International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

A BIG DATA–DRIVEN FRAMEWORK FOR ANALYSIS OF LARGE AND COMPLEX DATA SETS IN LARGE-SCALE ENVIRONMENTS

Mayana Arifulla Khan

Arifulla93@yahoo.com, makhan@elm.sa

Abstract

This essay describes how the value of Large Data analytics in the Large-Scale Environments is progressive in areas of contribution to operational efficiency, decision making and strategic flexibility. The research examines the opportunities of applying logistics, inventory and customer satisfaction with the help of data-driven insights based on DataCo supply chain at Kaggle. The paper discusses the 5Vs framework i.e., Volume, Velocity, Variety, Veracity and Value with regards to which these aspects are employed to establish the nature and opportunities of Large Data in SCM. It also outlines the key sources of Large-Scale Environments data, i.e., IoT sensors, ERP, e-commerce, and logistics trackers. The article presents the concept of a stratified Large Data system in the process of data ingestion, data storage, data processing, data analysis as well as data visualization, which in turn can be supported by recent tools. In addition, the article speaks on the machine learning algorithms, including Linear Regression and Random Forest, to predict the distribution of the delivery performance and sales. The results indicate how Large Data may be applied in predictive forecasting and monitoring, automating and easing the problems such as data integration, quality, and security. Finally, the article defines the current tendencies in the future of intelligent, and it involves AI-based decision-making, blockchain traceability, and sustainability analytics.

Keywords: Large Data, Machine Learning, Predictive Analytics, IoT, Data Architecture, Blockchain, Artificial Intelligence

1. Introduction

The Large Data issue has been turned into an essential asset of the modern innovative Large-Scale Environments. The analysis of large and complex data is referred to as Large Data that assists organisations to handle the massive amount of data generated in the course of the procurement process, production, warehousing, transportation and delivery to customers. Such Large-Scale Environments activities data sets are enormous and complicated and can be described as high volume, high velocity, and high-variety that require sophisticated analytics technology, sophisticated databases, and machine learning applications in order to identify meaningful insights. The benefits associated with the use of Large Data are operational visibility in real-time, proactive anticipation of disruptions in the Large-Scale Environments, increased accuracy and scientific decision-making and efficiency, cost reduction and profit growth.

The Large-Scale Environments boom data would be accompanied by the growth of Internet of Things (IoT) sensors, Enterprise Resource Planning (ERP) applications, logistics tracking

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

applications and e-commerce applications. The IoT sensors constantly transmit information about the performance of the machines, the conditions of the vehicles, the conditions of the environment and the efficiency of the routing on the other hand, the ERP systems consolidate the information on the transactions included in procurement, sales, and inventory departments. Similarly, the international logistics systems and online shops are also generating customer behaviour and shipping data than ever. Together these interrelated digital resources form an interrelated ecosystem of The Large-Scale Environments operations which can be analysed to maximise processes, improve forecasting and customer satisfaction.

A practically useful example of the application of the Large Data to SCM is the Kaggle DataCo Smart Large-Scale Environments Dataset with detailed information on orders, delivery times, customer types, and product lines. Through the statistical models and machine learning applications to such kinds of large and complex data sets, business is able to establish the trends in performance, future delays in deliveries, customer satisfaction, and also to establish the high-performing product lines. The given insights will allow the managers to optimize the logistics, change the inventory policy, and enhance the overall strategic decision-making.

Most of the time, the concept of Large Data in SCM is described with the help of the five Vs, namely Volume, Velocity, Variety, Veracity, and Value. Volume is the sheer bulk of data being generated every day, Velocity is the rate of data generation and processing, Variety is the types of data generated by sensors, transactions, and customer feedback, Veracity is how precise and reliable the data is, and Value is getting something useful done as a business [4]. These features, as a combination, predetermine the effectiveness of Large Data strategies used in Large-Scale Environments analytics.

Lastly, the Large-Scale Environments activities have been revolutionized by making decisions based on data. The automation and predictive analytics have the potential to assist the companies in enhancing the precision of the predictions, resource allocation as well as competitiveness. The Large Data is also useful in terms of operational efficiencies as well as agility in strategy which enable firms to adapt in a brief time to new market conditions and customer needs.

2. Literature Review

2.1 Technological Evolution

The developments of Large Data technologies over the last 10 years have played an important role in the changes in the Large-Scale Environments management (SCM). The Large-Scale Environments which has been in use is more dependent on the relational databases and manual reporting functions which could not scale and they did not have real time analysis software. However, with the introduction of distributed information processing engines such as Hadoop and Apache Spark, it is possible to approach the processing and handling of Large and complex data sets, which are unstructured in nature. Hadoop MapReduce structure offers the opportunity to handle Massive data in parallel clusters to offer the business an opportunity to challenge the Volume and Variety of data offered by the global logistics networks [5]. Similarly, Apache Spark may be better than Hadoop because it offers in-memory computing, which the service

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

drastically reduces the latency and Velocity increase, therefore, better fits an operation of a Large-Scale Environments time sensitivity like order tracking or route optimization.

It has been noted in the study that has recently taken place that the distributed computing systems have proven effective in handling complex data within the Large-Scale Environments. Spark streaming and Kafka have been used as an example of real-time monitoring of the Large-Scale Environments, as it becomes possible to detect and report any bottlenecks and delays in the transportation process in time [6]. Furthermore, these frameworks could be supplemented with data vision and BI systems, as per which the concerned parties will be able to receive engaging data about the performance parameters and key Large-Scale Environments indicators. A combination of these technological advances has allowed to manage the end-to-end Large-Scale Environments data more effectively leading to improved, faster, and more accurate decision-making and flexibility in the strategic sense.

2.2 Industrial Applications

The application of the Large Data analytics within the various operations of the Large-Scale Environments has provided radical results in logistics, transportation, inventory management, and procurement management. In transportation and logistics, Large Data can be used to plan the routes and track the deliveries, and manage the fleet in accordance with sensor data in real-time. The predictive analytics models that are developed through machine learning algorithms test the state of the traffic, fuel consumption, and actions of the driver to enhance the outcome of the delivery and reduce the costs of operation [7]. As an example, DHL and FedEx, logistics firms, can use predictive route optimization algorithms to ensure that the deliveries are both timely and with the least pollution rates.

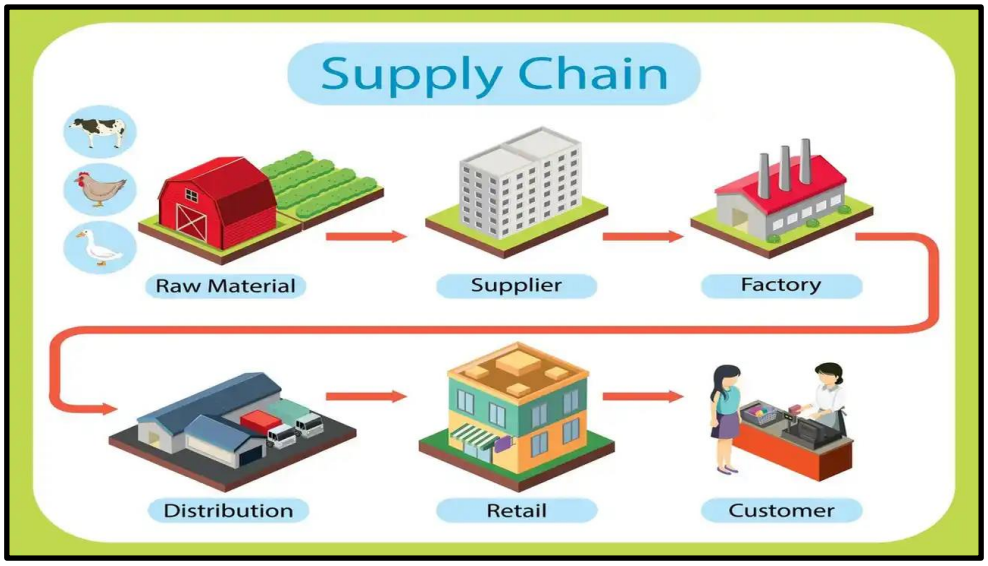


Figure 1: Fundamentals of Large-Scale Environments Management

(Source: wikiemanagement.com)

Deploying Artificial Intelligence (AI) and Machine Learning (ML) on SCM has provided an extra level to the process of decision-making. The predictive models can identify potential interruptions in the Large-Scale Environments and other available suppliers, and can even

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

automate the purchase orders. The anomaly detection algorithms are implemented over the fraud detection systems that detect suspicious transactions or patterns of abnormal procurement acts. Taking the example of combining ML and natural language processing (NLP), by which a company may analyze supplier contracts with compliance and ethical violations, one can be integrated. Moreover, a rather new branch in the field of Large-Scale Environments application is customer analytics, which utilizes the power of Large Data to customize services and predict customer satisfaction. Clickstream analysis and sentiment mining are used on e-commerce websites to make specific recommendations and predict demand spikes [8]. On the whole, the synergy between Large Data, AI, and cloud computing is creating an environment of a smart Large-Scale Environments ecosystem, in which decisions are becoming more and more based on data, predictive, and automated.

2.3 Emerging Trends

With the rapid evolution of digital technologies, new tendencies that are changing the modern Large-Scale Environmentss are many. Internet of Things (IoT) already is among the backgrounds of the next-generation SCM, with all devices able to be fed with real-time information about shipment locations, warehouse temperature, and equipment location [9]. The sensors and RFID tags are IoT-based devices that enhance visibility and tracking and help managers to monitor every step of the Large-Scale Environments. In combination with the Large Data analytics, IoT can be employed to facilitate predictive maintenance, dynamical route optimization, and energy efficiency.

In the meantime, decision systems that are driven by AI are becoming smart assistants in Large-Scale Environmentss activities. These systems combine ERP systems, sensors, and market analytics to execute strategic decisions automaticaly, such as who the suppliers are and when inventory is needed to be renewed. The Large-Scale Environments situations, automation of the procurement strategies, and elimination of risks are simulated more effectively by the algorithms of reinforcement learning and deep learning.

Figure 2: Large-Scale Environments Management Process

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

(Source: www.linkedin.com)

There is a significant research gap in the field of real-time analytics and data governance in the multi-source Large-Scale Environments background. Although the existing systems are very good in historical analysis, most of them are incapable of providing real-time insights that can be utilized in dynamic decision-making [10]. In addition, standards of data and interoperability among global systems are not similar, thus impeding collaboration. Future studies must therefore aim at coming up with scaled yet secure and ethically appropriate data-driven models to incorporate real-time intelligence in Large-Scale Environments networks.

3. Defining Large Data and the 5Vs Framework

The term Large Data in Large-Scale Environments Management denotes the massive amount of data produced by each of the Large-Scale Environments processes of procurement, production, warehousing, logistics, and customer communication. The 5Vs Framework, which consists of 5 concepts: Volume, Velocity, Variety, Veracity, and Value, is a framework that underlies the definition and analysis of the qualities of Large Data. Within the framework of the DataCo Smart Large-Scale Environments Dataset, the five dimensions are quite vivid and essential in the process of data analytics and decision-making.

Volume is characterized by the enormous quantity of data or numerous orders, customers, and shipments. The DataCo dataset is made of thousands of records of transactions connected with sales, delivery details, and product information, which reflects the size of the operations characteristic of the processes in the global Large-Scale Environmentss. This Large Data allows statistical and predictive tools to be used to reveal business trends and performance indicators.

Velocity emphasizes the rate at which data is produced and processed. In the actual context, Large-Scale Environments systems constantly receive live information about online transactions, logistics tracking, and customer service interaction [11]. This is a high-speed data flow that can be used in real-time decision-making, e.g., monitoring the delays in shipment or dynamically changing the inventory volume.

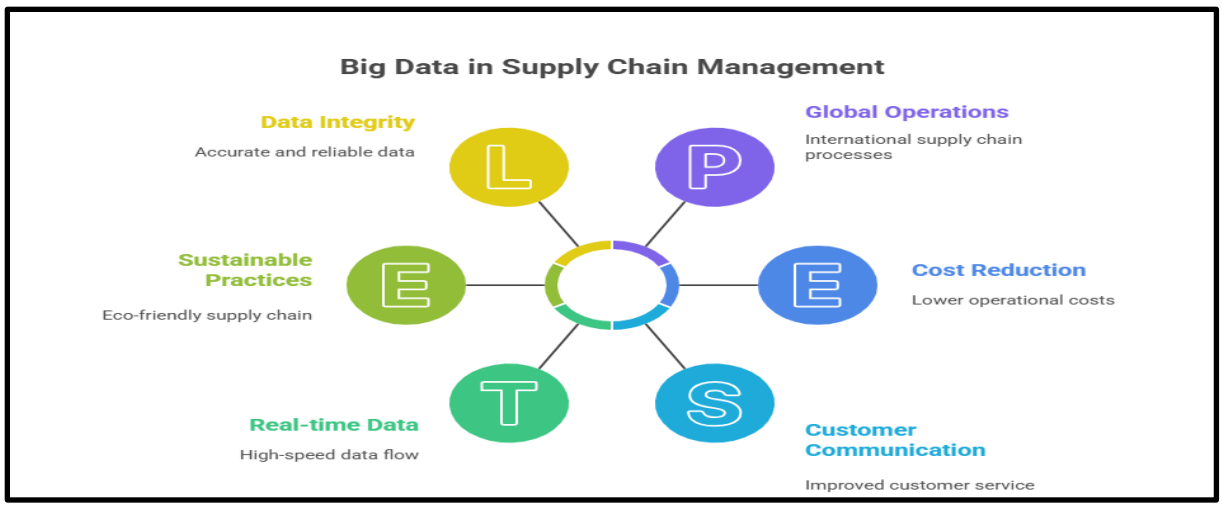


Figure 3: Large Data in Large-Scale Environments Management

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

(Source: Self-Created)

Variety refers to a variety of data formats. The DataCo data is mostly structured data (sales, orders, and inventory), but in real business conditions, it can be combined with semi-structured (e.g., JSON or XML logs) and non-structured data (e.g., emails, reviews, or feedback). Dealing with such diverse sources of data improves the level of analysis, but it also adds complexity.

Veracity is concerned with the quality and reliability of data. The data used in DataCo contains a few inconclusive or missing records, which is characteristic of the real world, where data may be duplicated, or some customer information may be missing. To have the right insights, it is crucial to ensure that the information is accurate and reflective of data integrity. Lastly, Value is the act of analysis applied to action, i.e,. lower costs of operations, faster delivery times, and better forecasting of demand that leads to organizational success.

4. Sources of Large Data in Large-Scale Environments

In the modern day Large-Scale Environments ecosystem, information is being generated constantly and through a myriad of sources all of which combined together provide a general idea of operations. DataCo Smart Large-Scale Environments Dataset is one of such data sources, which contains the information regarding customers, vendors, logistics, and transactional systems. This is due to the nature of these heterogeneous streams of data that allow the Large-Scale Environments network level performance to be forecasted and optimized.

IoT Sensors are some of the most vital sources, and it will measure information of transportation fleet, warehouses, and manufacturing gear in real-time. These sensors monitor the points of shipment, the temperature, the humidity, and conditions of the vehicles and guarantee the quality of the products, as well as their incursion in time. The obtained data is utilized to assume possible disruptions and help in the proactive process of decision-making.

Enterprise Systems: Enterprise resource planning (ERP) and Customer relationship management (CRM) systems are new to generate formal data that can be available to store business transactions, inventory as well as supplier relationships [12]. These systems provide the sales, vendor as well as customer information in detail as structured information in the DataCo dataset does. Now, with a combination of ERP and CRM data, one can improve the process of coordination in procurement, production, and customer service in organizations.

Transactions of e-commerce are the other important source of the data. The online sales platforms generate real time data on orders, payment and their returns, which is invaluable in demand perception and revenue analytics. Such data streams will be included in the Velocity element of Large Data that assists enterprises to fulfill the needs of consumers in a timely manner.

The Customer Feedback and Social Media provide useful data which is not structured. Mentions of brands, rating of the product, and comments provide information regarding customer satisfaction and the satisfaction of the market about it. Such data with text mining and sentiment analysis can help companies to enhance their products by product refinement and customer relationship among others.

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

The logistics Data including the time of transportation, paths and the use of fuel assists in maximizing on the transportation efficiency. It is a mixture of this data that will allow the companies to save money, reduce emissions, and enhance delivery processes. Lastly, we have Open Data Sources, such as weather forecasts and traffic forecasts, and economic forecasts, these forecasts drive granting of Large-Scale Environments plans and risk management [13]. One of the examples is the integration of weather/traffic-related information that can be utilized to predict potential disruptive events and alternate shipments in an efficient manner. All of the mentioned sources can create an integrated Large Data ecosystem, which will assist the organizations to view the whole process of Large-Scale Environmentss with a more accurate and quicker response.

5. Large Data Architecture

The pipeline of the Large-Scale Environments analysis within DataCo is centred on the architecture of the Large Data framework and it is composed of a layered pipeline to facilitate the data flow process in the pipeline starting with ingestion and up to visualisation. This architecture will ensure that there are accuracy, scalable and efficient processing of large data volumes to be utilized in analysis. The initial layer is the Data Ingestion Layer and it involves the importation of a CSV data of the DataCo dataset using Python scripts, API or batch ingestion app like Apache NiFi or Kafka. This is the step where information is collected and condensed using various sources, including logistics sensors, e-commerce applications, ERP systems, and others and sent to the processing environment.

The processed and raw data are then stored in the next Layer, which is the Storage Layer. Data may be stored in a data warehouse (i.e., PostgreSQL or Snowflake) or in a data lake (i.e., HDFS or AWS S3). Data lakes are suitable when it comes to manipulating all-sized structured and unstructured data without a predefined schema.

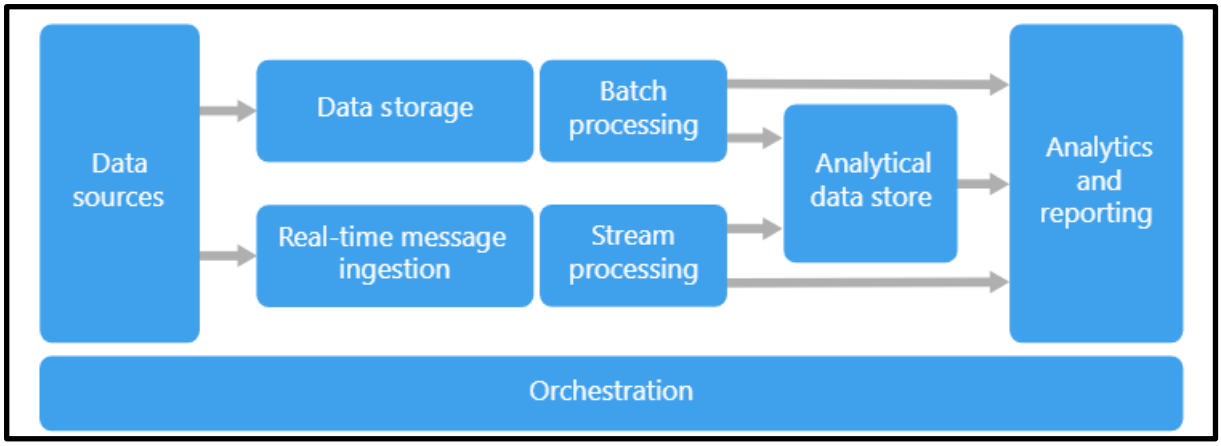


Figure 4: Large Data Architecture Style - Azure Architecture Center

(Source: learn.microsoft.com)

The Processing Layer makes the data clean, transforms, and integrates with the help of such tools as Apache Spark, PySpark, or Pandas. This stage normalizes formats and eliminates duplicates, and gets the dataset ready to be analyzed. Within the Analysis Layer, machine

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

learning and statistical models, i.e., Linear Regression or Random Forest Classifier, are used to forecast the performance of delivery, customer satisfaction, or sales trends. These analytics help businesses to arrive at data-driven decisions [14]. The Visualization Layer transforms the outputs of the analysis to actionable insights via Matplotlib, Seaborn, Power BI, or Tableau dashboards, both of which are useful in the executive decision process. Lastly, the Orchestration and Management Layer is the layer that orchestrates and automates the workflows with the help of such tools as Apache Airflow or Kubernetes, and makes sure that pipelines run without issues and the system remains reliable.

6. Large Data Tools and Technologies

The Large-Scale Environments management analytics based on Large Data is based on a wide range of tools and technologies that perform on the data life cycle- ingestion to visualization. These tools and their uses are shown in the following table:

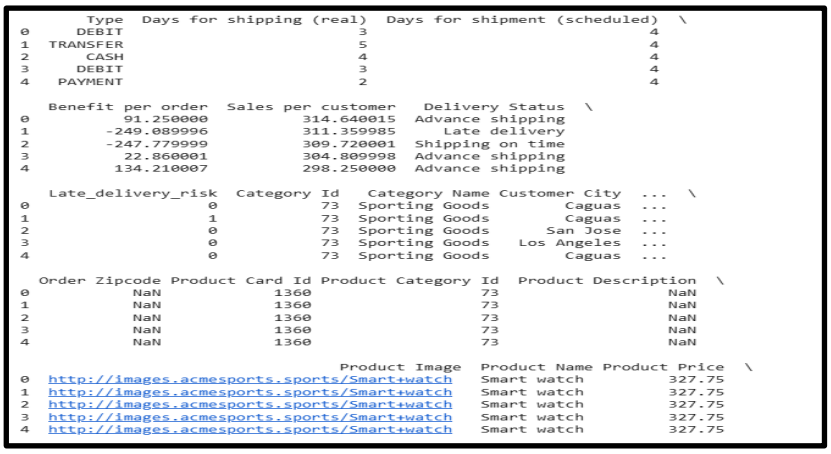
Category	Tools / Technologies	Purpose
Data Ingestion	Apache Kafka, Apache NiFi	Enable real-time or batch data collection from multiple sources such as IoT devices, ERP systems, and e-commerce platforms.
Storage	HDFS, Amazon S3, MongoDB	Provide scalable, distributed storage for structured and unstructured data within data lakes or warehouses.
Processing	Apache Spark, Hadoop, Apache Flink	Facilitate large-scale data transformation, cleansing, and stream/batch analytics.
Analysis	Python, R, TensorFlow	Support predictive and prescriptive analytics using machine learning and deep learning techniques.
Visualization	Power BI, Tableau, D3.js, Matplotlib	Transform analytical results into visual dashboards and insights for managerial decision-making.
Cloud Platforms	AWS, Microsoft Azure, Google Cloud	Offer scalable computing environments, hosting Large Data services and machine learning tools.

The combination of such technologies is the basis of modern Large-Scale Environments analytics. Data ingestion tools collect huge volumes of data consisting of operational systems, and storage structures handle and arrange them effectively. The processing and analysis layers make use of machine learning algorithms in order to infer insights, identify abnormalities, and

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

forecast trends. The insights are then conveyed to the decision-makers through visualization tools that promote efficiency and agility. Lastly, clouds provide elasticity, security, and accessibility globally, which is essential in managing Large-Scale Environmentss that are dynamic and driven by data.

7. Large-Scale Environments data analysis



	Type	Days for shipping (real)	Days for shipment (scheduled)	\
0	DEBIT	3	4	
1	TRANSFER	5	4	
2	CASH	4	4	
3	DEBIT	3	4	
4	PAYMENT	2	4	

	Benefit per order	Sales per customer	Delivery Status	\
0	91.250000	314.640015	Advance shipping	
1	-249.089996	311.359985	Late delivery	
2	-247.779999	309.720001	Shipping on time	
3	22.860001	304.809998	Advance shipping	
4	134.210007	298.250000	Advance shipping	

	Late_delivery_risk	Category Id	Category Name	Customer City	...	\
0	0	73	Sporting Goods	Caguas	...	
1	1	73	Sporting Goods	Caguas	...	
2	0	73	Sporting Goods	San Jose	...	
3	0	73	Sporting Goods	Los Angeles	...	
4	0	73	Sporting Goods	Caguas	...	

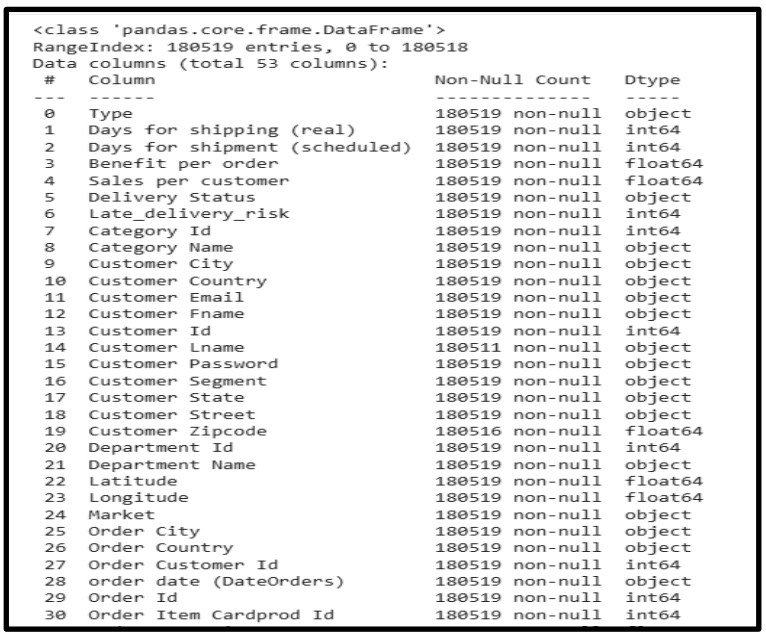
	Order Zipcode	Product Card Id	Product Category Id	Product Description	\
0	NaN	1360	73	NaN	
1	NaN	1360	73	NaN	
2	NaN	1360	73	NaN	
3	NaN	1360	73	NaN	
4	NaN	1360	73	NaN	

	Product Image	Product Name	Product Price	\
0	http://images.acmesports.sports/Smart+watch	Smart watch	327.75	
1	http://images.acmesports.sports/Smart+watch	Smart watch	327.75	
2	http://images.acmesports.sports/Smart+watch	Smart watch	327.75	
3	http://images.acmesports.sports/Smart+watch	Smart watch	327.75	
4	http://images.acmesports.sports/Smart+watch	Smart watch	327.75	

Figure 5: Few Rows of the dataset

(Source: Self-Created)

They consist of the data e-commerce orders which comprises of mode of payment delivered, days of shipment, the delivery, and the product data. It keeps track of learning characteristics of customers and orders such as city, type, sales, and the benefits/order. Each record contains shipping risk, mode, and date. Products like, the name, price and image linkages are also provided. Overall, it helps to analyze the performance within the delivery sphere, customer behavior, and the sales efficiency within online retailing company.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180519 entries, 0 to 180518
Data columns (total 53 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   Type                           180519 non-null  object
 1   Days for shipping (real)       180519 non-null  int64
 2   Days for shipment (scheduled)  180519 non-null  int64
 3   Benefit per order              180519 non-null  float64
 4   Sales per customer             180519 non-null  float64
 5   Delivery Status                180519 non-null  object
 6   Late_delivery_risk             180519 non-null  int64
 7   Category Id                    180519 non-null  int64
 8   Category Name                  180519 non-null  object
 9   Customer City                  180519 non-null  object
 10  Customer Country               180519 non-null  object
 11  Customer Email                 180519 non-null  object
 12  Customer Fname                 180519 non-null  object
 13  Customer Id                    180519 non-null  int64
 14  Customer Lname                 180511 non-null  object
 15  Customer Password              180519 non-null  object
 16  Customer Segment               180519 non-null  object
 17  Customer State                 180519 non-null  object
 18  Customer Street                180519 non-null  object
 19  Customer Zipcode               180516 non-null  float64
 20  Department Id                  180519 non-null  int64
 21  Department Name                180519 non-null  object
 22  Latitude                       180519 non-null  float64
 23  Longitude                      180519 non-null  float64
 24  Market                         180519 non-null  object
 25  Order City                     180519 non-null  object
 26  Order Country                  180519 non-null  object
 27  Order Customer Id              180519 non-null  int64
 28  order date (DateOrders)        180519 non-null  object
 29  Order Id                       180519 non-null  int64
 30  Order Item Cardprod Id         180519 non-null  int64
```

Figure 6: Dataset Information

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

(Source: Self-Created)

This dataset has 180,519 records and 53 columns, which include the data of the customer, order, product, and shipment. It has numeric, nominal fields, as well as date fields. The columns are full with most of them, except the Product Description, which is totally vacant, and Order Zipcode, which is full of nulls. The important fields are order ID, profit, sales, customer information, delivery status, shipping mode, and it is the right field to analyze the sales performance, efficiency in delivery, and customer behavior.

	Days for shipping (real)	Days for shipment (scheduled)	Benefit per order	Sales per customer	Late_delivery_risk	Category Id	Customer Id	Customer Zipcode	Department Id	Latitude	...	Order Item Quantity	Sales	Order Item Total	Order Profit Per Order
count	180519.000000	180519.000000	180519.000000	180519.000000	180519.000000	180519.000000	180519.000000	180516.000000	180519.000000	180519.000000	...	180519.000000	180519.000000	180519.000000	180519.000000
mean	3.497654	2.931847	21.974989	183.107609	0.548291	31.851451	6691.379495	35921.126914	5.443460	29.719955	...	2.127638	203.772096	183.107609	21.974989
std	1.623722	1.374449	104.433526	120.043670	0.497664	15.640064	4162.918106	37542.461122	1.629246	9.813646	...	1.453451	132.273077	120.043670	104.433526
min	0.000000	0.000000	-4274.979980	7.490000	0.000000	2.000000	1.000000	603.000000	2.000000	-33.937553	...	1.000000	9.990000	7.490000	-4274.979980
25%	2.000000	2.000000	7.000000	104.379997	0.000000	18.000000	3258.500000	725.000000	4.000000	18.265432	...	1.000000	119.980003	104.379997	7.000000
50%	3.000000	4.000000	31.520000	163.990005	1.000000	29.000000	6457.000000	19380.000000	5.000000	33.144863	...	1.000000	199.919998	163.990005	31.520000
75%	5.000000	4.000000	64.800003	247.399994	1.000000	45.000000	9779.000000	78207.000000	7.000000	39.279617	...	3.000000	299.950012	247.399994	64.800003
max	6.000000	4.000000	911.799988	1939.989990	1.000000	76.000000	20757.000000	99205.000000	12.000000	48.781933	...	5.000000	1999.989990	1939.989990	911.799988

Figure 7: Descriptive Statistics

(Source: Self-Created)

The statistical summary of the dataset indicates that there are 180,519 records and different numerical fields. The mean shipping time is approximately 3.5 days, with the majority of the orders being made within 3 days. Sales per customer are mean 183.10, profit per order is mean 21.97, with an extensive range of values. The data of latitude and longitude show that it is global. The field of ProductDescription is completely absent, and the value of the status of all the products is 0.

```
Late_delivery_risk                0
Category Id                       0
Category Name                     0
Customer City                     0
Customer Country                  0
Customer Email                    0
Customer Fname                    0
Customer Id                       0
Customer Lname                    8
Customer Password                 0
Customer Segment                  0
Customer State                    0
Customer Street                   0
Customer Zipcode                  3
Department Id                     0
Department Name                   0
Latitude                          0
Longitude                         0
Market                            0
Order City                        0
Order Country                     0
Order Customer Id                 0
order date (DateOrders)           0
Order Id                          0
Order Item Cardprod Id            0
Order Item Discount               0
Order Item Discount Rate          0
Order Item Id                     0
Order Item Product Price          0
Order Item Profit Ratio           0
Order Item Quantity               0
Sales                             0
Order Item Total                  0
Order Profit Per Order            0
Order Region                      0
Order State                       0
Order Status                      0
Order Zipcode                155679
Product Card Id                   0
Product Category Id               0
Product Description          180519
Product Image                     0
Product Name                      0
Product Price                     0
Product Status                    0
shipping date (DateOrders)        0
Shipping Mode                     0
dtype: int64
```

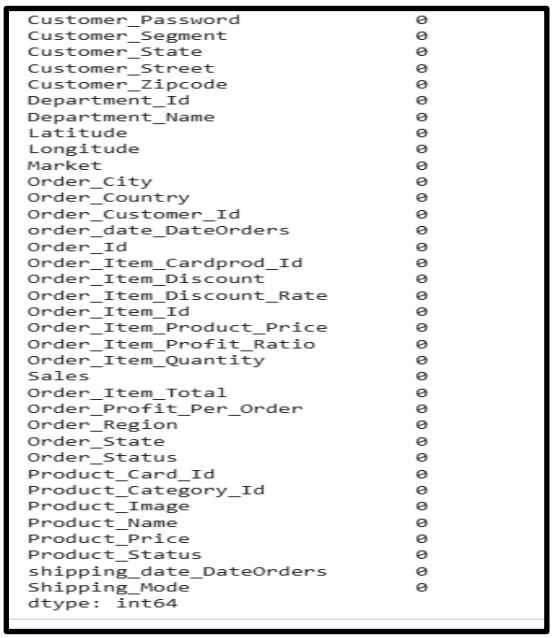
Figure 8: Checking Null Values

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

(Source: Self-Created)

There are very few missing data points in the dataset, with the exception of two large columns. Product Description is empty and contains 180, 519 values equal to null, and Order Zipcode has 155, 679 values equal to null. There are minor blank points in Customer Lname (8 missing) and Customer Zipcode (3 missing). Other fields are filled, meaning that most of the features have good data quality, but zip code and product description should be cleaned or imputed.



```
Customer_Password              0
Customer_Segment               0
Customer_State                 0
Customer_Street                0
Customer_Zipcode               0
Department_Id                  0
Department_Name                0
Latitude                       0
Longitude                      0
Market                         0
Order_City                     0
Order_Country                  0
Order_Customer_Id              0
order_date_DateOrders          0
Order_Id                       0
Order_Item_Cardprod_Id         0
Order_Item_Discount            0
Order_Item_Discount_Rate       0
Order_Item_Id                  0
Order_Item_Product_Price       0
Order_Item_Profit_Ratio        0
Order_Item_Quantity            0
Sales                          0
Order_Item_Total               0
Order_Profit_Per_Order         0
Order_Region                   0
Order_State                    0
Order_Status                   0
Product_Card_Id                0
Product_Category_Id            0
Product_Image                  0
Product_Name                   0
Product_Price                  0
Product_Status                 0
shipping_date_DateOrders       0
Shipping_Mode                  0
dtype: int64
```

Figure 9: Checking Null Values after Cleaning

(Source: Self-Created)

Following data cleaning, all the missing values have been removed or imputed. Each column is filled with all the information and the null values of 180, 519 records. This means that it has clean data that can be further analyzed or modeled. Data quality and consistency have greatly improved, as there are no gaps in the data, and the business intelligence, customer behavior, and delivery performance evaluations can be performed with accurate insights.

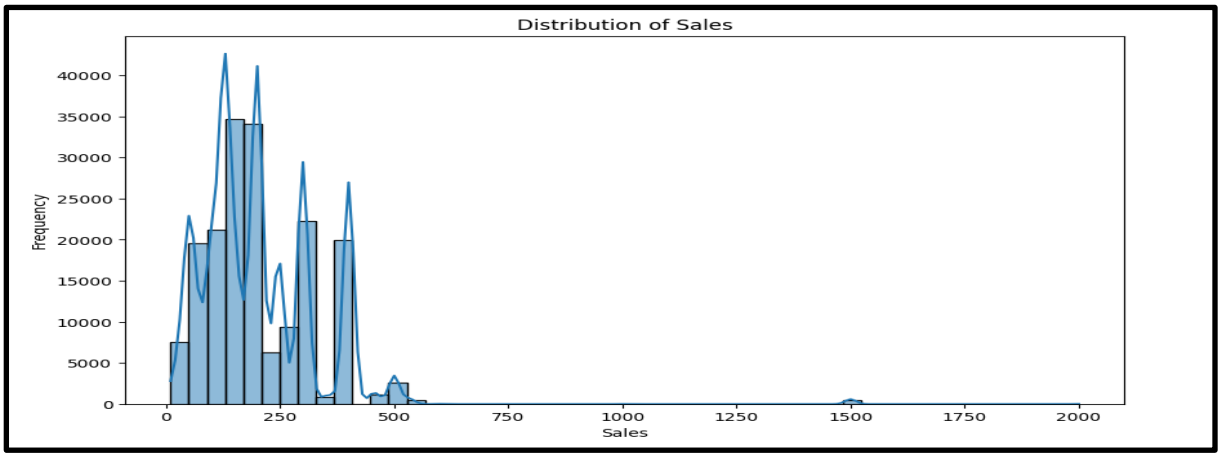


Figure 10: Distribution of Sales

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

(Source: Self-Created)

The histogram indicates the distribution of the sales values, which are much skewed to the right. The majority of the sales are below 500, meaning that there are low to moderate amounts of transaction deals. There are a few extreme values that are higher than 100,0, which are high-value outliers. The density curve also indicates the uneven distribution and indicates that as you have more small sales, the large transactions are fewer. This imbalance can affect the work of the model and has to be normalized or transformed in the course of the data preprocessing.

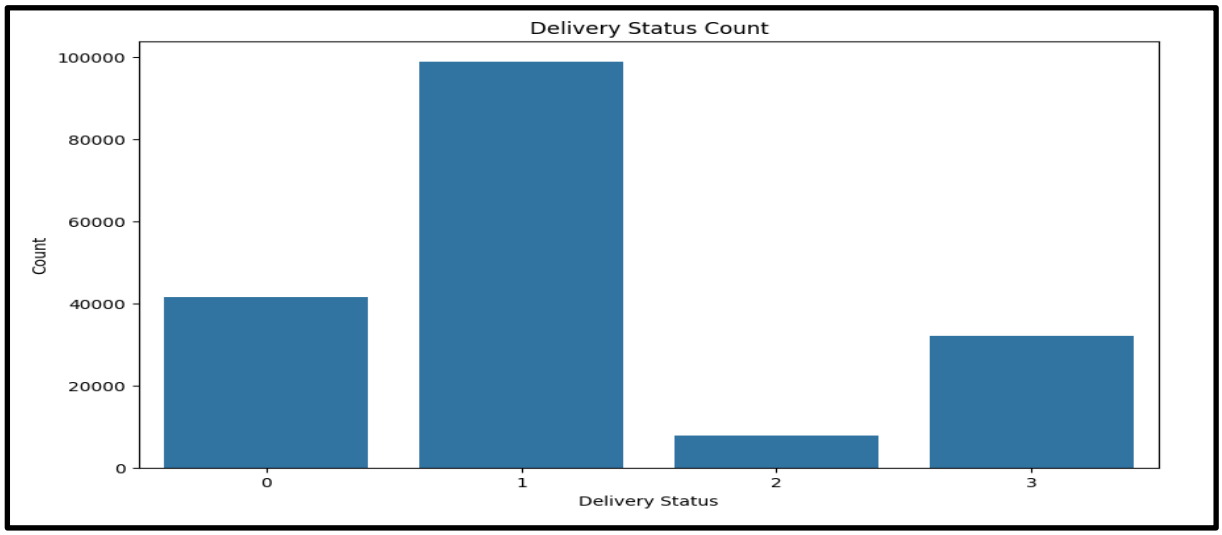


Figure 11: Delivery Status Count

(Source: Self-Created)

In the bar chart, it is demonstrated how often the various delivery statuses occur in the dataset. Most of the orders belong to one superior category status, meaning that most deliveries were timely or usual. The few orders fall under other categories, which may be early, late, or pending shipments. This distribution shows the good performance of good overall delivery, but the imbalance indicates that there can be areas of concentration on the unfocused or late delivery cases.

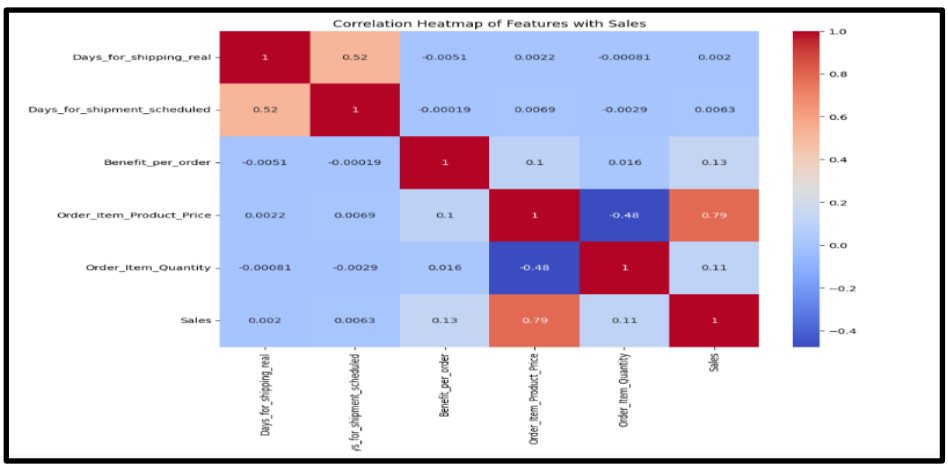


Figure 12: Delivery Status Count

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

(Source: Self-Created)

The heatmap of correlation displays mutual dependencies between the important numerical variables. Order Item Product Price is strongly positively correlated (0.79) with Sales, and this means that the value of sales increases with higher product prices. Shipping days (real) and shipping days (scheduled) have a moderate correlation (0.52). Other attributes, such as Benefit per order and Order item quantity, depict weak correlations, implying that they do not have a significant direct impact on the overall performance of the sales.

```
# Linear Regression for Sales
lr = LinearRegression()
lr.fit(X_train_reg_scaled, y_train_reg)
y_pred_reg = lr.predict(X_test_reg_scaled)

# R² Score
r2 = r2_score(y_test_reg, y_pred_reg)
print("R² Score:", r2)

# Mean Squared Error
mse = mean_squared_error(y_test_reg, y_pred_reg)
print("Mean Squared Error:", mse)

# Root Mean Squared Error
rmse = np.sqrt(mse)
print("Root Mean Squared Error:", rmse)
```

```
R² Score: 0.9243822348930943
Mean Squared Error: 1311.2838807263606
Root Mean Squared Error: 36.211653935250744
```

Figure 13: Linear Regression for Sales

(Source: Self-Created)

The sales prediction of the linear regression model was an excellent predictor. The value of the R 2 = 0.924 shows that 92.4 per cent. The variation in sales can be attributed to the characteristics of the model. The mean squared error of (1311.28) and the root mean squared error are very low, indicating that there are very small errors in prediction. All in all, the model is highly accurate and reliable for predicting sales in this dataset.

```
Random Forest Accuracy: 0.9363228451141148
Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.98      0.97      8282
           1       0.96      0.98      0.97     19797
           2       0.04      0.02      0.03      1558
           3       0.95      0.97      0.96      6467

    accuracy                           0.94     36104
   macro avg       0.73      0.74      0.73     36104
weighted avg       0.92      0.94      0.93     36104

Confusion Matrix:
 [[ 8098     0   184     0]
 [    0 19372   425     0]
 [  332   890    30   306]
 [    0     0   162  6305]]
```

Figure 14: Linear Regression for Sales

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

(Source: Self-Created)

The accuracy of the Random Forest model was high at 93.63% which shows that the overall predictive performance was good. It showed high accuracy and reliability of the classes 0, 1, and 3, which indicates high accuracy of most of the classes. Nevertheless, the performance on class 2 was abnormally poor, with poor accuracy and recall, which means that it is not easy to find cases of minorities or rarity. The confusion matrix indicates that there are not many misclassifications between dominant classes, which indicates that the model is robust. The macro-average F1-score of "0.73" and weighted F1-score of "0.93" also verify the fact that the model is very effective, but it has problems with the lack of balance in classes.

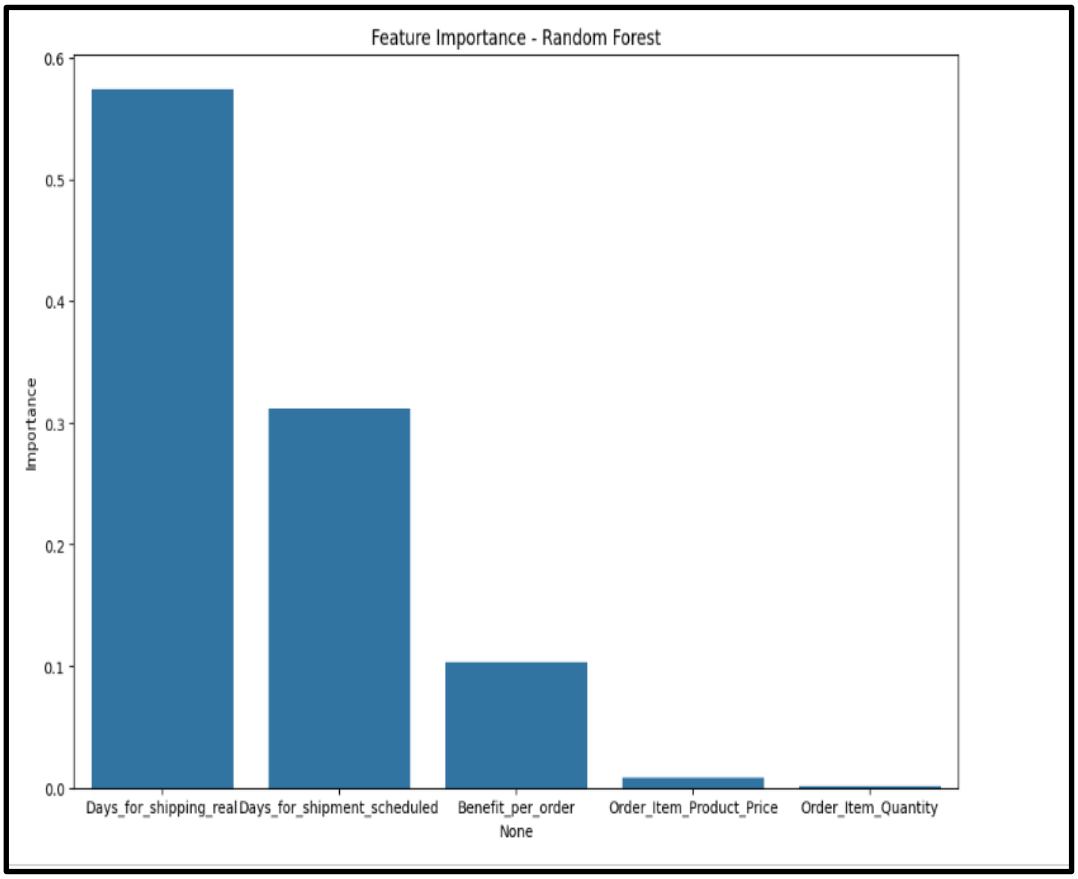


Figure 15: Feature Importance - Random Forest

(Source: Self-Created)

The feature importance chart indicates that the variable with the largest impact on the model decision-making process is the Days_for_shipping_real variable, which contributes close to 60 percent of the decision-making process. Days for shipment are scheduled thereafter with a value of approximately 30% meaning that it is very relevant to the predictions. The significant effect is experienced in Benefit_per_order, although there is a slight contribution by Order_Item_Product_Price and Order_Item_Quantity. It implies that the main determinant of the classification results of the Random Forest model is shipping time and schedule.

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

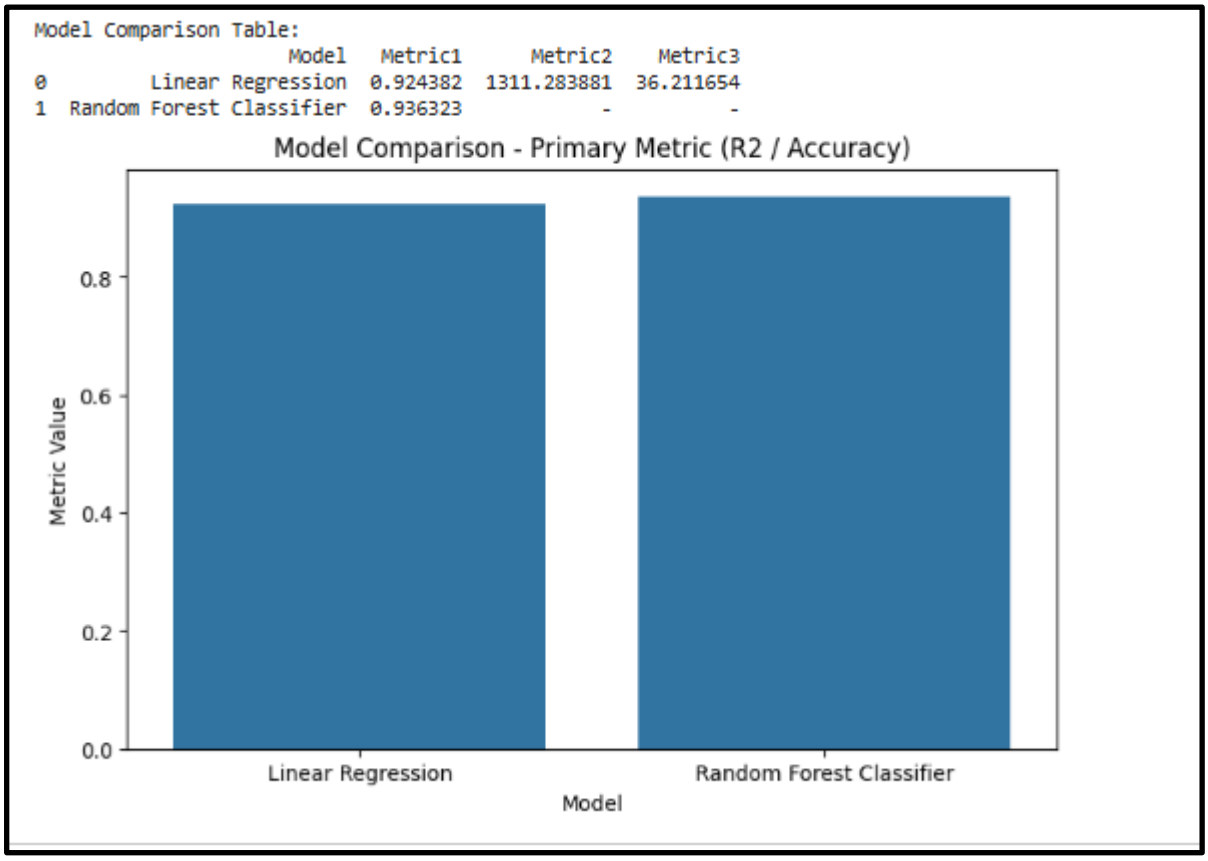


Figure 16: Model Comparison

(Source: Self-Created)

The model chart of comparison reveals that the performance of the Linear Regression and the random Forest Classifier have been pegged on their primary metrics. Linear Regression represented a score reading of 0.924 in the R 2 and the Random Forest Classifier represented a marginally better level of 0.936 in the measure of accuracy. This demonstrates that the two models are very predictive models, yet in general, the Random Forest is a bit higher in the accuracy and strength.

8. Large Data Analytics Approaches

Large Data analytics of Large-Scale Environments Management makes use of a blend of descriptive, predictive and prescriptive approaches to extract actionable information out of large and complex data. The analytics techniques in DataCo Smart Large-Scale Environments Dataset are practical to identify key performance indicators, it is expected to proceed, full of delivery delays or sales, and consequently propose interventions to encourage the efficiency and profitability.

The foundation is descriptive analytics which provides a historical view of the operations as per data summarization, visualization and pattern recognition. The usage of such tools as Pandas and Power BI is designed to analyze the trends in sales, delivery rates, and customer preferences in order to give some idea of the existing operation efficiency.

Received: August 10, 2025

International Journal of Applied Mathematics
Volume 38 No. 11s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Predictive analytics uses models of statistical and machine learning, including Linear Regression and Random Forest, to predict the future based on history. An example is that regression models will forecast the cost of sales and shipping time, and classification models will forecast the delivery status or possible delays. Such methods allow making decisions beforehand and reducing risks.

The most sophisticated step is referred to as prescriptive analytics, which incorporates models of optimization and simulation, and suggests the most optimal actions [15]. Based on the analysis of a variety of constraints (cost, time, and resource availability), algorithms can propose the best shipping schedules, inventory restocking points, or procurement plans.

Combined, these Large Data analytics strategies can help organizations shift to become data-driven. Using these techniques on data such as DataCo, companies will be able to predict more accurately, reduce logistics, and enhance consumer satisfaction, eventually having a wiser and more robust Large-Scale Environments ecosystem.

9. Applications of Large Data in Large-Scale Environments

Large Data has now become part of the optimization of all processes in the Large-Scale Environments Management, including procurement and ultimate delivery. Data analytics is used in the logistics field to optimize routes, maximize fuel efficiency, and track movements in real-time based on the IoT and GPS data. Predictive analytics assist in the forecasting of changes in demand so that there are optimum stock levels and avert stockouts or excessive inventory levels. Large Data can be used in procurement to assist in the evaluation of suppliers and analyzing risks, where a reliable supplier may be identified using performance data and market trends.

Predictive maintenance in manufacturing operations can also be used, where sensor information is used to detect any fault in equipment in advance of its failure to avoid downtime. On the same note, quality control analytics oversees production metrics in order to ensure consistency and minimize defects[16].

Furthermore, combining Large Data with AI and blockchain will create a larger traceability and transparency, which will guarantee the authenticity of products over the global networks. All in all, the Large Data applications enable the organizations to make descriptive decisions, cut operational expenses, and provide their customers with outstanding experiences whilst remaining agile in the face of a swiftly evolving market.

10. Challenges in Large Data Adoption

Although it has potential, numerous technical and organizational challenges to adopting Large Data in the Large-Scale Environmentss. The first is the problem of data integration because the Large-Scale Environments data tends to be a product of different systems, such as ERP, CRM, IoT sensors, and social media platforms. It can be complicated to guarantee compatibility and synchronization of these heterogeneous sources.

The quality and validity of data are also issues, as there can be missing, duplicate, or contradictory records, which result in erroneous conclusions. Also, infrastructure and shortages

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

of skilled workforce are too expensive to adopt, especially among small and medium enterprises [17]. The threat of cybersecurity and data privacy also makes the implementation process even more challenging because sensitive data about suppliers and customers should be secured against attacks.

In addition, the absence of unified frameworks of data governance and interoperability complicates collaboration between global partners of the Large-Scale Environments. To overcome these challenges, businesses ought to invest in cloud computing systems, develop successful policies of data governance, and train professional data specialists on complex analytics to ensure successful, trustworthy, and scalable Large Data.

11. Future Trends in Large-Scale Environments Large Data

Significant IoT networks will constantly feed data on logistics fleets, warehouses, and retail stores, and provide an instant understanding and responsive decision-making. Machine Learning (ML) and artificial intelligence (AI) will be developed to become autonomous Large-Scale Environmentss where predictive and prescriptive models optimize inventory, routing, and procurement with the limited involvement of humans.

The blockchain technology will enhance the integrity and traceability of the data, which will be secured and verifiable records of all the transactions and shipments [18]. In the meantime, the implementation of edge computing will decrease the processing time because there will be data processing nearer to the source, which will be needed during time-sensitive processes, such as cold chain temperature monitoring.

It is also innovation that will be propelled by sustainability since the analytics platforms will be incorporated with carbon footprint monitoring and green logistics optimization to assist in achieving environmental objectives. Lastly, self-service BI tools will democratize the data to enable all employees to access and interpret data on their own, and this will enable a culture of data-driven decision-making to be adopted throughout the Large-Scale Environments ecosystem.

12. Conclusion

The concept of Large Data has changed the face of Large-Scale Environments Management, which has become a proactive, data-driven field, rather than reactive and experience-driven. Combining information provided by various sources, including IoT sensors, ERP solutions, e-compositions, and logistic networks, companies can have a full picture of the whole Large-Scale Environments [19]. Using Large Data analytics allows the company to increase the accuracy of predictions, resource distribution, and overall operational efficiency.

Now it is possible to process large volumes of data, with technologies such as Hadoop, on-the-fly, and cloud data lakes, to make smarter decisions. Predictive and prescriptive analytics can enable organizations to foresee any possible disruption, unearth cost reduction opportunities, and establish responsive strategies to market fluctuations. As it was demonstrated with the help of the DataCo Smart Large-Scale Environments Dataset, machine learning modeling, including Linear Regression and Random Forest results, could be successfully utilized to predict the level

Received: August 10, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

of sales performance and the delivery outcomes, which will give the organizations an opportunity to act best and increase customer satisfaction levels.

However, not everything is smooth on the road to full integration of Large Data. The challenges to mass adoption have been such obstacles as data quality, privacy, interoperability, and/or infrastructure expenses [20]. Good data governance policies, human resources, and investments in scalable systems including cloud and edge computing are a way to address the challenges in the areas.

SCM will be increasingly digitalized in the future with traceability using blockchain technology, real-time analytics, and automation using AI. These technologies will create more intelligent, open and sustainable Large-Scale Environments ecosystems. Lastly, strategic utilisation of the Large Data analytics will positively increase the operational efficiency, along with resilience, innovation, and competitive advantage in a vibrant global market. Organizations can do this by transforming their Large-Scale Environmentss to be smart systems that will propel them to expansion and long run success in business through data as a competitive advantage.

References

[1] Elkliny, A., Mahmoudi, A. and Deng, X., 2025. Large Data-Driven Implementation in International Construction Large-Scale Environments Management: Framework Development, Future Directions, and Barriers. Buildings, 15(13), p.2167.

[2] Zhou, C., Stephen, A., Cao, X. and Wang, S., 2021. A data-driven business intelligence system for large-scale semi-automated logistics facilities. International Journal of Production Research, 59(8), pp.2250-2268.

[3] Tan, Y., Gu, L., Xu, S. and Li, M., 2024. Large-Scale Environments Inventory Management from the Perspective of “Cloud Large-Scale Environments”—A Data Driven Approach. Mathematics, 12(4), p.573.

[4] Tan, D., Su, Y., Peng, X., Chen, H., Zheng, C., Zhang, X. and Zhong, W., 2023. Large-scale data-driven optimization in deep modeling with an intelligent decision-making mechanism. IEEE Transactions on Cybernetics, 54(5), pp.2798-2810.

[5] Selvarajan, G.P., 2022. Leveraging SnowflakeDB in Cloud Environments: Optimizing AI-driven Data Processing for Scalable and Intelligent Analytics. International Journal of Enhanced Research in Science, Technology & Engineering, 11(11), pp.257-264.

[6] Onukwulu, E.C., Agho, M.O. and Eyo-Udo, N.L., 2023. Developing a framework for AI-driven optimization of Large-Scale Environmentss in energy sector. Global Journal of Advanced Research and Reviews, 1(2), pp.82-101.

[7] Bechtsis, D., Tsolakis, N., Iakovou, E. and Vlachos, D., 2022. Data-driven secure, resilient and sustainable Large-Scale Environmentss: gaps, opportunities, and a new generalised data sharing and data monetisation framework. International Journal of Production Research, 60(14), pp.4397-4417.

Received: August 10, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[8] Ikegwu, A.C., Nweke, H.F., Anikwe, C.V., Alo, U.R. and Okonkwo, O.R., 2022. Large Data analytics for data-driven industry: a review of data sources, tools, challenges, solutions, and research directions. Cluster Computing, 25(5), pp.3343-3387.

[9] Ekundayo, F., 2024. Leveraging AI-driven decision intelligence for complex systems engineering. Int J Res Publ Rev, 5(11), pp.1-10.

[10] Jahin, M.A., Shovon, M.S.H., Shin, J., Ridoy, I.A. and Mridha, M.F., 2023. Large Data-Large-Scale Environments management framework for forecasting: Data preprocessing and machine learning techniques. arXiv preprint arXiv:2307.12971.

[11] Nabeel, M.Z., 2024. AI-enhanced project management systems for optimizing resource allocation and risk mitigation: Leveraging Large Data analysis to predict project outcomes and improve decision-making processes in complex projects. Asian Journal of Multidisciplinary Research & Review, 5(5), pp.53-65.

[12] Hosen, M.S., Islam, R., Naeem, Z., Folorunso, E.O., Chu, T.S., Al Mamun, M.A. and Orunbon, N.O., 2024. Data-driven decision making: Advanced database systems for business intelligence. Nanotechnology Perceptions, 20(3), pp.687-704.

[13] Niu, Y., Ying, L., Yang, J., Bao, M. and Sivaparthipan, C.B., 2021. Organizational business intelligence and decision making using Large Data analytics. Information Processing & Management, 58(6), p.102725.

[14] Kommisetty, P.D.N.K. and Dileep, V., 2022. Leading the future: Large Data solutions, cloud migration, and AI-driven decision-making in modern enterprises. Educational Administration: Theory and Practice, 28(03), pp.352-364.

[15] Abaku, E.A., Edunjobi, T.E. and Odimarha, A.C., 2024. Theoretical approaches to AI in Large-Scale Environments optimization: Pathways to efficiency and resilience. International Journal of Science and Technology Research Archive, 6(1), pp.092-107.

[16] Ji, E., Wang, Y., Xing, S. and Jin, J., 2025. Hierarchical reinforcement learning for energy-efficient API traffic optimization in large-scale advertising systems. IEEE Access.

[17] Lazaroiu, G., Androniceanu, A., Grecu, I., Grecu, G. and Neguriță, O., 2022. Artificial intelligence-based decision-making algorithms, Internet of Things sensing networks, and sustainable cyber-physical management systems in Large Data-driven cognitive manufacturing. Oeconomia Copernicana, 13(4), pp.1047-1080.

[18] Mourtzis, D., 2021. Towards the 5th industrial revolution: A literature review and a framework for process optimization based on Large Data analytics and semantics. Journal of Machine Engineering, 21(3), pp.5-39.

[19] Azad, M.A., 2025. Leveraging Large-Scale Environments analytics for real-time decision making in apparel manufacturing. Authorea Preprints.

[20] Jahin, M.A., Shovon, M.S.H., Shin, J., Ridoy, I.A. and Mridha, M.F., 2023. Large Data-Large-Scale Environments management framework for forecasting: Data preprocessing and machine learning techniques. arXiv preprint arXiv:2307.12971.

Received: August 10, 2025