

# **Customer Credit and Buying Profile in E-commerce**

**A PROJECT REPORT**

*Submitted by,*

**Mr. Arnav Gupta - 20201CEI0158**  
**Mr. R. Jatin Kumar -20201CEI0155**  
**Ms. Kumari Puja Sharma -20201CEI0180**  
**Ms. Chowdam Likitha -20201CEI0175**  
**Mr. Viraj Mahavir Magdum -20201CEI0109**

*Under the guidance of,*  
**Dr. Pallavi M**

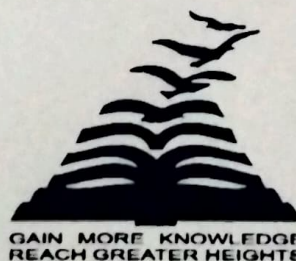
*in partial fulfillment for the award of the degree of*

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER ENGINEERING [ARTIFICIAL INTELLIGENCE AND  
MACHINE LEARNING]**

**At**



**PRESIDENCY UNIVERSITY**  
**BENGALURU**  
**JANUARY 2024**

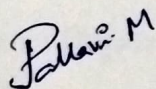


# PRESIDENCY UNIVERSITY

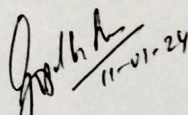
## SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

### CERTIFICATE

This is to certify that the Project report “Customer Credit and Buying Profile in E-commerce” being submitted by “Arnav Gupta, R. Jatin Kumar, Kumari Puja Sharma, Chowdam Likitha, Viraj Mahavir Magdum” bearing roll number(s) “20201CEI0158, 20201CEI0155, 20201CEI0180, 20201CEI0175, 20201CEI0109” in partial fulfilment of requirement for the award of degree of Bachelor of Technology in Computer Engineering [Artificial Intelligence and Machine Learning] is a bonafide work carried out under my supervision.



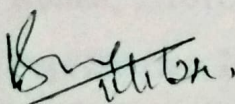
**Dr. Pallavi M.**  
Asst. Professor  
School of CSE & IS  
Presidency University



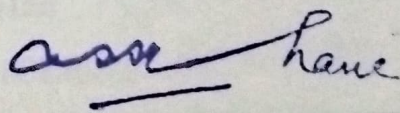
**Dr. Gopal K. Shyam**  
Prof. & HoD  
School of CSE  
Presidency University



**Dr. C. KALAIARASAN**  
Associate Dean  
School of CSE&IS  
Presidency University



**Dr. L. SHAKKEERA**  
Associate Dean  
School of CSE&IS  
Presidency University



**Dr. Md. SAMEERUDDIN KHAN**  
Dean  
School of CSE&IS  
Presidency University



# PRESIDENCY UNIVERSITY

## SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

### DECLARATION

We hereby declare that the work, which is being presented in the project report entitled **Customer Credit and Buying Profile in E-commerce** in partial fulfilment for the award of Degree of **Bachelor of Technology in Computer Engineering [Artificial Intelligence and Machine Learning]**, is a record of our own investigations carried under the guidance of **Dr. Pallavi M, Assistant professor, School of Computer Science & Engineering, Presidency University, Bengaluru.**

We have not submitted the matter presented in this report anywhere for the award of any other Degree.

**Arnav Gupta, 20201CEI0158**

*Arnav Gupta*

**R. Jatin Kumar, 20201CEI0155**

*R. Jatin Kumar*

**Kumari Puja Sharma, 20201CEI0180**

*Kumari Puja Sharma*

**Chowdam Likitha, 20201CEI0175**

*C. Likitha*

**Viraj Mahavir Magdum, 20201CEI0109**

*Viraj*

## **ABSTRACT**

The advent of e-commerce has ushered in a paradigm shift in the way businesses operate, presenting both opportunities and challenges. In this dynamic landscape, challenges such as returns, cancellations, and the diverse array of payment preferences have become pivotal considerations for businesses seeking to ensure optimal customer satisfaction. This paper addresses these challenges through the proposition of an innovative methodology that seamlessly integrates big data analytics, advanced Machine Learning Techniques, and rigorous statistical analysis.

Central to this methodology is the construction of detailed and nuanced customer profiles, achieved through the aggregation and anonymization of vast datasets. By harnessing the power of big data, businesses can gain profound insights into customer behavior, preferences, and creditworthiness. The emphasis on individual privacy remains a cornerstone of this approach, with a commitment to anonymization techniques that safeguard sensitive information.

The system devised through this methodology goes beyond mere profiling; it facilitates the generation of personalized credit and purchasing profiles for individual customers. This data-driven approach offers a multitude of benefits, including the ability to tailor delivery options and provide Equated Monthly Installment (EMI) choices that align with the financial capacities and preferences of customers. Notably, the system categorizes customers based on their credit and purchasing behavior, enabling strategic decisions such as limiting Cash on Delivery for users with unfavorable buying profiles and withholding EMI choices for those with less favorable credit profiles.

One of the key distinguishing features of this solution is its commitment to privacy protection. In optimizing e-commerce operations, the system carefully navigates the delicate balance between customization and confidentiality. By prioritizing privacy, businesses can build trust with their customers while harnessing the analytical power of Machine Learning to enhance service offerings.

In conclusion, this paper introduces a comprehensive and ethically grounded approach to tackling the challenges of e-commerce. By fusing big data analytics, Machine Learning Techniques, and privacy-preserving methodologies, businesses can not only navigate complexities related to returns, cancellations, and payment preferences but also elevate the customer experience through personalized services. This innovative methodology represents a forward-looking solution that harmonizes the imperatives of data-driven decision-making and the ethical considerations inherent in preserving individual privacy in the digital age.

## ACKNOWLEDGEMENT

First of all, we indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time.

We express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, Dean, School of Computer Science and Engineering & School of Information Science, Presidency University for getting us permission to undergo the project.

We record our heartfelt gratitude to our beloved Associate **Deans Dr. Kalaiarasan C and Dr. Shakkeera L**, School of Computer Science and Engineering & School of Information Science, Presidency University and Dr. Gopal K. Shyam, Head of the Department, School of Computer Science and Engineering, Presidency University for rendering timely help for the successful completion of this project.

We are greatly indebted to our guide Dr. Pallavi M., Asst. Professor, School of Computer Science and Engineering, Presidency University for her inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the project work.

We would like to convey our gratitude and heartfelt thanks to the University Project-II **Coordinators Dr. Sanjeev P Kaulgud, Dr. Mrutyunjaya MS** and also the department Project Coordinators Ms. Yogeetha B R, Dr. Sudha P, and Dr. Sasidhar Babu.

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

**Arnav Gupta**

**R. Jatin Kumar**

**Kumari Puja Sharma**

**Chowdam Likitha**

**Viraj Mahavir Magdum**

## LIST OF TABLES

Sl. No.	Table Name	Table Caption	Page No.
1	Table 1	Initial Dataset	11

## LIST OF FIGURES

Sl. No.	Figure Name	Caption	Page No.
1	Figure 1	Initial Data frame (A)	12
2	Figure 2	Initial Data frame (B)	13
3	Figure 3	Final Data frame	14
4	Figure 4	Timeline	19
5	Figure 5	Analysis of KNN with outliers	35
6	Figure 6	Analysis of Decision Tree with outliers	35
7	Figure 7	Analysis of SVM with outliers	36
8	Figure 8	Analysis of Random Forest with outliers	36
9	Figure 9	Analysis of KNN without outliers	37
10	Figure 10	Analysis of Decision Tree without outliers	37
11	Figure 11	Analysis of Random Forest without outliers	38
12	Figure 12	Analysis of SVM without outliers	38
13	Figure 13	Analysis of Random Forest with Polynomial Features	39



## **TABLE OF CONTENTS**

<b>CHAPTER NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
	<b>CERTIFICATE</b>	<b>ii</b>
	<b>DECLARATION</b>	<b>iii</b>
	<b>ABSTRACT</b>	<b>iv- v</b>
	<b>ACKNOWLEDGEMENT</b>	<b>vi</b>
<b>1.</b>	<b>INTRODUCTION</b>	<b>2</b>
<b>2.</b>	<b>LITERATURE SURVEY</b>	<b>4</b>
<b>3.</b>	<b>RESEARCH GAPS OF EXISTING METHODS</b>	<b>7</b>
	3.1 Algorithmic Transparency and Interpretability	7
	3.2 Incorporation of Advanced Machine Learning Techniques	7
	3.3 Dynamic and Adaptive Models	7
	3.4 Integration of Alternative Data Sources	8
	3.5 Ethical Considerations and Bias Mitigation	8
	3.6 Real-Time Processing and Adaptation	8
	3.7 Evaluation Metrics and Benchmarking	9
	3.8 Generalizability Across Industries and Regions	9
	3.9 Privacy-Preserving Techniques	9
	3.10 User-Centric Approaches	9
<b>4.</b>	<b>PROPOSED METHODOLOGY</b>	<b>11</b>
	4.1 Data Preprocessing Methods	11
	4.2 Feature Extraction Methods	12
	4.3 Outlier Detection and Removal	13
<b>5.</b>	<b>OBJECTIVES</b>	<b>15</b>

<b>6.</b>	<b>SYSTEM DESIGN &amp; IMPLEMENTATION</b>	<b>17</b>
<b>7.</b>	<b>TIMELINE FOR EXECUTION OF PROJECT</b>	<b>19</b>
<b>8.</b>	<b>OUTCOMES</b>	<b>20</b>
<b>9.</b>	<b>RESULTS AND DISCUSSIONS</b>	<b>22</b>
	9.1 Performance Analysis	22
	9.2 Outlier Sensitivity	22
	9.3 Feature Engineering Impact	23
	9.4 Model Robustness	23
	9.5 Practical Implications	23
	9.6 Future Directions	23
<b>10.</b>	<b>CONCLUSION</b>	<b>25</b>
<b>11.</b>	<b>REFERENCES</b>	<b>26</b>
<b>12.</b>	<b>APPENDIX-A: PSEUDOCODE</b>	<b>28</b>
<b>13.</b>	<b>APPENDIX-B: SCREENSHOTS</b>	<b>32</b>
<b>14.</b>	<b>APPENDIX-C: ENCLOSURES</b>	<b>37</b>

## **CHAPTER-1**

### **INTRODUCTION**

In response to the dynamic challenges posed by the evolving landscape of e-commerce, this research introduces a cutting-edge system designed to address the intricacies of personalized credit and purchasing profiles. The overarching objective is to empower e-commerce platforms with a sophisticated tool that harnesses the capabilities of advanced technologies, including big data analytics, artificial intelligence, and statistical analysis. By delving into these technological frontiers, the system aims to unravel profound insights into customer behaviors, allowing businesses to tailor their services with a heightened degree of precision.

A central tenet of the proposed system is the emphasis on individualized profiles, seeking to enhance the overall shopping experience for customers. The approach goes beyond the conventional understanding of customer segmentation, aiming to create nuanced profiles that reflect not only purchasing patterns but also creditworthiness. By doing so, businesses can strategically customize their offerings, such as devising tailored delivery benefits and Equated Monthly Installment (EMI) options, aligning with the financial proclivities of individual customers.

Crucially, amidst the drive for enhanced personalization, the research underscores a paramount commitment to customer privacy. The system incorporates state-of-the-art anonymization and aggregation techniques to ensure that sensitive customer information is safeguarded. Recognizing the ethical dimensions of deploying such technology, the research places a strong emphasis on transparency, user consent, and compliance with privacy regulations. In doing so, it seeks to establish a model that not only delivers technological sophistication but also aligns with the ethical imperatives of the digital age.

Furthermore, the research is attuned to the complexities associated with navigating the legal landscape surrounding data privacy and protection. It acknowledges the need to operate within established legal frameworks and advocates for a principled approach to technology deployment. By recognizing the potential pitfalls and challenges, the research endeavors to provide a comprehensive roadmap for e-commerce platforms, guiding them towards the implementation of a privacy-respecting yet highly effective system for customer profiling and service customization.

In conclusion, this research serves as a beacon for businesses navigating the nuanced terrain of e-commerce. Through its innovative system, it not only offers a technological edge in understanding customer behaviors but also sets a benchmark for ethical considerations in the realm of personalization. By prioritizing privacy, transparency, and legal compliance, the proposed system represents a pioneering solution that harmonizes technological prowess with ethical responsibility, ensuring a sustainable and customer-centric future for e-commerce.



## **CHAPTER-2**

### **LITERATURE SURVEY**

Credit scoring plays a pivotal role in the financial industry, aiding in risk assessment and decision-making processes. Wang and Yang (2020)[1] contribute to this field by achieving a high accuracy of 85.71% in credit scoring for bank credit cards using an unspecified machine learning (ML) model. While the specific algorithm remains undisclosed, the remarkable result suggests the effectiveness of ML in financial risk assessment. In a similar vein, Zhang et al. (2020)[2] focus on constructing a credit scoring model by leveraging multiple data sources, including personal information and credit card data. Although accuracy reports are lacking, the exploration of multi-source data holds the potential to enhance model accuracy and provide a more comprehensive financial picture.

Maheshwari et al. (2019)[3] delve into the credit behavior of e-commerce customers, presenting an intriguing avenue for credit profiling. The study focuses on utilizing the RFM strategy and advanced K-means clustering for effective e-commerce customer segmentation. The proposed methodology involves data cleaning, RFM-based credit point calculation, and classification into customer categories. Support Vector Machine (SVM) and Random Forest algorithms demonstrate high accuracy, reaching 87% and 99%, respectively. The approach aims to minimize losses by identifying genuine and fraud customers.

Customer segmentation is a crucial aspect of understanding and targeting diverse customer groups effectively. Wu et al. (2022)[4] introduce an improved k-medoids clustering algorithm for e-commerce customer segmentation, achieving moderate clustering quality with a silhouette coefficient of 0.585. This underscores the potential of customized clustering approaches to outperform standard algorithms in the context of e-commerce. Dursun and Caber (2016)[5] apply RFM (Recency, Frequency, Monetary) and hierarchical clustering to

segment hotel customers, showcasing the effectiveness of traditional methods in specific contexts, even though accuracy metrics are absent.

Christy et al. (2021)[6] utilize RFM ranking for segmentation, but the unspecified ML methods and lack of accuracy data limit a comprehensive understanding of their approach. This highlights the ongoing challenge of integrating ML to refine segmentation based on RFM in e-commerce. Dogan et al. (2018)[7] and Kabasakal (2020)[10] similarly employ RFM and k-means for retail and e-retail segmentation, respectively, without providing accuracy metrics, suggesting the enduring applicability of these methods in offline and online retail environments.

In contrast, Cheng and Chen (2009)[8] achieve an impressive accuracy of 85.6% by combining RFM with rough set theory for customer value segmentation. This emphasizes the potential of hybrid approaches for improved segmentation and customer value assessment, shedding light on the effectiveness of integrating different methodologies. Brahmana et al. (2020)[9] conduct a comparative analysis of clustering algorithms, concluding that k-means performs best for RFM-based segmentation, providing valuable insights into selecting the most suitable algorithm based on data characteristics.

Expanding the scope of customer segmentation, Firdaus and Utama (2021)[11] develop a bank customer segmentation model using RFM+B (Behavior), incorporating broader behavioral data to expand the traditional RFM approach. Although details on ML techniques and accuracy are lacking, their work suggests the importance of incorporating diverse behavioral data for a more comprehensive understanding of customer segments.

Hossain et al. (2023)[12] explore the role of technology in boosting e-commerce adoption for small and medium enterprises (SMEs). Analyzing factors such as ICT adoption,

internet connectivity, and business data management, they find that these factors significantly promote e-commerce success for SMEs. The paper underscores the need for further research on organizational and environmental factors influencing SME e-commerce adoption, offering valuable insights for SMEs and policymakers on leveraging technology for growth in the digital marketplace.

In summary, the literature review highlights the diverse approaches to credit scoring and customer segmentation, emphasizing the potential of machine learning in enhancing accuracy and refining segmentation methods. While certain studies present impressive results, the field continues to evolve, with opportunities for future research to explore advanced algorithms, multi-source data integration, and the integration of behavioral data for a more nuanced understanding of customer behaviors in the context of e-commerce.

## **CHAPTER-3**

### **RESEARCH GAPS OF EXISTING METHODS**

Despite the advancements in credit scoring and customer segmentation methods discussed in the literature review, several research gaps persist, presenting opportunities for further exploration and improvement:

#### **3.1. Algorithmic Transparency and Interpretability:**

While machine learning models, such as those used in credit scoring, often demonstrate high accuracy, the lack of transparency and interpretability remains a critical concern. Many models operate as "black boxes," making it challenging to understand the decision-making process. Future research could focus on developing more interpretable models to enhance trust and regulatory compliance.

#### **3.2. Incorporation of Advanced Machine Learning Techniques:**

Existing studies primarily utilize traditional machine learning techniques, such as k-means clustering and decision trees. There is a gap in exploring more advanced and complex machine learning algorithms, such as deep learning and ensemble methods, which may provide superior performance in capturing intricate patterns within e-commerce datasets.

#### **3.3. Dynamic and Adaptive Models:**



Credit scoring models often assume static conditions, neglecting the dynamic nature of customer behavior and financial markets. Future research could delve into developing dynamic and adaptive models that can continuously learn and adjust to evolving customer preferences and market dynamics, enhancing the robustness of credit scoring systems.

### **3.4. Integration of Alternative Data Sources:**

The majority of studies focus on conventional data sources, such as credit card data and personal information. There is a gap in research exploring the integration of alternative data sources, such as social media activity, online behavior, or non-traditional financial data, to provide a more comprehensive and real-time assessment of customer creditworthiness.

### **3.5. Ethical Considerations and Bias Mitigation:**

Ethical considerations, including fairness and bias in credit scoring, remain under-addressed. Future research should investigate methods to mitigate biases and ensure fairness, especially concerning protected characteristics like race, gender, or socioeconomic status, to prevent discriminatory outcomes and promote ethical AI practices.

### **3.6. Real-Time Processing and Adaptation:**

The evolving nature of e-commerce demands real-time processing capabilities. Current models may not adequately address the need for instant decision-making. Future research could focus on developing models that can process and adapt to new

data in real-time, providing timely insights for credit scoring and customer segmentation in dynamic e-commerce environments.

### **3.7. Evaluation Metrics and Benchmarking:**

There is a lack of standardized evaluation metrics and benchmarking procedures across studies, making it challenging to compare the effectiveness of different models. Future research should establish standardized metrics for credit scoring and customer segmentation, facilitating better comparisons and providing clearer insights into model performance.

### **3.8. Generalizability Across Industries and Regions:**

Many studies focus on specific industries or regions, limiting the generalizability of their findings. Future research could explore the development of models that are adaptable across various e-commerce sectors and geographical locations, considering the diverse factors that influence customer behaviour and creditworthiness.

### **3.9. Privacy-Preserving Techniques:**

With increasing concerns about data privacy, there is a need for research on privacy-preserving techniques in credit scoring and customer segmentation. Developing methods that can extract valuable insights from sensitive data while ensuring individual privacy and compliance with regulations is crucial for the ethical deployment of these systems.

### **3.10. User-Centric Approaches:**

Incorporating user-centric perspectives in the design and evaluation of credit scoring models and customer segmentation strategies is often overlooked. Future research should involve end-users in the development process to ensure that the models align with customer expectations and provide user-friendly experiences.

Addressing these research gaps can contribute to the refinement and advancement of credit scoring and customer segmentation methods, making them more reliable, ethical, and adaptable to the evolving landscape of e-commerce.

## CHAPTER-4

### PROPOSED METHODOLOGY

#### 4.1 Data Preprocessing Methods

- **Concatenation:** The data from two distinct CSV files, d1.csv and d2.csv shown in Table [1], Figure. [1] and [2], is loaded into separate dataframes (d1 and d2). These dataframes are concatenated along the rows using the `pd.concat` function, resulting in the creation of the combined dataframe `df`.
- **Handling Missing Values:** The `df` dataframe is examined for missing values using the `df.isnull().sum()` function. Rows containing missing values are then eliminated from the dataframe using `df = df.dropna()`.
- **Date Conversion and Time-Based Feature Engineering:** The 'InvoiceDate' column is converted to datetime format using `pd.to_datetime`. Subsequently, new features related to recency, last purchase date, purchase period, and frequency are engineered based on date information.

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	12/1/2009 7:45	6.95	13085.0	United Kingdom
1	489434	79323P	PINK CHERRY LIGHTS	12	12/1/2009 7:45	6.75	13085.0	United Kingdom
2	489434	79323W	WHITE CHERRY LIGHTS	12	12/1/2009 7:45	6.75	13085.0	United Kingdom
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	12/1/2009 7:45	2.10	13085.0	United Kingdom
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	12/1/2009 7:45	1.25	13085.0	United Kingdom
...	...	...	...	...	...	...	...	...
1067366	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0	France
1067367	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0	France
1067368	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0	France
1067369	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0	France
1067370	581587	POST	POSTAGE	1	12/9/2011 12:50	18.00	12680.0	France

TABLE 1 – INITIAL DATASET



## 4.2 Feature Extraction Methods

- **Grouping and Aggregation:** Data is grouped by relevant columns such as 'Customer ID', 'Invoice', 'DaysSinceLastPurchase', and 'Frequency'. Aggregation is performed on the 'TotalCost' column using the sum() function.
- **Return Count Calculation:** The total return count for each customer is computed, and this information is merged back into the original dataframe.
- **Payment Method Frequency:** A new feature 'paymtd' is calculated as the frequency of the payment method for each customer.
- **Credit Score Calculation (c\_s):** The credit score ('c\_s') is calculated for each customer using features such as Recency, TotalCost, Frequency, paymtd, and Return\_Count.
- **Categorization into Classes:** A new feature 'Class' is derived by categorizing customers based on quantiles of the credit score ('c\_s').

	Invoice	StockCode	Description	Quantity
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12
1	489434	79323P	PINK CHERRY LIGHTS	12
2	489434	79323W	WHITE CHERRY LIGHTS	12
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24
...	...	...	...	...
1067366	581587	22899	CHILDREN'S APRON DOLLY GIRL	6
1067367	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4
1067368	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4
1067369	581587	22138	BAKING SET 9 PIECE RETROSPOT	3
1067370	581587	POST	POSTAGE	1

	InvoiceDate	Price	Customer ID	Country
0	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
1	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
2	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
3	2009-12-01 07:45:00	2.10	13085.0	United Kingdom
4	2009-12-01 07:45:00	1.25	13085.0	United Kingdom
...	...	...	...	...
1067366	2011-12-09 12:50:00	2.10	12680.0	France
1067367	2011-12-09 12:50:00	4.15	12680.0	France
1067368	2011-12-09 12:50:00	4.15	12680.0	France
1067369	2011-12-09 12:50:00	4.95	12680.0	France
1067370	2011-12-09 12:50:00	18.00	12680.0	France

**FIGURE 1 – INITIAL DATAFRAME(A)**

	DaysSinceLastPurchase	LastPurchaseDate	PurchasePeriod	Frequency	\
0	168	2011-07-05 12:11:00	581	58.1	
1	168	2011-07-05 12:11:00	581	58.1	
2	168	2011-07-05 12:11:00	581	58.1	
3	168	2011-07-05 12:11:00	581	58.1	
4	168	2011-07-05 12:11:00	581	58.1	
...	...	...	...	...	
1067366	11	2011-12-09 12:50:00	112	28.0	
1067367	11	2011-12-09 12:50:00	112	28.0	
1067368	11	2011-12-09 12:50:00	112	28.0	
1067369	11	2011-12-09 12:50:00	112	28.0	
1067370	11	2011-12-09 12:50:00	112	28.0	
	TotalCost				
0	83.40				
1	81.00				
2	81.00				
3	100.80				
4	30.00				
...	...				
1067366	12.60				
1067367	16.60				
1067368	16.60				
1067369	14.85				
1067370	18.00				

FIGURE 2 – INITIAL DATAFRAME(B)

### 4.3 Outlier Detection and Removal

- **Z-Score and IQR Methods:** Z-Score and Interquartile Range (IQR) methods are applied to identify outliers in specific columns, including 'Recency', 'Frequency', 'paymtd', 'Return\_Count', and 'TotalCost'. Rows identified as outliers are subsequently removed from the dataframe, resulting in the creation of the final processed dataframe df4 as shown in Figure. [3].

This comprehensive methodology ensures the effective handling of data preprocessing challenges, meaningful feature extraction, and the removal of outliers, setting the foundation for subsequent analyses and modeling tasks.

	Customer ID	Recency	Frequency	paymtd	Return_Count	TotalCost	Class
2	12348.0	0.914	72.400000	0.4	0	2019.40	2
3	12349.0	0.971	143.200000	0.8	1	4404.54	3
4	12350.0	0.679	0.000000	1.0	0	334.40	1
5	12351.0	0.614	0.000000	0.0	0	300.93	0
7	12353.0	0.785	102.000000	0.5	0	406.76	1
...	...	...	...	...	...	...	...
5937	18283.0	0.986	29.727273	0.5	0	2736.65	2
5938	18284.0	0.560	1.000000	0.0	1	436.68	0
5939	18285.0	0.329	0.000000	1.0	0	427.00	0
5940	18286.0	0.513	82.333333	0.0	1	1188.43	1
5941	18287.0	0.947	86.875000	0.5	1	4177.89	3

FIGURE 3 – FINAL DATAFRAME

## **CHAPTER-5**

### **OBJECTIVES**

The primary objective of this research is to devise an innovative methodology for customer credit scoring specifically designed for e-commerce platforms. The ever-evolving landscape of online transactions presents unique challenges, including returns, cancellations, and diverse payment preferences, which need to be effectively addressed to ensure customer satisfaction. This study aims to harness the power of big data analytics, machine learning techniques, and statistical analysis to develop a sophisticated credit scoring system. The goal is to construct detailed and personalized credit and purchasing profiles for customers, providing valuable insights that can be utilized to enhance the overall e-commerce experience.

In pursuit of these objectives, the research emphasizes a privacy-centric approach to data handling. The aggregation and anonymization of customer data play a pivotal role in generating personalized profiles while respecting individual privacy. This, in turn, facilitates the tailoring of delivery benefits and the provision of suitable EMI options for customers. The categorization of customers based on their credit and purchasing behaviors is a key aspect, allowing for strategic limitations on certain payment options for users with less favorable profiles. Striking a balance between service customization and privacy preservation is a crucial objective, as the research seeks to optimize e-commerce operations without compromising customer confidentiality.

Furthermore, ethical considerations are a core component of the research objectives. Transparency, consent, and adherence to privacy regulations are prioritized to ensure responsible data handling and system implementation. The research aims to establish a comprehensive and robust credit scoring system that not only meets the specific needs of e-commerce platforms but also adheres to the highest ethical standards. By addressing the challenges posed by the dynamic nature of online transactions and placing privacy at the

forefront, the research endeavors to contribute to the development of a reliable and responsible credit scoring solution for the e-commerce industry.

In summary, the research has multifaceted objectives, encompassing the development of a sophisticated credit scoring system, the construction of personalized customer profiles, and the optimization of e-commerce operations. The emphasis on privacy, ethical considerations, and a tailored approach to customer credit and purchasing behaviors underscores the comprehensive nature of the research objectives, aiming to make a significant contribution to the evolving field of e-commerce analytics and credit scoring.

## **CHAPTER-6**

### **SYSTEM DESIGN & IMPLEMENTATION**

The experimental setup for developing a robust customer credit scoring system in the realm of e-commerce involved a meticulous and comprehensive approach. The process began with the preprocessing of the dataset, transforming it into a refined form denoted as df4. This preprocessing step encompassed critical methods such as concatenation of data from different CSV files (d1.csv and d2.csv), handling missing values, and date conversion with time-based feature engineering. This ensured that the dataset was suitably prepared for the subsequent application of machine learning models.

In the heart of our experimental design, we employed a variety of machine learning models known for their efficacy in classification tasks. These models included k-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), Random Forest, and Random Forest with Polynomial Features. The selection of these models was driven by their suitability for credit scoring predictions and their diverse approaches to handling data. The features chosen for prediction comprised Recency, Frequency, TotalCost, paymtd (payment method), and Return\_Count (return count), providing a comprehensive set of variables for evaluating customer creditworthiness.

To gauge the performance of each machine learning model, we adopted a set of standard metrics, including accuracy, precision, recall, F1 score, and Cohen's Kappa. These metrics provided a nuanced evaluation of the models' effectiveness in predicting customer credit scores. The utilization of multiple metrics ensured a holistic assessment, considering aspects such as true positives, false positives, and false negatives. This multifaceted evaluation allowed for a thorough understanding of each model's strengths and potential limitations in the context of e-commerce credit scoring.

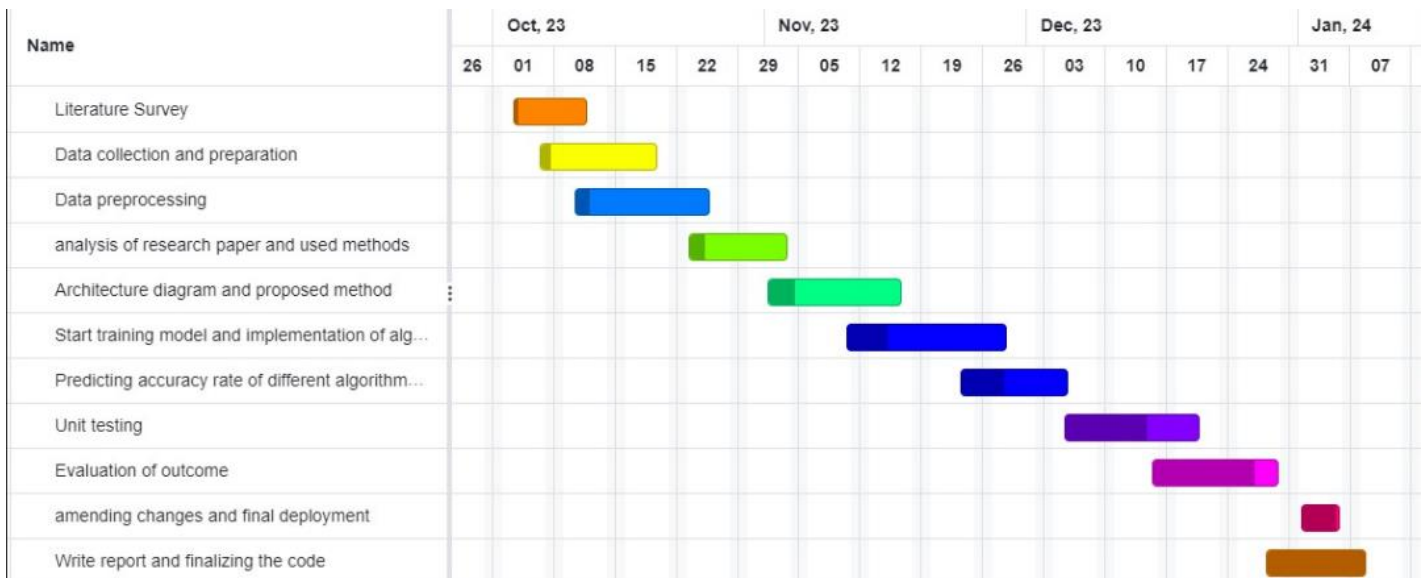
The implementation phase of the experimental setup involved the systematic application of each machine learning model to the prepared dataset. The models were trained on a subset of the data and tested on a separate set to simulate real-world scenarios. The systematic evaluation using diverse metrics facilitated a comparative analysis of model performance, allowing us to identify the most effective approach for predicting customer credit scores in the e-commerce domain.

In summary, our experimental setup was designed with a focus on diversity in machine learning models, comprehensive feature selection, and thorough performance evaluation metrics. This approach ensures the reliability and applicability of the customer credit scoring system in dynamic e-commerce environments. The systematic design and implementation of the experimental setup laid the foundation for deriving meaningful insights into the effectiveness of various machine learning models in predicting customer credit scores.



## CHAPTER-7

### TIMELINE FOR EXECUTION OF PROJECT (GANTT CHART)



**FIGURE 4 - TIMELINE**

## **CHAPTER-8**

### **OUTCOMES**

The outcomes of our comprehensive experimentation with various machine learning models for predicting customer credit scores in the e-commerce domain reveal valuable insights into their performance and effectiveness under different conditions.

The k-Nearest Neighbors (KNN) classifier, despite demonstrating a reasonable accuracy of 80.15%, exhibited sensitivity to outliers, as evidenced by a notable drop in accuracy to 78.74% after their removal. This underscores the importance of outlier handling in enhancing the robustness of the KNN model for credit scoring tasks in e-commerce.

The Decision Tree model emerged as a standout performer, achieving a high accuracy of 92.26%, even with outliers, and maintaining a commendable accuracy of 92.52% without outliers. This consistency highlights the reliability of the Decision Tree approach in predicting customer credit scores in the e-commerce context.

The Support Vector Machine (SVM) model, while effective with outliers, experienced a significant decline in performance without them, with an accuracy dropping from 87.30% to 41.46%. This emphasizes the importance of proper handling of outliers and cautious consideration of data preprocessing techniques when employing SVM for credit scoring tasks.

The Random Forest model consistently demonstrated robust performance, achieving an accuracy of 94.62% with outliers and an even higher accuracy of 95.44% without outliers. This underscores the model's resilience and effectiveness in handling diverse credit scoring scenarios.

Notably, the introduction of Polynomial Features in the Random Forest model led to a remarkable improvement, with an outstanding accuracy of 96.64%. This highlights the potential benefits of feature engineering in enhancing the predictive capabilities of machine learning models for customer credit scoring in e-commerce.

In summary, our outcomes emphasize the significance of model selection, outlier handling, and feature engineering in developing reliable credit scoring systems for e-commerce platforms. While Decision Tree and Random Forest models offer consistent and high-performance results, further exploration into outlier handling and advanced feature engineering techniques could enhance the overall robustness of credit scoring models. The findings provide valuable insights for practitioners and researchers aiming to deploy effective credit scoring systems in dynamic e-commerce environments.

## **CHAPTER-9**

### **RESULTS AND DISCUSSIONS**

The results of our experimentation with various machine learning models for predicting customer credit scores in the e-commerce domain provide substantial insights into the efficacy of different approaches. The discussion below synthesizes the outcomes and delves into the implications of our findings.

#### **9.1. Performance Analysis:**

The KNN classifier demonstrated reasonable accuracy (80.15%) but revealed sensitivity to outliers, highlighting the need for robust outlier handling techniques. The Decision Tree model emerged as a consistent and high-performing choice, maintaining accuracy at 92.26% with outliers and 92.52% without outliers. SVM exhibited effectiveness with outliers but suffered a significant performance drop without proper handling, emphasizing its sensitivity to data preprocessing. The Random Forest model showcased resilience, achieving accuracies of 94.62% with outliers and 95.44% without outliers. The incorporation of Polynomial Features in the Random Forest model resulted in an exceptional accuracy of 96.64%, emphasizing the potential benefits of feature engineering.

#### **9.2. Outlier Sensitivity:**

The sensitivity of models, particularly SVM, KNN, and the performance drop observed without outliers, underscores the importance of meticulous data preprocessing. Outlier removal significantly impacted SVM's accuracy, reinforcing the critical role of identifying and handling outliers for effective credit scoring.

### **9.3. Feature Engineering Impact:**

The introduction of Polynomial Features in the Random Forest model led to a substantial performance boost, highlighting the significance of feature engineering in enhancing predictive capabilities. This suggests that exploring advanced feature engineering methods could further optimize credit scoring models.

### **9.4. Model Robustness:**

The robustness of the Decision Tree and Random Forest models across different scenarios positions them as reliable choices for credit scoring in e-commerce. However, further investigations into outlier handling and feature engineering could enhance their overall robustness.

### **9.5. Practical Implications:**

Our findings offer practical insights for developing credit scoring systems in e-commerce platforms. Decision Tree and Random Forest models can provide reliable performance, but careful consideration of outlier handling and feature engineering is crucial. Practitioners should be aware of the sensitivity of models to outliers and explore advanced feature sets for improved predictive accuracy.

### **9.6. Future Directions:**

Future work should focus on refining preprocessing steps, exploring advanced feature engineering methods, and potentially integrating ensemble models to harness the strengths of multiple algorithms. Additionally, real-time data integration, dynamic feature adjustments, and the incorporation of deep learning techniques could further enhance predictive accuracy and adaptability in the dynamic e-commerce landscape.

In summary, our results contribute to bridging gaps in existing literature by providing a nuanced understanding of model effectiveness and the importance of outlier handling and feature engineering in developing robust credit scoring systems for e-commerce platforms. The insights gained from this study can inform the design and implementation of effective credit scoring solutions, ensuring their reliability in dynamic and evolving e-commerce environments.

## **CHAPTER-10**

### **CONCLUSION**

In conclusion, our research into predicting customer credit scores within the e-commerce domain offers valuable insights into the effectiveness of diverse machine learning models. The consistent high performance of the Decision Tree and Random Forest models, with accuracies of 92.26% and 94.62%, respectively, underscores their suitability for credit scoring tasks. The SVM model's efficacy, particularly with outliers, emphasizes the critical role of data preprocessing in ensuring model effectiveness. Moreover, the introduction of Polynomial Features in the Random Forest model showcases the potential advantages of feature engineering, resulting in an impressive accuracy of 96.64%.

The sensitivity to outliers observed in the KNN classifier highlights the importance of careful model selection and a nuanced understanding of the dataset's characteristics. Our research underscores the significance of thoughtful model selection and thorough data preprocessing in the development of reliable credit scoring systems for e-commerce platforms. The findings contribute to guiding practitioners in choosing appropriate models and preprocessing techniques tailored to their specific datasets.

Looking ahead, future research should delve into advanced feature engineering methods and explore the integration of ensemble models to further enhance the robustness and adaptability of credit scoring systems in the dynamic landscape of e-commerce. By continuing to refine and innovate in these areas, we can contribute to the development of more sophisticated and reliable credit scoring systems that meet the evolving demands of the e-commerce industry.

## REFERENCES

- [1] Maoguang Wang and Hang Yang. Research on customer credit scoring model based on bank credit card. In *Intelligent Information Processing X: 11th IFIP TC 12 International Conference, IIP 2020, Hangzhou, China, July 3–6, 2020, Proceedings 11*, pages 232–243. Springer, 2020.
- [2] Haichao Zhang, Ruishuang Zeng, Linling Chen, and Shangfeng Zhang. Research on personal credit scoring model based on multi source data. In *Journal of Physics: Conference Series*, volume 1437, page 012053. IOP Publishing, 2020.
- [3] Kirti Maheshwari, Ria Khapekar, Anmol Bahl, and Kunal Bhatia. Credit profile of e-commerce customer. 2019.
- [4] Zengyuan Wu, Lingmin Jin, Jiali Zhao, Lizheng Jing, and Liang Chen. Research on segmenting e-commerce customer through an improved k-medoids clustering algorithm. *Computational Intelligence and Neuroscience*, 2022, 2022.
- [5] Aslihan Dursun and Meltem Caber. Using data mining techniques for profiling profitable hotel customers: An application of rfm analysis. *Tourism management perspectives*, 18:153–160, 2016.
- [6] A Joy Christy, A Umamakeswari, L Priyatharsini, and A Neyaa. Rfm ranking—an effective approach to customer segmentation. *Journal of King Saud University-Computer and Information Sciences*, 33(10):1251–1257, 2021. `
- [7] Onur Dogan, Ejder Ayçin, and Zeki Bulut. Customer segmentation by using rfm model and clustering methods: a case study in retail industry. *International Journal of Contemporary Economics and Administrative Sciences*, 8, 2018.
- [8] Ching-Hsue Cheng and You-Shyang Chen. Classifying the segmentation of customer value via rfm model and rs theory. *Expert systems with applications*, 36(3):4176–4184, 2009.
- [9] RW Sembiring Brahmana, Fahd Agodzo Mohammed, and K Chairuang. Customer



segmentation based on rfm model using k means, k-medoids, and dbscan methods. Lontar Komput. J. Ilm. Teknol. Inf, 11(1):32, 2020.

[10] İnanç Kabasakal. Customer segmentation based on recency frequency monetary model: A case study in e-retailing. Bilişim Teknolojileri Dergisi, 13(1):47–56, 2020.

[11] Uus Firdaus and D Utama. development of bank’s customer segmentation model based on rfm+ b approach. Int. J. Innov. Comput. Inf. Cont, 12(1):17–26, 2021.

[12] Md Billal Hossain, Nargis Dewan, Aslan Amat Senin, and Csaba Balint Illes. Evaluating the utilization of technological factors to promote e-commerce adoption in small and medium enterprises. Electronic Commerce Research, pages 1–20, 2023.

## APPENDIX-A

### PSUEDOCODE

```
# Import necessary libraries

import pandas as pd

import numpy as np

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification_report, confusion_matrix, accuracy_score,
precision_score, recall_score, f1_score, roc_curve, auc, make_scorer, cohen_kappa_score

from sklearn.preprocessing import LabelEncoder, PolynomialFeatures

from sklearn.datasets import make_classification

import matplotlib.pyplot as plt

import seaborn as sns

from datetime import datetime

import random

from tabulate import tabulate

from scipy.stats import zscore


# Load datasets

d1 = pd.read_csv("/content/data0.csv", encoding='unicode_escape')

d2 = pd.read_csv("/content/d2.csv", encoding='unicode_escape')


# Merge datasets

df = pd.concat([d1, d2], ignore_index=True, join="inner")
```

```
# Display information about missing values
print(df.isnull().sum())

# Drop rows with missing values
df = df.dropna()

# Feature extraction
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'], format='%m/%d/%Y %H:%M')
df['DaysSinceLastPurchase'] = (datetime.now() - df.groupby('Customer ID')['InvoiceDate'].transform('max')).dt.days - 4400
df['LastPurchaseDate'] = df.groupby('Customer ID')['InvoiceDate'].transform('max')
df['PurchasePeriod'] = (df.groupby('Customer ID')['InvoiceDate'].transform('max') - df.groupby('Customer ID')['InvoiceDate'].transform('min')).dt.days
df['Frequency'] = df['PurchasePeriod'] / df.groupby('Customer ID')['Invoice'].transform('nunique')
df['TotalCost'] = df['Quantity'] * df['Price']

# Extract total cost feature for each invoice per customer
df2 = df.groupby(['Customer ID', 'Invoice', 'DaysSinceLastPurchase', 'Frequency'])['TotalCost'].sum().reset_index()
df1 = df.groupby(['Customer ID', 'Invoice', 'LastPurchaseDate', 'Frequency'])['TotalCost'].sum().reset_index()

# Extracting return, total return, and assigning pay method features
y = []
return_list = []

# Generate random values for pay method and return
```

---

```
for i in range(len(df1['Invoice'])):
    y += [random.randint(0, 1)]

for i in df1['Invoice']:
    try:
        x = int(i)
        return_list += [0,]
    except:
        return_list += [1,]

df1['pay_mtd'] = y
df1['return'] = return_list

# Group by 'Customer ID' and calculate return count
return_count = df1.groupby('Customer ID')['return'].sum().reset_index()
return_count.rename(columns={'return': 'Return_Count'}, inplace=True)

# Merge the return count with the original DataFrame
df1 = df1.merge(return_count, on='Customer ID', how='left')

# Extract pay method as a normalized value for each customer
df1['paymtd'] = df1.groupby('Customer ID')['pay_mtd'].transform(lambda x: x.sum() /
len(x))

# Adding pay method to the dataset
df3 = pd.merge(df2, df1[['Invoice', 'paymtd', 'return', 'Return_Count']], on='Invoice',
how='left')
```

---

---

```

df3.rename(columns={'DaysSinceLastPurchase': 'Recency'}, inplace=True)

df3 = df3.groupby(['Customer ID', 'Recency', 'Frequency', 'paymtd',
'Return_Count'])['TotalCost'].sum().reset_index()

# Assigning classes to customers based on credit score

df3['Recency'] = 1 - (df3['Recency'] / 1000)

df3['c_s'] = df3['Recency'] + (df3['TotalCost'] / 1000) + (df3['Frequency'] / 100) +
df3['paymtd'] - (df3['Return_Count'] / 10)

df3['c_s'] = pd.to_numeric(df3['c_s'], errors='coerce')

df3['Class'] = pd.qcut(df3['c_s'], q=[0, 0.25, 0.65, 0.85, 1.0], labels=['0', '1', '2', '3'])

df3 = df3.drop(columns=['c_s'])

# KNN Classifier with outliers

features = ['Recency', 'Frequency', 'TotalCost', 'paymtd', 'Return_Count']

target = 'Class'

X = df3[features]

y = df3[target]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

knn_model = KNeighborsClassifier(n_neighbors=3)

knn_model.fit(X_train, y_train)

y_pred = knn_model.predict(X_test)

conf_matrix = confusion_matrix(y_test, y_pred)

# (Code for plotting confusion matrix and metrics calculation)

# ...

# Decision Tree Classifier with outliers

features = ['Recency', 'Frequency', 'TotalCost', 'paymtd', 'Return_Count']

target = 'Class'

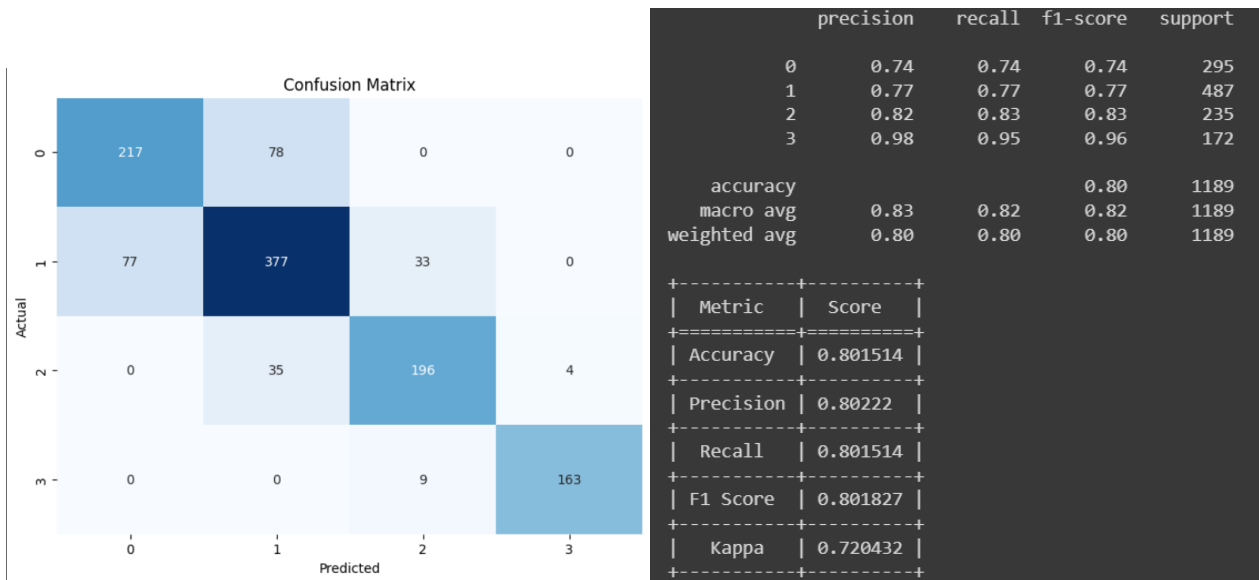
```

---

## APPENDIX-B

### SCREENSHOTS

The below figures contain the confusion matrix, classification reports, accuracy, precision, Recall, f1- score, kappa score for 4 models (SVM, KNN, Decision Tree, Random Forest). The Figures [5], [6], [7], [8] contain the analysis of the models when dataset still contained outliers and Figures [9], [10], [11], [12] contain analysis of these models with dataset after excluding outliers. Figure [13] shows performance of Random Forest model trained with a Polynomial Feature transformation applied to the outlier free dataset to further improve accuracy.



**FIGURE 5 – CONFUSION MATRIX, CLASSIFICATION REPORT, EVALUATION METRICS**

KNN Classifier with outliers

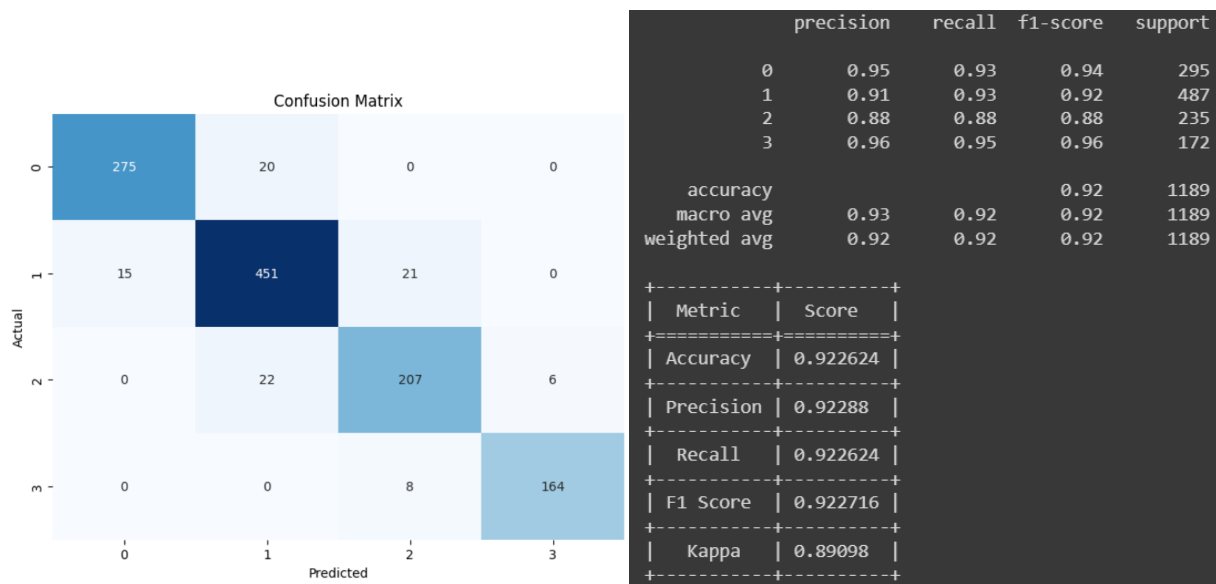


FIGURE 6 – CONFUSION MATRIX, CLASSIFICATION REPORT, EVALUATION METRICS

## Decision Tree Classifier with outliers

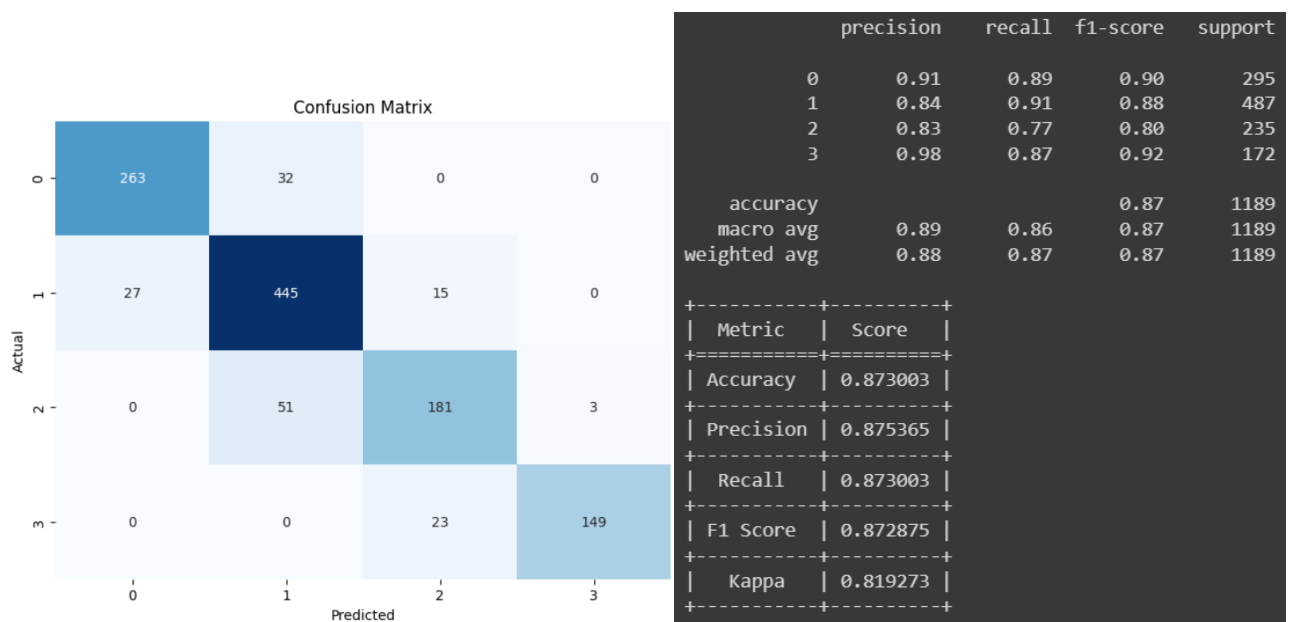


FIGURE 7 – CONFUSION MATRIX, CLASSIFICATION REPORT, EVALUATION METRICS

## SVM with outliers

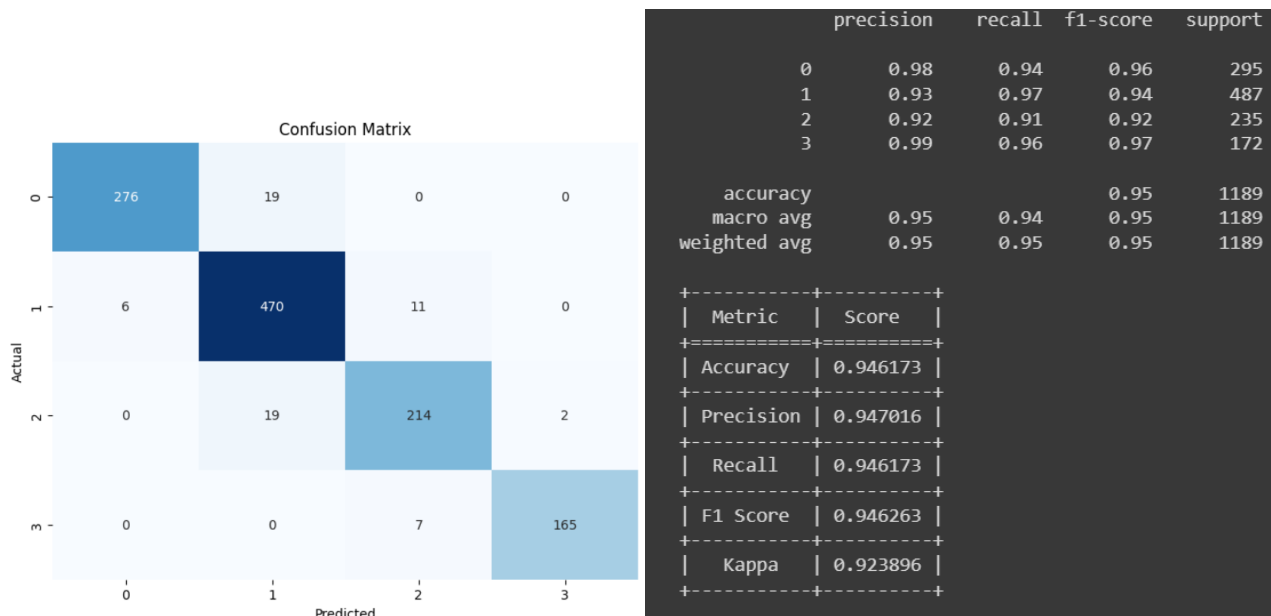


FIGURE 8 – CONFUSION MATRIX, CLASSIFICATION REPORT, EVALUATION METRICS

## Random Forest with outliers

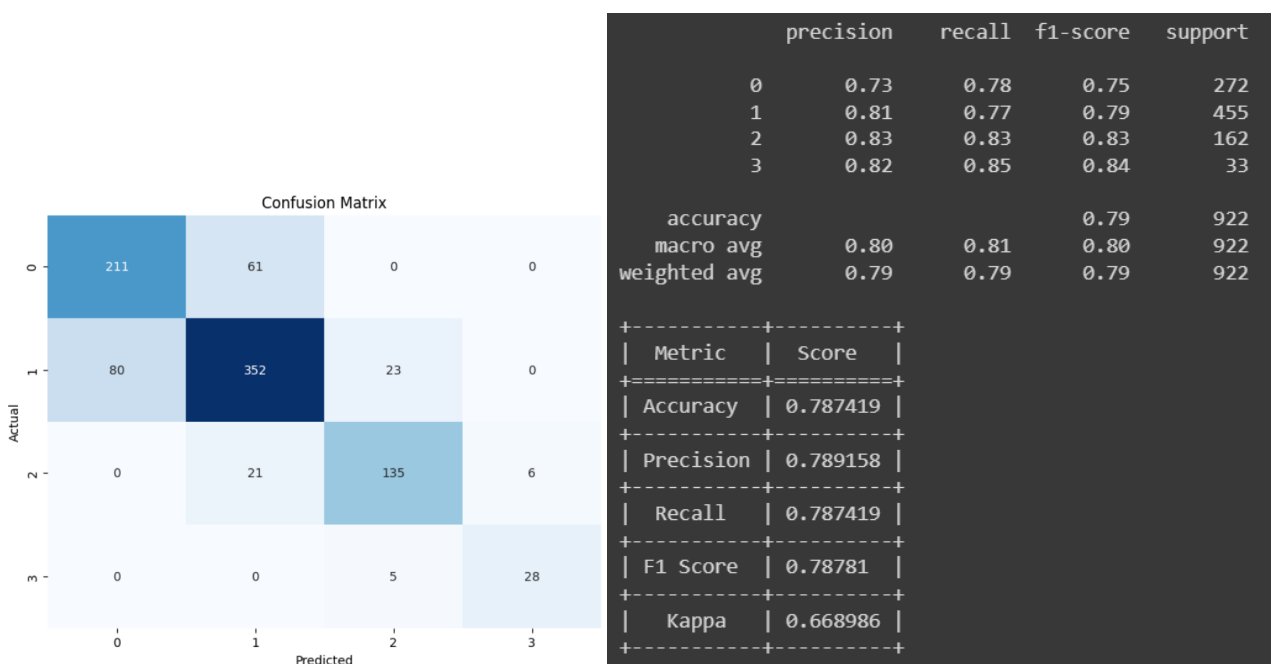


FIGURE 9 – CONFUSION MATRIX, CLASSIFICATION REPORT, EVALUATION METRICS

## KNN Classifier without outliers



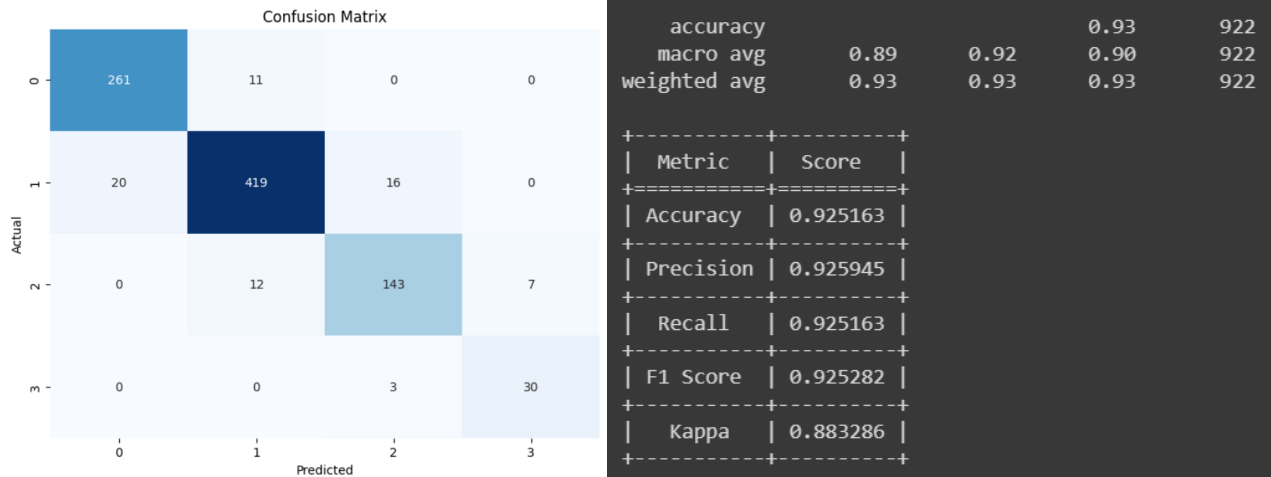


FIGURE 10 – CONFUSION MATRIX, CLASSIFICATION REPORT, EVALUATION METRICS

## Decision Tree Classifier without outliers

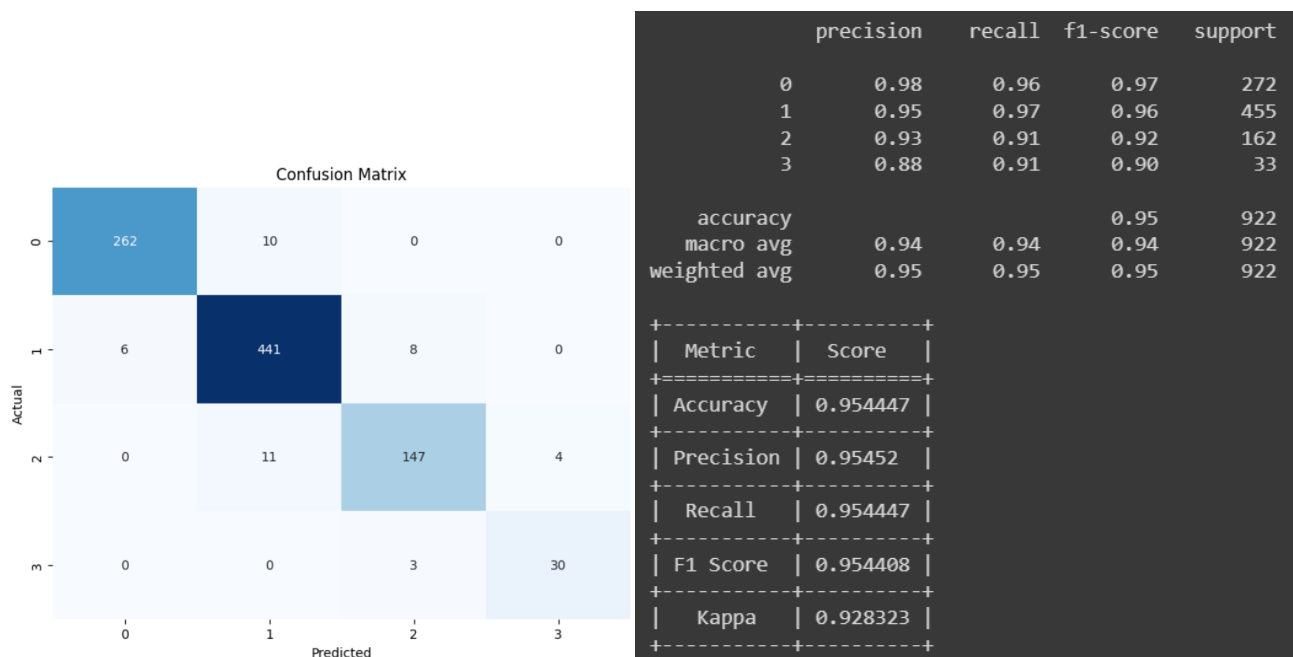


FIGURE 11 – CONFUSION MATRIX, CLASSIFICATION REPORT, EVALUATION METRICS

## Random Forest without outliers

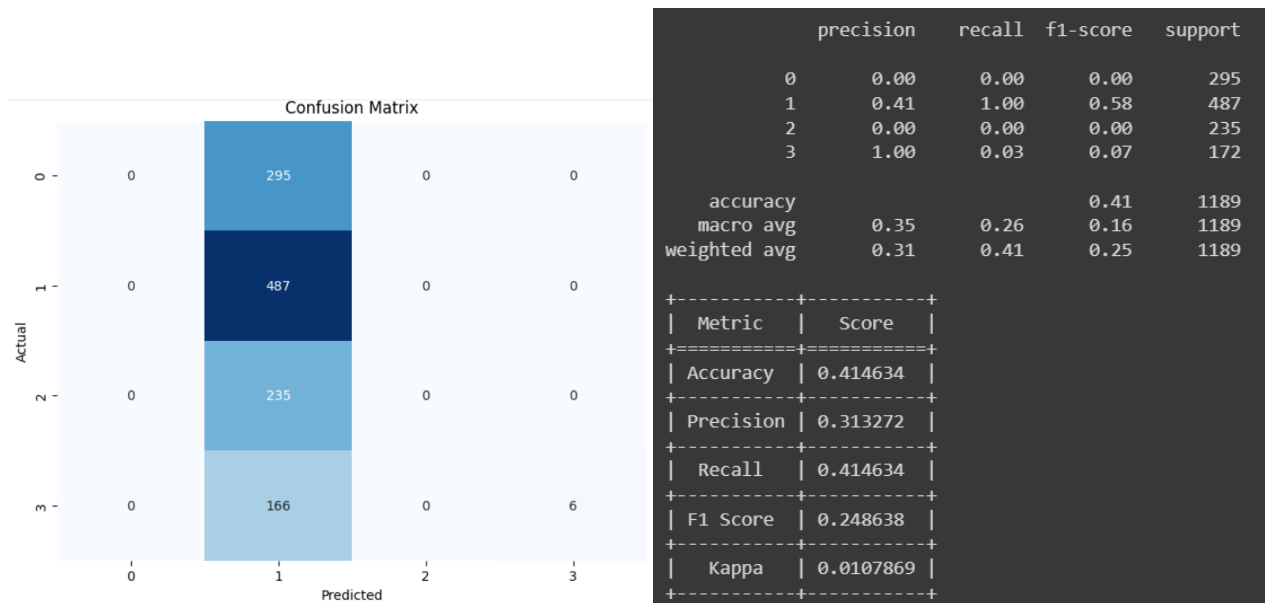


FIGURE 12 – CONFUSION MATRIX, CLASSIFICATION REPORT, EVALUATION METRICS

## SVM without outliers

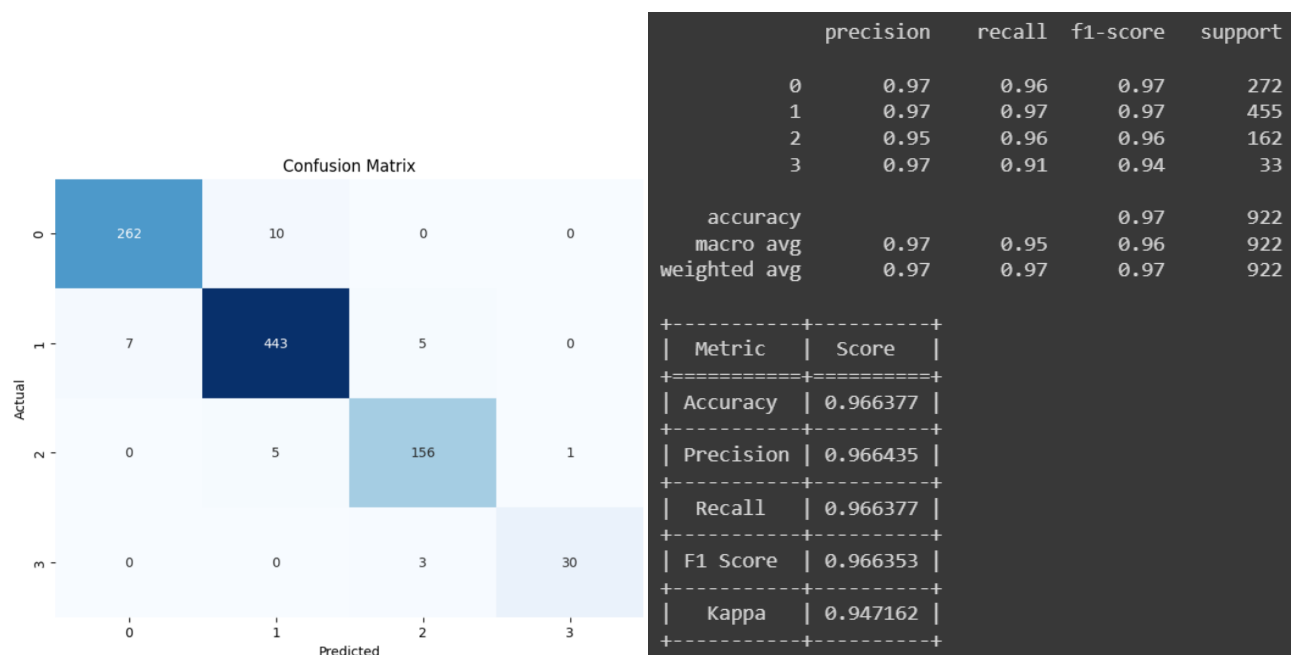


FIGURE 13 – CONFUSION MATRIX, CLASSIFICATION REPORT, EVALUATION METRICS

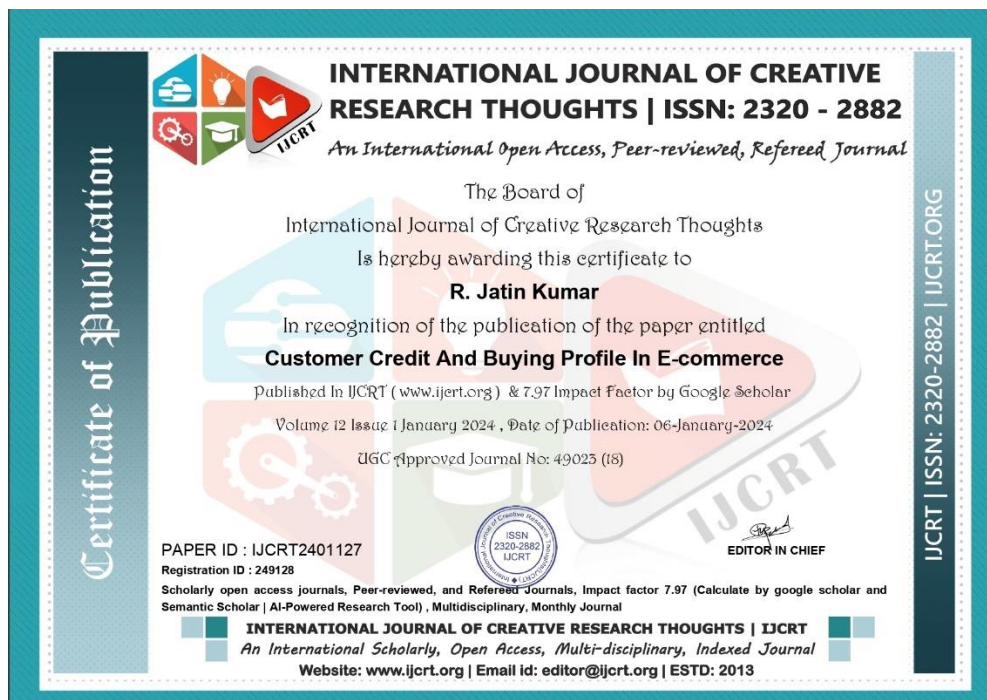
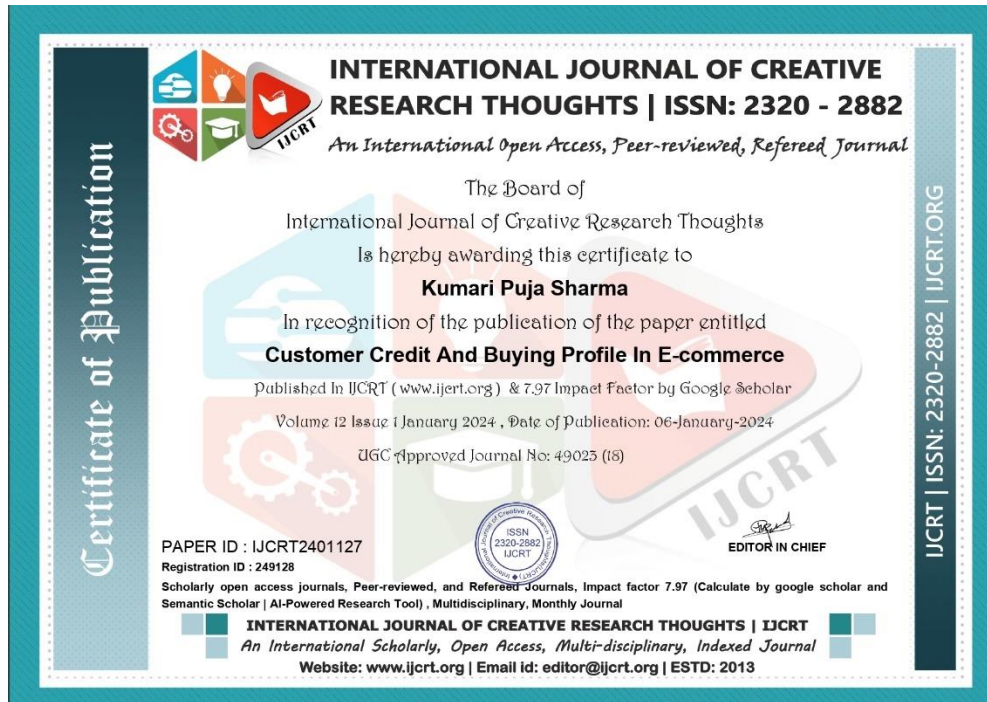
## Random Forest with Polynomial Features

## APPENDIX-C

### ENCLOSURES

#### 1. Conference Paper Presented Certificates of all students.









**2. Certificate(s) of any Achievement/Award won in any project related event.**

-- N – A –

### 3. Similarity Index / Plagiarism Check report Percentage (%).





**The project work carried out here is mapped to SDG – 9 Industry, Innovation, and Infrastructure.**

This research aligns with SDG 9 by leveraging advanced technologies, including big data analytics and machine learning, to propose an innovative methodology for constructing personalized credit and purchasing profiles in e-commerce. The focus on industry optimization, technological innovation, and ethical considerations contributes to the goal of fostering inclusive and sustainable industrialization. The proposed system enhances efficiency in e-commerce operations, promoting the responsible use of technology for economic growth and improved services while respecting individual privacy.