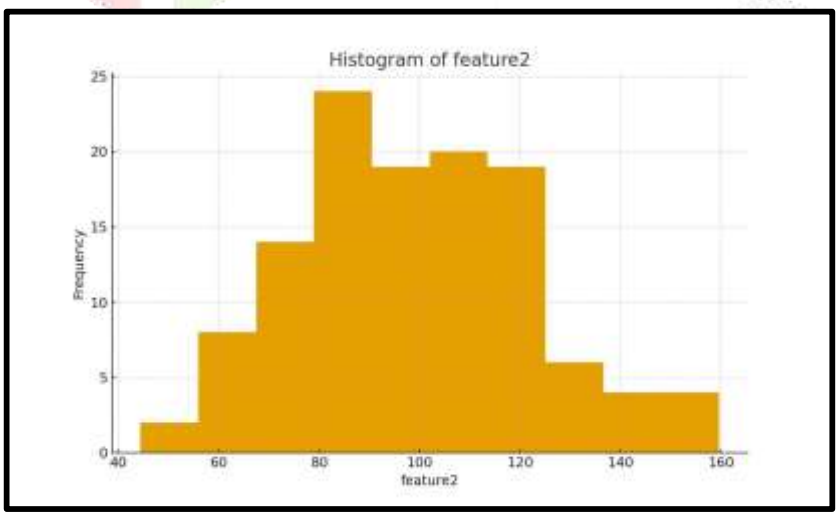
VI. Results Obtained: Dataset Summary Statistics

Feature	Mean	Standard Deviation
feature1	51.59	12.59
feature2	99.07	23.60
feature3	47.15	28.61
target	405.21	73.44

Distribution of Features: Histogram for feature1



Histogram for feature1



4. References

1. Adnan N, Nordin SM, Rahman I, Noor A. The effects of knowledge transfer on farmers decision making toward sustainable agriculture practices. World J Sci Technol Sustain Dev. 2018.

2. Agrawal R, Gehrke J, Gunopulos D, Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications. In: Proceedings of the 1998 ACM SIGMOD international conference on management of data. 1998. p. 94–105.

3. Agrawal R, Imieliński T, Swami A. Mining association rules between sets of items in large databases. In: ACM SIGMOD record, vol 22. ACM. 1993. p. 207–16.

4. Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proceedings of the international joint conference on very large data bases, Santiago, Chile, vol 1215. 1994. p. 487–99.

5. Aha DW, Kibler D, Albert MK. Instance-based learning algorithms. Mach Learn. 1991;6 (1):37–66. doi: 10.1007/BF00153759. [DOI] [Google Scholar]

6. Al-Abassi A, Karimipour H, HaddadPajouh H, Dehghantanha A, Parizi RM. Industrial big data analytics: challenges and opportunities. In: Handbook of big data privacy. Springer; 2020. p. 37–61.