**SSN COLLEGE OF ENGINEERING KALAVAKKAM -603110**

**Department of Computer Science and Engineering**

**Mini Project**

**UCS2612 – Machine Learning Lab**

**E-commerce Customer Data For Behavior Analysis**

**TEAM MEMBERS:**

1. **Ashwin R - 3122 21 5001 014**
2. **Deepti Gowtham:- 3122215001023**
3. **Gayathri Venkatesan:- 3122 21 5001 026**
4. **Karthik Vijayakumar:- 3122 21 5001 040**

**1. Introduction**

In this report, we delve into the realm of sentiment analysis within the context of e-commerce platforms, where understanding customer sentiment plays a pivotal role in shaping business strategies and enhancing customer satisfaction. Leveraging natural language processing (NLP) techniques, we explore the use of machine learning models to decipher the sentiments expressed in customer reviews and feedback. By analyzing textual data sourced from online reviews, comments, and ratings, we aim to extract valuable insights into customer perceptions, preferences, and satisfaction levels. Through this analysis, we seek to shed light on the efficacy of sentiment analysis in discerning nuanced sentiments and its implications for e-commerce businesses in terms of improving customer engagement, optimizing product offerings, and fostering long-term relationships with customers.

**2. Problem Statement**

The problem statement revolves around the need for businesses to understand and predict customer behavior for future purchases. With the vast amount of data generated by customer interactions, there is a pressing need to leverage this data effectively to gain actionable insights. This involves analyzing past purchasing patterns, browsing behavior, feedback, and demographics to anticipate future actions and preferences. However, the challenge lies in the complexity and volume of data, as well as the dynamic nature of consumer behavior. Businesses need robust data collection, integration, and analysis techniques to uncover meaningful patterns and trends. Additionally, ethical considerations regarding data privacy and consent must be carefully addressed to ensure customer trust and compliance with regulations. Ultimately, the goal is to develop predictive models, segmentation strategies, and personalized marketing initiatives that enhance customer satisfaction, loyalty, and long-term profitability.

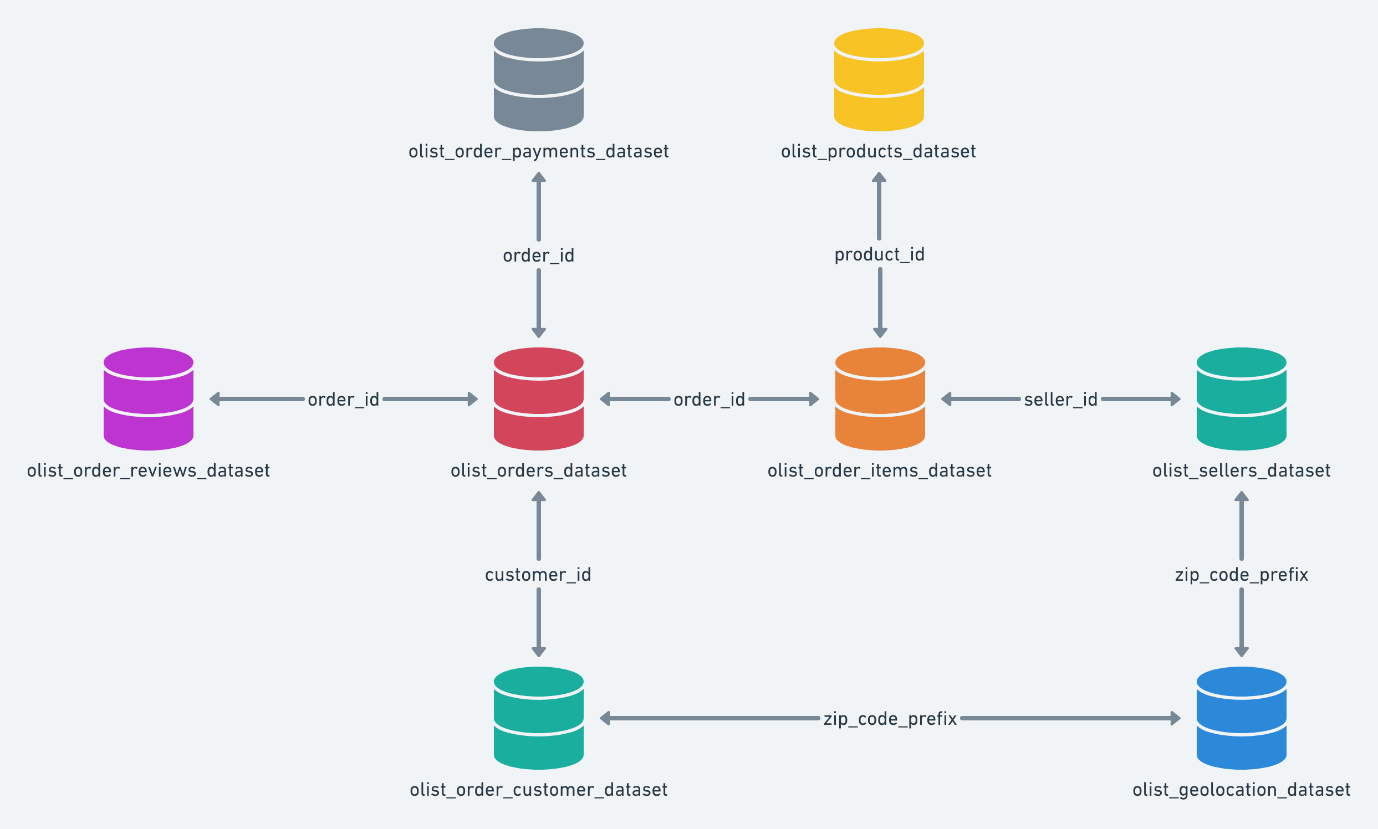
In the realm of modern commerce, the quest to understand and forecast customer behavior for future purchases has become an essential strategic imperative for businesses across industries. With the proliferation of digital touchpoints and the exponential growth of data generated by consumer interactions, companies have an unprecedented opportunity to delve deep into customer preferences, habits, and decision-making patterns. By harnessing advanced analytics techniques such as predictive modeling, machine learning, and data mining, businesses can unlock valuable insights from this wealth of information, enabling them to anticipate market trends, personalize offerings, and optimize marketing strategies for enhanced customer engagement and satisfaction.Top of Form

**3. Development Environment**

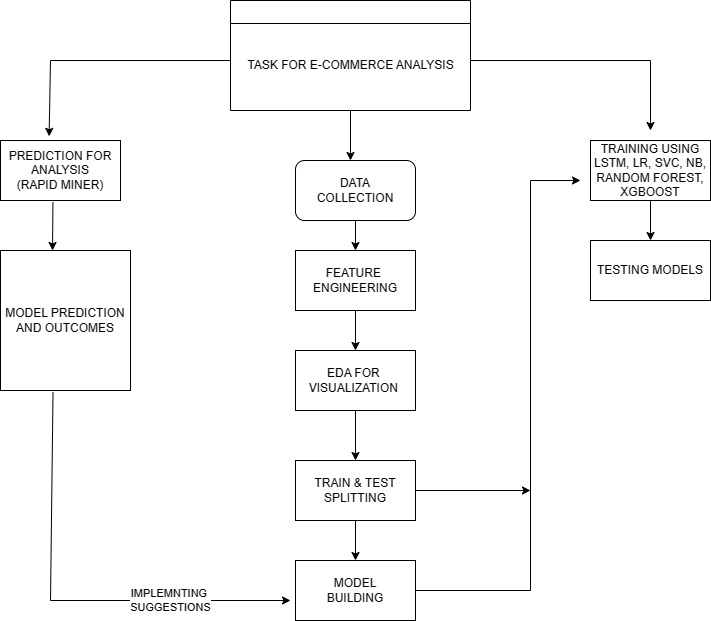
The project was developed using Python programming language along with several essential libraries, including pandas for data manipulation, NumPy for numerical computing, NLP, Keras, scikit-learn for machine learning algorithms and model evaluation, seaborn and Matplotlib for data visualization. These libraries provided a robust framework for data analysis and model development. Additionally, Google Colaboratory was utilized as the primary development environment for data preprocessing, model training, and evaluation. The collaborative nature of Google Colaboratory allowed for seamless collaboration among team members and provided access to powerful computing resources, enabling efficient experimentation and iteration throughout the project lifecycle. Overall, leveraging Python and these libraries in conjunction with Google Colaboratory facilitated the development of a scalable and effective machine learning solution for predicting customer churn in the e-commerce sector.

**4. System Architecture diagram**

**a) Database**



**b) System Overview**

****

**5. Dataset collection**

The dataset was collected from[**https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce**](https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce) **.**

It has nine datasets.

* **olist\_customers\_dataset.csv**
* **olist\_geolocation\_dataset.csv**
* **olist\_order\_items\_dataset.csv**
* **olist\_order\_payments\_dataset.csv**
* **olist\_order\_reviews\_dataset.csv**
* **olist\_orders\_dataset.csv**
* **olist\_products\_dataset.csv**
* **olist\_sellers\_dataset.csv**
* **product\_category\_name\_translation.csv**

**And each of it’s description:-**

#### olist\_customers\_dataset.csv - Information about customers who made purchases.

#### olist\_geolocation\_dataset.csv - Geolocation data related to Brazilian postal codes.

#### olist\_order\_items\_dataset.csv - Details of items included in each order.

#### olist\_order\_payments\_dataset.csv - Payment information for each order.

#### olist\_order\_reviews\_dataset.csv - Reviews provided by customers for their orders.

#### olist\_orders\_dataset.csv - General information about orders.

#### olist\_products\_dataset.csv - Information about products available for purchase.

#### olist\_sellers\_dataset.csv - Details of sellers who provided products.

#### product\_category\_name\_translation.csv - Translation of product category names.

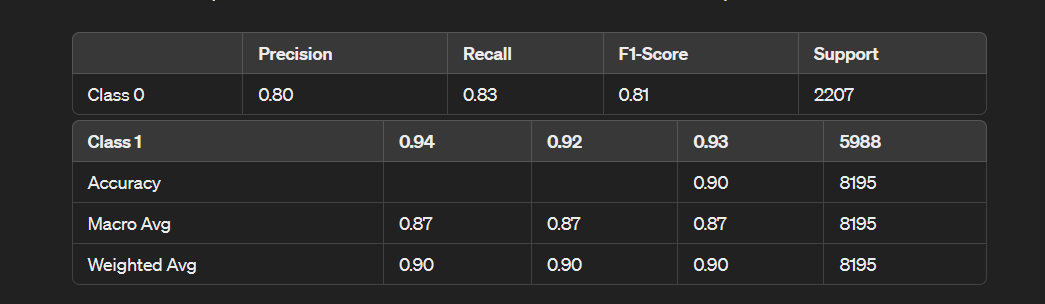
#### 6. Results and Inference

**A) LSTM model**

Model Score on Training Data : 91.08962416648865 %

Model Score on Testing Data : 89.68883752822876 %

**Classification report:-**

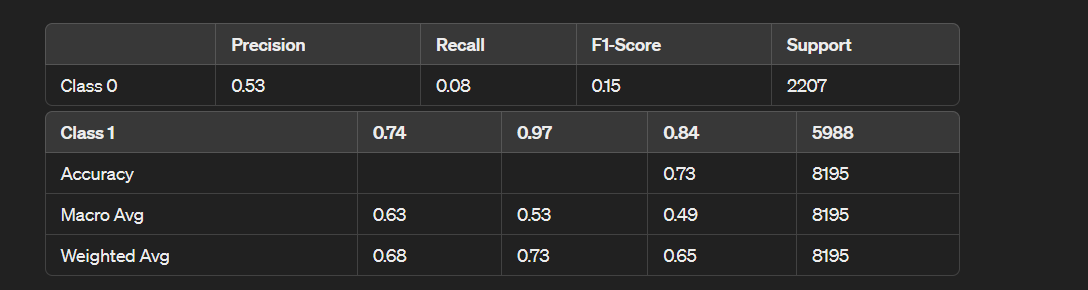
****

**B) Logistic regression**

Model score on Training Data = 74.4920993227991 %

Model score on Testing Data = 73.32519829164124 %

**Classification report:-**

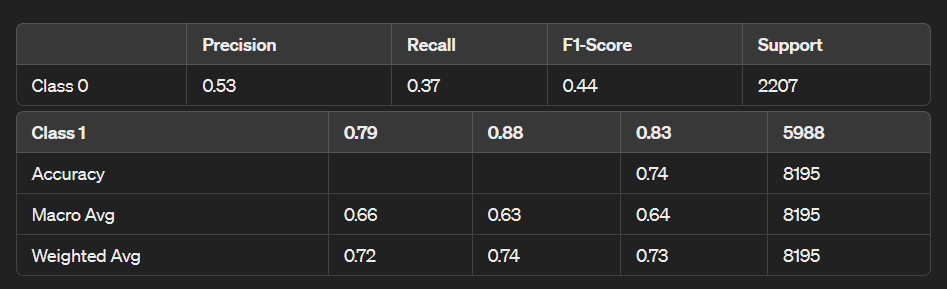
****

**C) Naive bayes**

Model score on Training Data = 74.55615886767129 %

Model score on Testing Data = 74.35021354484441 %

**Classification Report:-**

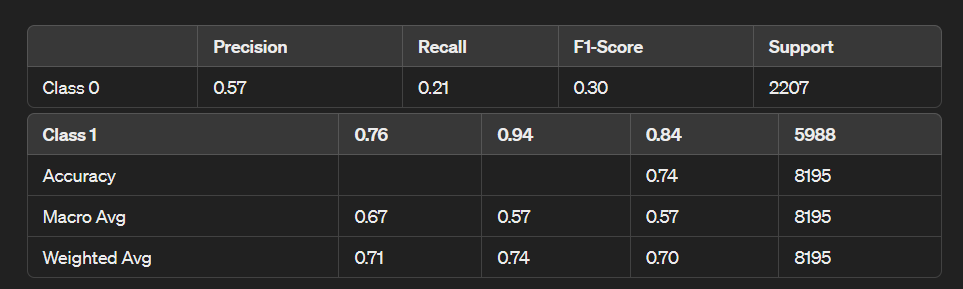


**D) SUPPORT VECTOR CLASSIFIER (SVC)**

Model score on Training Data = 79.7907388200842 %

Model score on Testing Data = 74.49664429530202 %

**Classification Report:-**

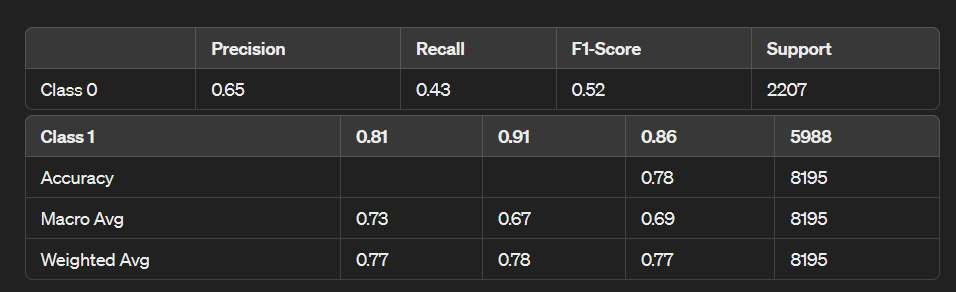


**E) RANDOM FOREST CLASSIFIER**

Model score on Training Data = 99.5790372765542 %

Model score on Testing Data = 78.40146430750458 %

**Classification Report:-**

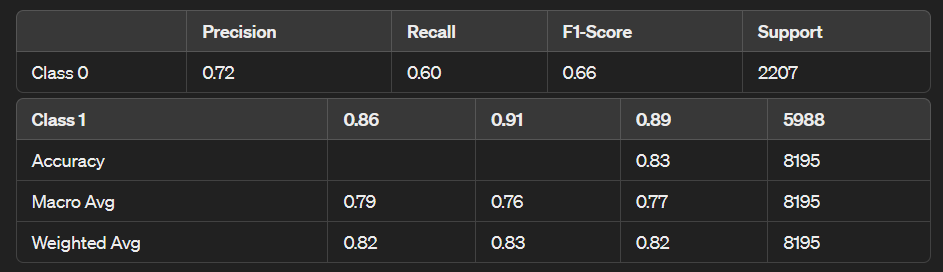
****

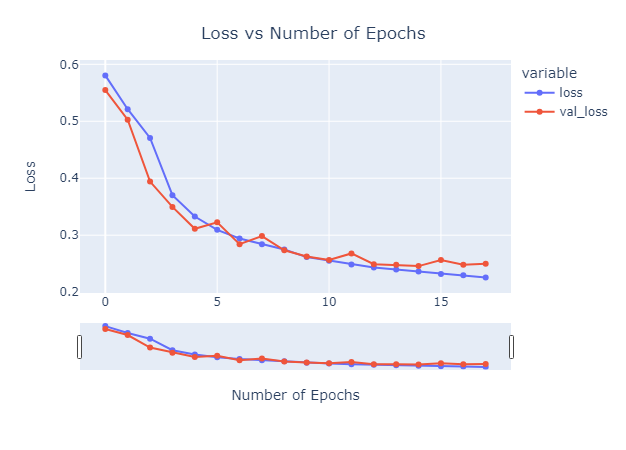
**F) XGBOOST CLASSIFIER**

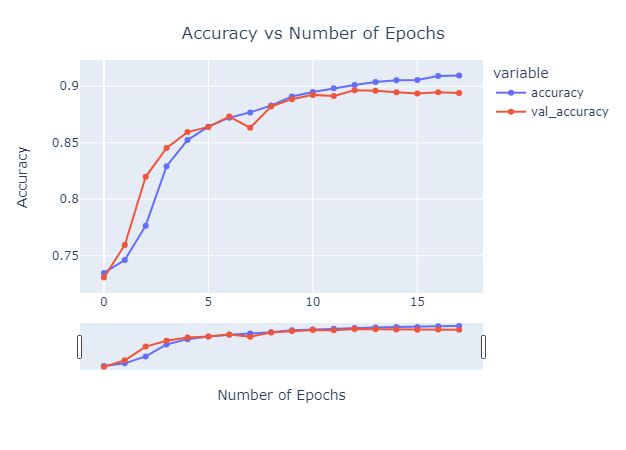
Model score on Training Data = 93.8167286925752 %

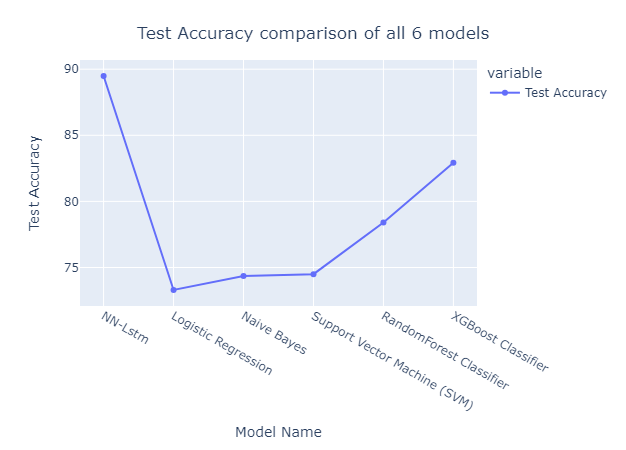
Model score on Testing Data = 82.91641244661379 %

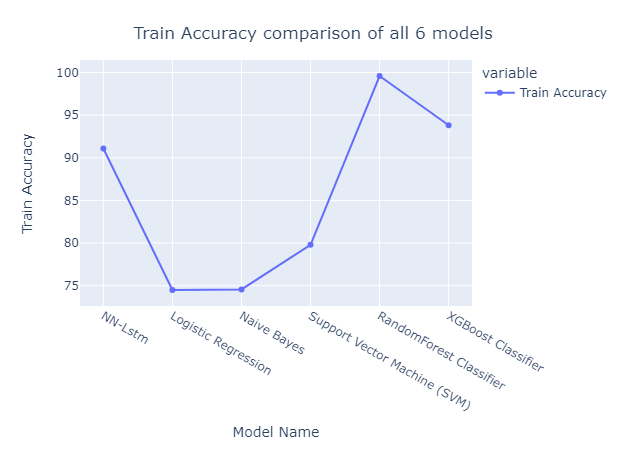
**Classification Report:-**

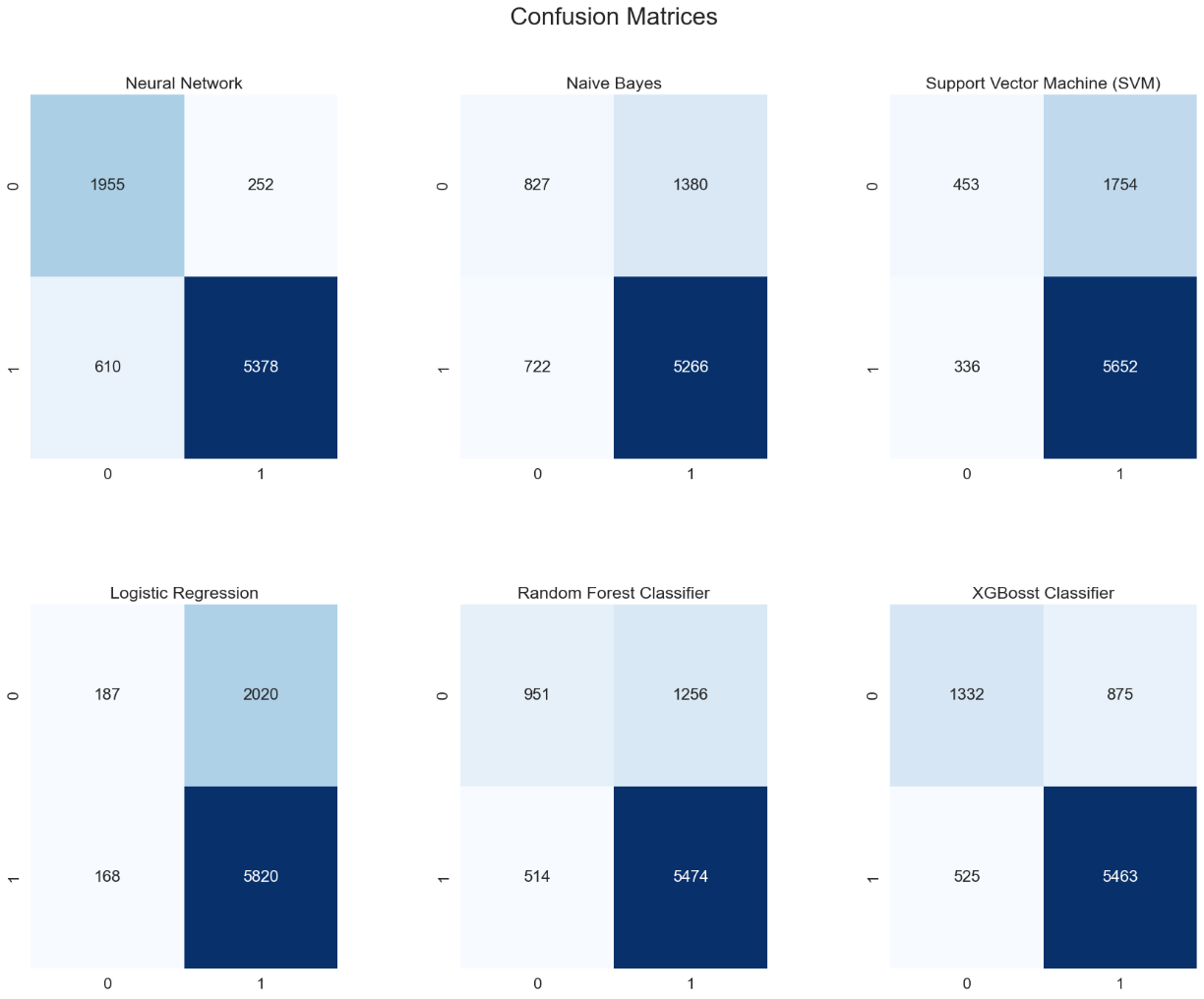


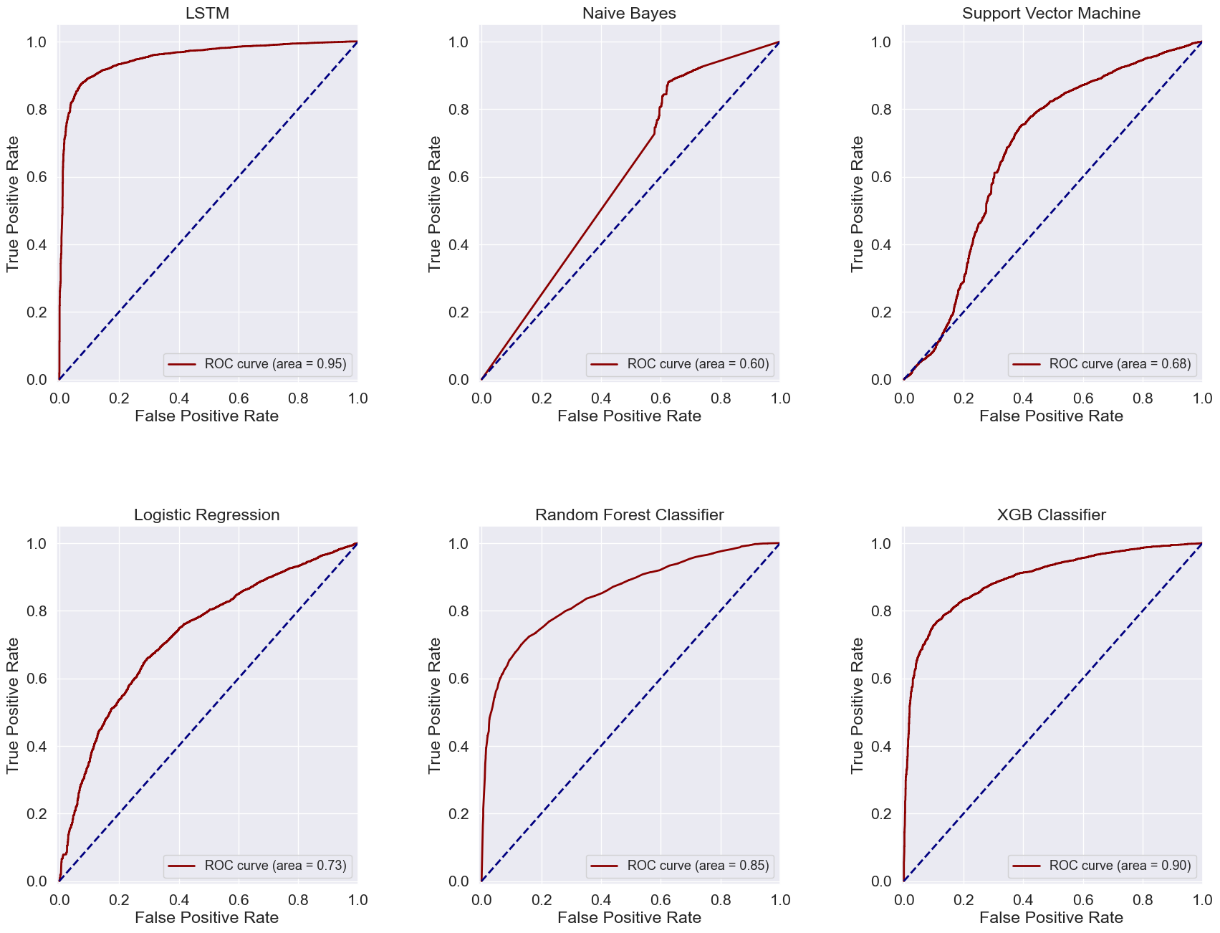
****

****

****

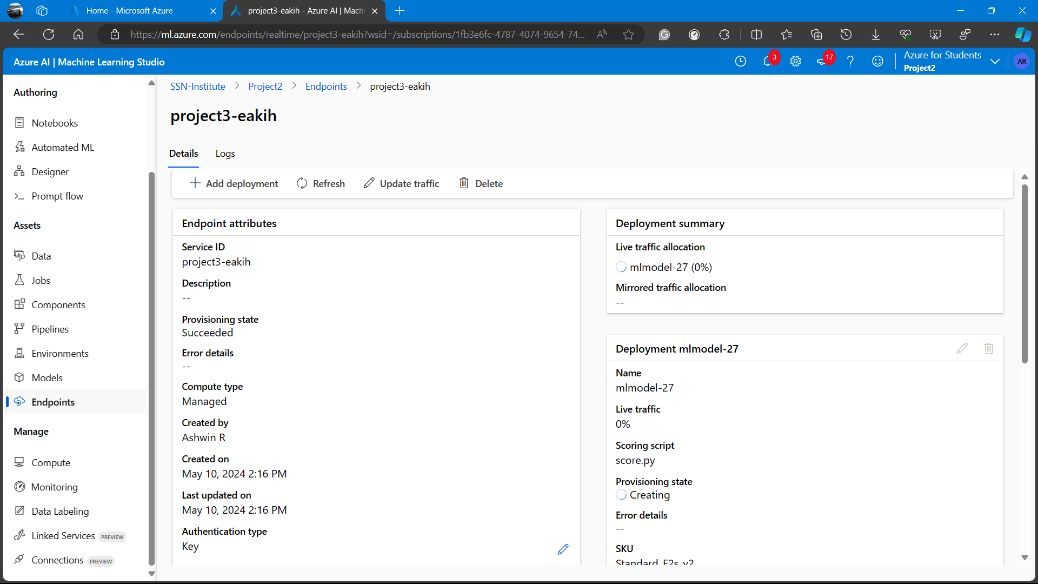
****

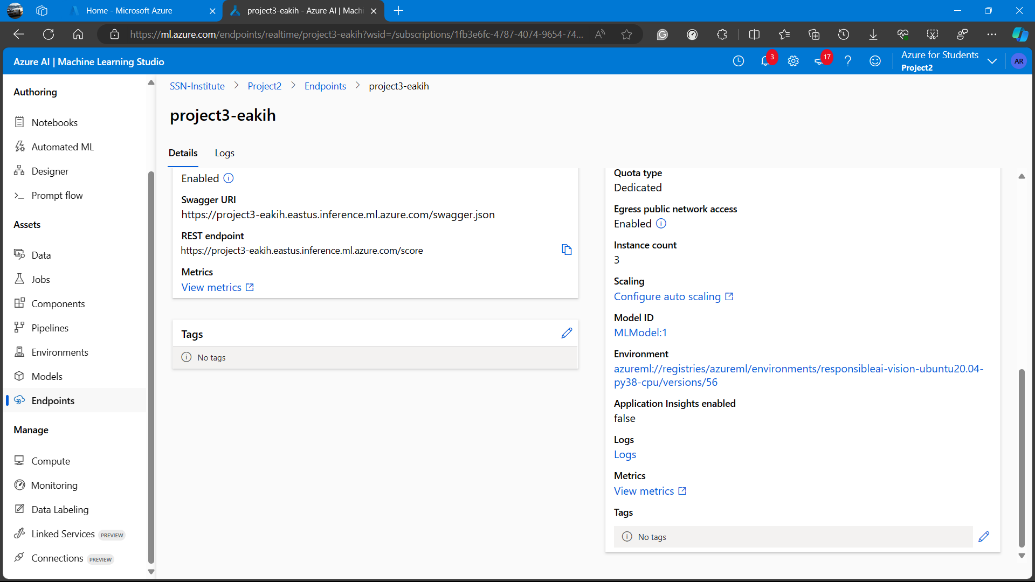
****

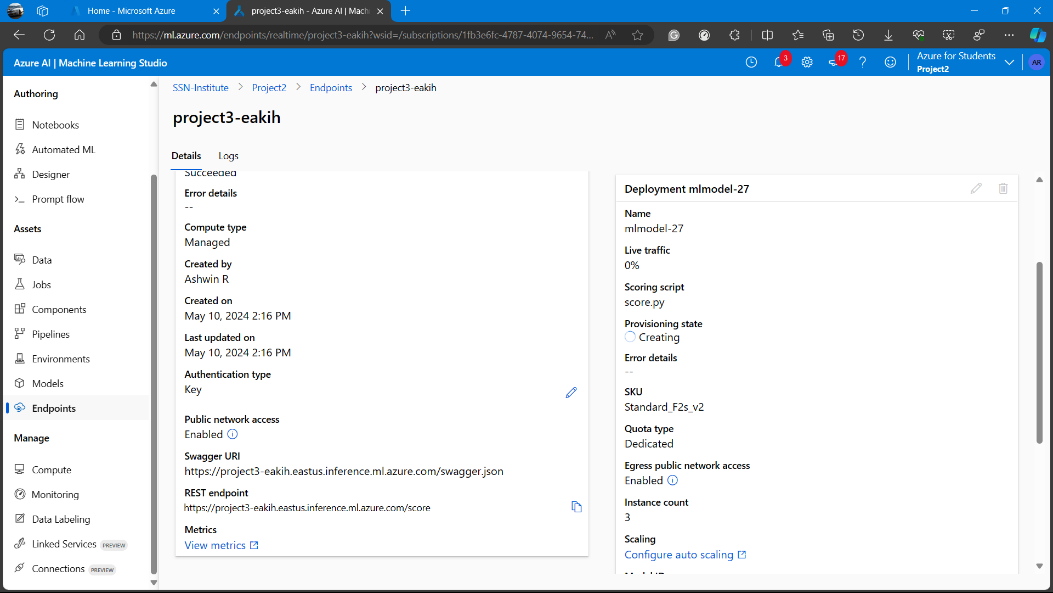
****

**Deployed on Cloud(Microsoft Azure)**

* [Swagger URI](https://project3-eakih.eastus.inference.ml.azure.com/swagger.json)
* [REST endpoint](https://project3-eakih.eastus.inference.ml.azure.com/score)

****

****

****

**Explanation of the above test result statistics of the models above:-**

1. **Long Short-Term Memory (LSTM) Model:** The LSTM model achieved the highest accuracy on the test data among the models listed, with an accuracy of approximately 89.69%. It demonstrated high precision, recall, and F1-score for both classes (0 and 1), indicating good performance in classifying instances from each class. The precision, recall, and F1-score for class 1 (positive class) are notably higher than those for class 0 (negative class), suggesting better prediction capability for positive instances.
2. **Logistic Regression**: The logistic regression model achieved moderate performance with an accuracy of around 73.33% on the test data. While it's a simple and interpretable model, it might not capture complex relationships in the data as effectively as more advanced models like LSTM.
3. **Naive Bayes**: The Naive Bayes classifier demonstrated similar performance to logistic regression, with an accuracy of about 74.35% on the test data. Naive Bayes assumes independence among features, which might not hold true in all datasets, potentially limiting its performance compared to more sophisticated models.
4. **Support Vector Classifier (SVC):** SVC achieved an accuracy of approximately 74.50% on the test data. While it generally performs well in high-dimensional spaces, it might struggle with large datasets due to its computational complexity.
5. **Random Forest Classifier:** The Random Forest model achieved the highest training accuracy of around 99.58%, indicating potential overfitting. Its test accuracy of approximately 78.40% suggests good generalization performance but lower than that of the LSTM and XGBoost models.
6. **XGBoost Classifier:** XGBoost demonstrated strong performance with an accuracy of about 82.92% on the test data. It's known for its scalability, efficiency, and effectiveness in handling large datasets, which might contribute to its performance compared to other models.

**Inference:**

1. **Accuracy**: The accuracy metric represents the overall correctness of the model's predictions. In this context, the LSTM model achieved the highest accuracy of approximately 89.69% on the test data, indicating that it correctly classified about 89.69% of the instances in the dataset. This suggests that the LSTM model is effective in distinguishing between different classes within the dataset.
2. **Precision and Recall**: Precision measures the proportion of correctly predicted positive instances among all instances predicted as positive, while recall measures the proportion of correctly predicted positive instances among all actual positive instances in the dataset. For the LSTM model, both precision and recall are high for class 1 (positive class), indicating that it effectively identifies instances belonging to this class. However, precision and recall for class 0 (negative class) are slightly lower but still acceptable, suggesting that the model performs reasonably well in identifying instances from both classes.
3. **F1-score**: The F1-score is the harmonic mean of precision and recall and provides a balance between the two metrics. For the LSTM model, the F1-score is high for both classes, indicating a good balance between precision and recall. This suggests that the model achieves both high precision and high recall simultaneously, making it suitable for this classification task.
4. **Overall Performance**: Overall, the LSTM model demonstrates superior performance compared to other traditional machine learning models such as logistic regression, Naive Bayes, Support Vector Classifier (SVC), Random Forest Classifier, and XGBoost Classifier. It achieves the highest accuracy, precision, recall, and F1-score among the listed models on the test data. This indicates that the LSTM model is the most effective in capturing the underlying patterns in the data and making accurate predictions for future purchases based on behavioral analysis.

**7. Impact of the Project**

Understanding customer sentiment in e-commerce is paramount for businesses as it transcends mere technical aspects, profoundly impacting society. By meticulously analyzing customer reviews and feedback, companies can glean invaluable insights into consumer preferences and concerns, enabling them to tailor their products and services to meet evolving demands effectively. This deep understanding not only enhances customer satisfaction but also fosters trust in online platforms, crucial for sustaining the growth and integrity of digital commerce. This trust is pivotal in encouraging more individuals to embrace online shopping with confidence, thereby bolstering the overall ecosystem of e-commerce.

Moreover, sentiment analysis in e-commerce data holds the potential to influence broader market trends and consumer behavior. By discerning patterns in customer sentiment, businesses can proactively anticipate shifts in preferences and adjust their strategies accordingly. This proactive approach not only empowers individual companies to stay ahead of the curve but also fosters innovation and competitiveness within the e-commerce landscape. Ultimately, by harnessing sentiment analysis effectively, businesses can create more personalized and impactful experiences for their customers, driving positive societal outcomes such as increased consumer satisfaction, trust, and market dynamism.

**8. Conclusion and Future Work**

In conclusion, the application of sentiment analysis in e-commerce represents a pivotal step towards enhancing customer satisfaction, fostering trust, and shaping market dynamics. By leveraging advanced analytics techniques to decipher customer feedback, businesses can adapt their strategies to meet evolving consumer needs, ultimately contributing to a more robust and customer-centric e-commerce ecosystem. The societal impact of such endeavors extends beyond individual transactions, influencing broader trends in online commerce and consumer behavior.

Looking ahead, future work in this domain could focus on refining sentiment analysis algorithms to capture nuanced emotions and cultural variations more effectively. Additionally, exploring the integration of sentiment analysis with other data sources, such as social media and demographic data, could provide deeper insights into consumer preferences and behaviors. Furthermore, efforts to democratize access to sentiment analysis tools and insights could empower smaller businesses and entrepreneurs to compete more effectively in the e-commerce landscape. Ultimately, continued innovation and collaboration in sentiment analysis hold the promise of driving further advancements in customer experience, trust-building, and market intelligence within the realm of e-commerce.

**9. Learning Outcomes**

* Understanding of sentiment analysis techniques and their application in the e-commerce domain.
* Proficiency in leveraging advanced analytics techniques to decipher customer feedback and adapt business strategies accordingly.
* Ability to analyze market dynamics and consumer behavior trends using sentiment analysis insights.
* Skills in refining sentiment analysis algorithms to capture nuanced emotions and cultural variations effectively.
* Knowledge of integrating sentiment analysis with other data sources for deeper insights into consumer preferences and behaviors.
* Understanding of the societal impact of sentiment analysis in shaping trust, customer satisfaction, and market competitiveness in e-commerce.
* Ability to collaborate and innovate in the development of sentiment analysis tools and methodologies for driving advancements in customer experience and market intelligence.

**10. References**

* [Brazilian E-Commerce Public Dataset by Olist](https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce)
* [E-Commerce Sentiment Analysis: EDA + Viz + NLP](https://www.kaggle.com/code/thiagopanini/e-commerce-sentiment-analysis-eda-viz-nlp)
* [E-Commerce System for Sale Prediction Using Machine Learning Technique](https://iopscience.iop.org/article/10.1088/1742-6596/1712/1/012042/meta)