A Novel Approach to Predict Customer Satisfaction Level Using ML Algorithm

Abstract

Customer satisfaction is an essential metric in the competitive e-commerce sector since it influences consumer loyalty and brand reputation. Lazada faces challenges in maintaining high customer satisfaction due to rising demand, requiring proactive ways to increase customer satisfaction. This project explores the various Machine Learning (ML) models for predicting customer satisfaction. The results show that the Random Forest (RF) model is the best for predicting customer satisfaction, having the highest accuracy score compared to other models. By predicting customer satisfaction, Lazada may improve its marketing strategies and discount guidelines, resulting in better customer satisfaction.

1.0 Introduction

The world e-commerce business has been quickly expanding since 2020, and it is expected to reach $5 trillion by 2024, with a 15% annual growth rate from 2019 to 2024 (Uzunoglu, 2024). Lazada's efforts on improving technology, logistics, and payment infrastructure ensure its ability to cater to the growing market, providing Southeast Asians with a delightful online shopping experience (Kathryn, 2023). The platform provides a wide choice of items on a business-to-consumer (B2C) basis and has strict constraints that vendors must follow. Lazada's e-commerce market has grown rapidly over the last decade, with the broadest assortment of brands and merchants (Kathryn, 2023).

Although technology innovation provides ease to customers, Lazada faces challenges in keeping a high customer satisfaction level. The satisfaction level of customers is a key indicator in the e-commerce industry, as it affects brand reputation and revenue. As a result, predicting customer satisfaction levels based on various factors is critical for taking proactive steps to improve the online shopping experience while maintaining a competitive advantage.

In this project, a combined methodology will be used to address Lazada's difficulties, integrating a literature survey and ML model development. A review of several approaches and ML models used by e-commerce enterprises to forecast and sustain high customer satisfaction is carried out. Next, a dataset is imported into Google Colab for visualization and preprocessing. After cleaning and preparing the data, ML models are developed to find key factors in predicting customer satisfaction. The findings are then evaluated and analysed to generate meaningful suggestions to Lazada in addressing customer satisfaction issues.

2.0 Literature Survey

Lazada confronts issues in maintaining high customer satisfaction as the number of customers increases, necessitating proactive measures to predict customer satisfaction levels by addressing various features. This section will discuss five significant studies on various ML algorithms used to estimate customer satisfactions in the e-commerce industry.

The study by Le et al. (2023) on the Shopee platform reveals that pricing has a substantial impact on customer satisfaction. They developed ML models to predict and classify customer satisfaction based on topic prediction results from deep learning models and other account criteria. In terms of accurately classifying customer satisfaction levels, the XGBoost model outperformed the Support Vector Machine (SVM), and RF models, with the greatest accuracy of 81.45 percent. However, there is a significant misclassification of moderate and high satisfaction levels. This is demonstrated by the misclassified cases in the confusion metrics, implying that some real occurrences of moderate satisfaction are projected as high satisfaction. Therefore, this model may struggle to distinguish between moderate and high satisfaction. By resolving the issues of misclassification, this paper can strengthen the reliability and robustness of its findings.

According to Wong & Marikannan (2020), customer insights contribute significantly to the success of e-commerce; thus, factors influencing customer satisfaction should be investigated thoroughly. Regardless of data preparation methods or feature engineering, the four ML algorithms used in this model consistently achieved accuracy rates ranging from 87.0 to 87.5 percent. RF is picked as the best model due to its high accuracy, sensitivity, and specificity. Nonetheless, it should increase in specificity since predicting and identifying unsatisfied customers is a key requirement in this paper. Additionally, it is critical to strike a balance between accuracy and computational speed because speed is critical in processing the large amount of real-world data accurately.

Noori (2021) mentions that customer reviews are useful sources of marketing intelligence and sentiment analysis. This paper compares five ML algorithms to provide a novel framework for identifying and predicting customer satisfaction. The Decision Tree (DT) model achieves the highest accuracy of 98.9 percent, which is quite high for a real-world situation. Thus, the DT algorithm is chosen as the best model as it has the best values of 0.98 for recall and 0.99 for precision in the dataset, resulting in a high F-measure. Moreover, it is important to ensure that the dataset is balanced before training the model. Although the model performs effectively with high accuracy in this paper, the impact of data imbalance has not been completely investigated. Hence, this paper can be improved by handling the imbalanced dataset to produce more robust and reliable findings.

According to Rahib et al. (2024), anticipating customer churn is critical for developing strategies to improve customer loyalty and revenue projection. To achieve this goal, various classification techniques is used to predict customer churn. To avoid overfitting, this paper will choose the algorithms that produces the smallest accuracy difference between train and test data. Therefore, SVM algorithm, which achieve the best performance with the lowest accuracy difference, high performance metrics and the maximum test data accuracy of 83.45 percent. While certain hyperparameters are specified, there is no evidence of extensive hyperparameter adjustment for ML models. Thus, further tuning with cross validation or grid search can increase the model's accuracy and robustness.

Lastly, the large amount of e-commerce transactions causes new concerns to arise, which are transaction fraud. To tackle this issue, Saputra & Suharjito (2019) employ ML algorithms to detect fraud in e-commerce enterprises. Various classification techniques are used to categorize e-commerce transactions as fraudulent or non-fraudulent. The Neural Network (NN) technique without SMOTE achieves 96 percent accuracy; however, the accuracy of the NN algorithm with SMOTE is 85 percent. This might be owing to the use of genetic algorithms in ANN, which determine the number of hidden nodes and layers and select important characteristics for NN model, resulting in increased accuracy. While the paper interpreted the results with standard metrics, it should incorporate the Area Under the Receiver Operating Characteristic (ROC-AUC) curve, which can better evaluate the model's effectiveness.

Based on the five papers reviewed, several ML algorithms were used to conduct various analyses in the e-commerce industry, including customer satisfaction, customer churn, and fraud detection prediction. Although the models work excellently with high accuracy, the papers examined did not include hyperparameter tuning, which can build a more robust prediction model and improve prediction accuracy. Hyperparameter tuning approaches, such as grid search and random search, will be used in this project to increase model performance.

Table 1: Summary of Classification Reports

Authors	Model		Result			
			Accuracy (%)	Precision	Recall	F1-Score
Le et al. (2023)	**XGBoost**		**81.4**	**0.81**	**0.81**	**0.81**
	RF		77.1	0.75	0.77	0.74
	SVM		78.9	0.73	0.79	0.76
Wong & Marikannan (2020)	DT		87.2	-	0.98	0.93
	RF		**87.5**	**-**	**0.97**	**0.93**
	ANN		87.3	-	0.98	0.93
	SVM		87.1	-	0.98	0.93
Noori (2021)	SVM		77.5	0.78	0.77	0.78
	NN		55.1	0.53	0.50	0.51
	NB		70.6	0.71	0.70	0.70
	DT		**98.9**	**0.99**	**0.98**	**0.99**
	KNN		72.0	0.75	0.72	0.73
Rahib et al. (2024)	XGBoost		83.3	0.80	0.83	0.77
	DT		83.3	0.80	0.83	0.77
	RF		82.1	0.79	0.82	0.80
	SVM		**83.4**	**0.84**	**0.83**	**0.76**
Saputra & Suharjito (2019)	Without SMOTE	DT	91.0	0.54	0.59	0.56
		NB	95.0	0.91	0.54	0.67
		RF	95.0	0.95	0.55	0.69
		ANN	**96.0**	**0.97**	**0.54**	**0.97**
	With SMOTE	DT	91.0	0.91	0.60	0.91
		NB	95.0	0.94	0.54	0.94
		RF	95.0	0.80	0.58	0.94
		ANN	85.0	0.92	0.76	0.85

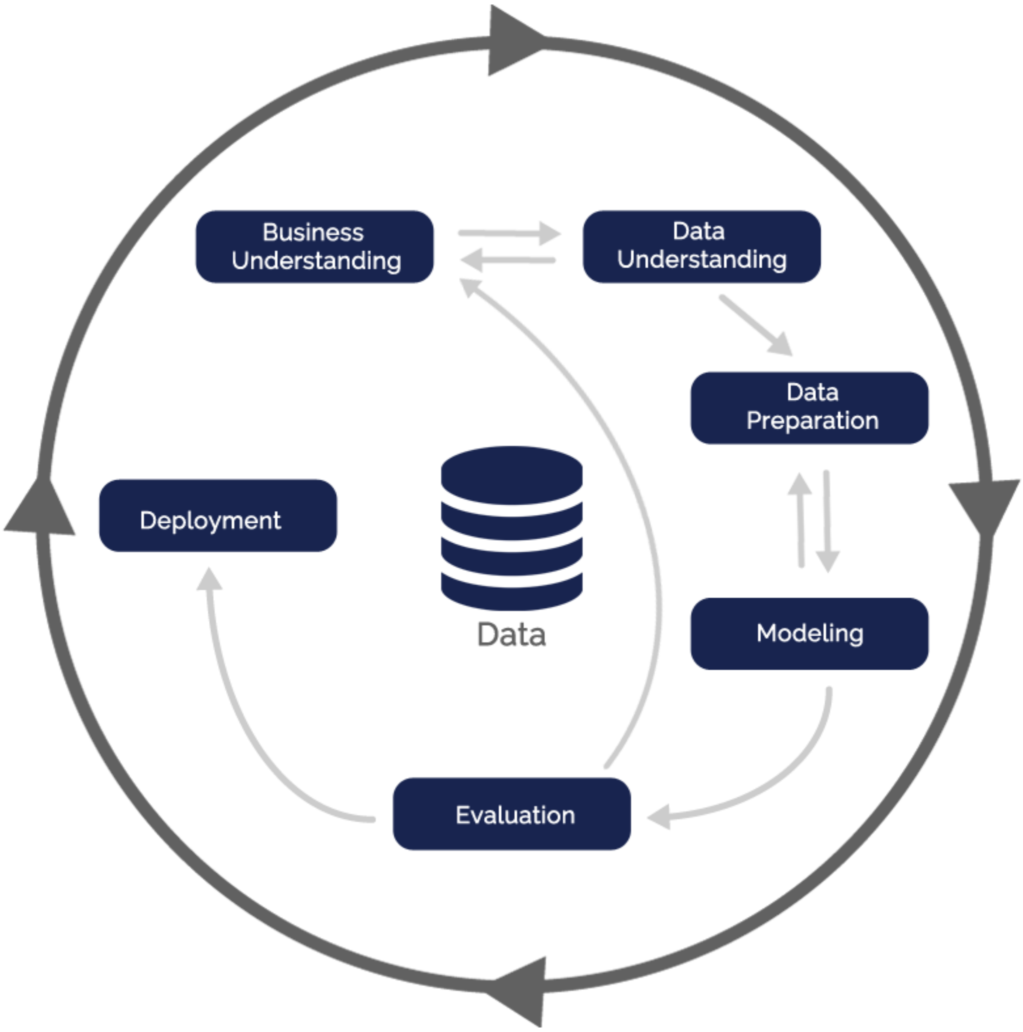
Notes. KNN stands for K-Nearest Neighbors.

3.0 Methodology

3.1 Standard Process used to Perform ML

Figure 1: CRISP-DM Methodology

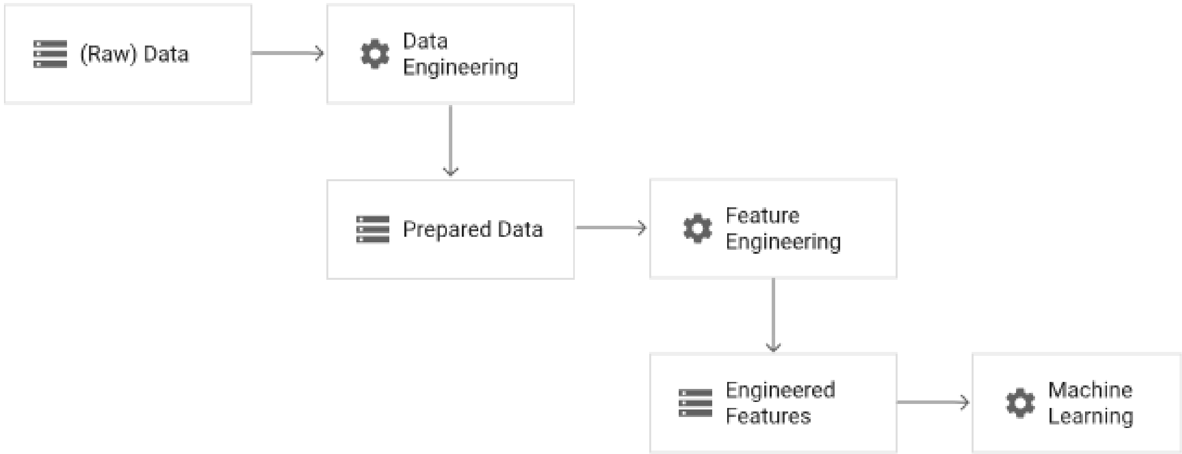
(Sources: https://analyticsindiamag.com/ai-origins-evolution/crisp-dm-data-science-project/)



According to Luna (2022), Cross Industry Standard Process for Data Mining (CRISP-DM) is a data mining approach divided into six phases as shown in Fig. 1, each with its own set of activities and relationships. The model is adaptable and adjustable, making it appropriate for a variety of purposes (Luna, 2022). Firstly, the business understanding phase focuses on project objectives and needs, whereas the data understanding phase involves obtaining and exploring data using visualizations (Schröer et al., 2021). Data preparation is the most crucial step, involving aggregating records, generating new features, and eliminating or replacing missing values, as seen in Fig. 2 (Schröer et al., 2021).

Figure 2: The Flow of Data Preprocessing

(Sources: https://www.tensorflow.org/tfx/guide/tft_bestpractices)

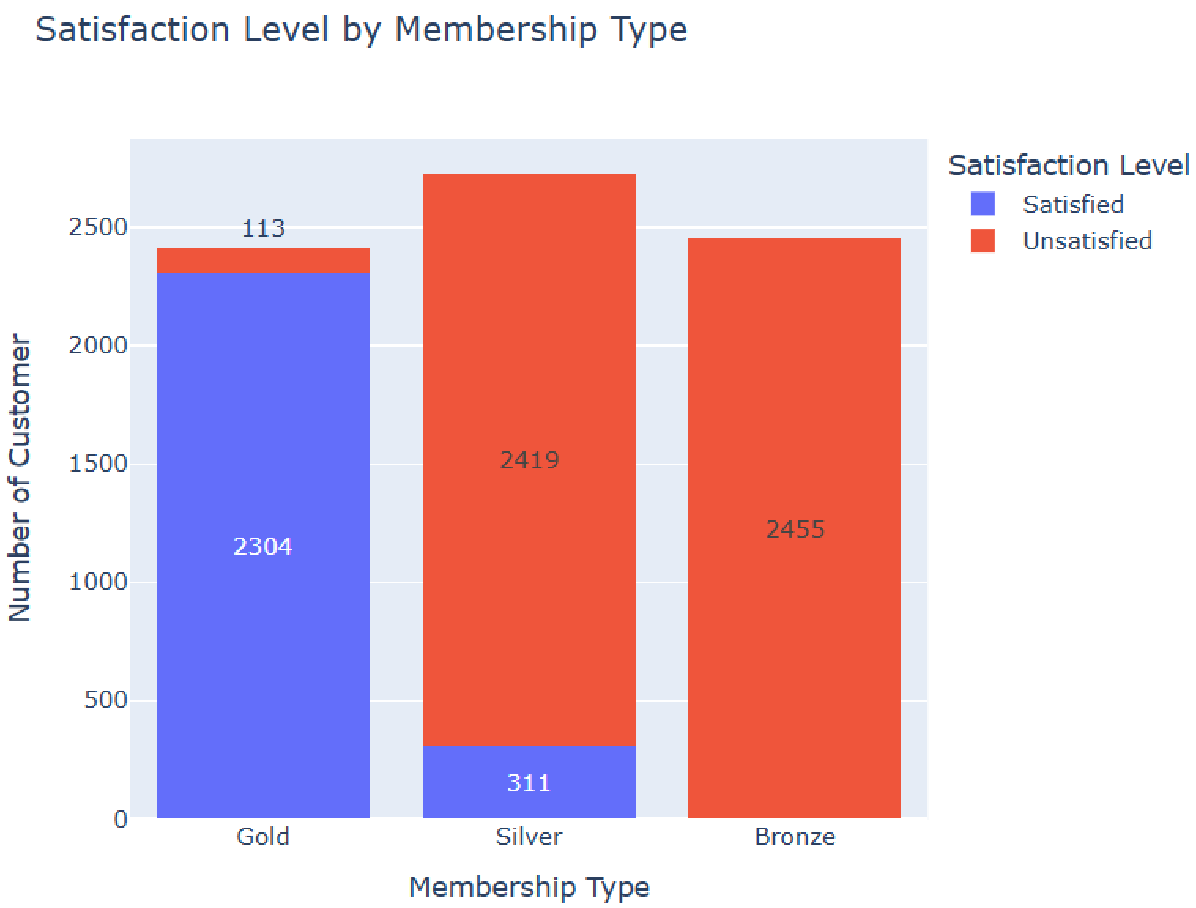


Modelling process includes introducing prepared data into analytic tools, which are Google Colab, and running numerous models with the default parameters, then fine-tuning the parameters or returning to the data preparation step for model-specific modifications (Hotz, 2024). The evaluation process then confirms that the models are functional, and the performance metrics of the models are measured against the business success criteria. Finally, during the deployment phase, new insights generated from the prediction model are used to enhance the marketing strategies (Hotz, 2024).

3.2 Exploratory Data Analysis

This project's dataset contains 7699 rows, nine variables, and a target variable that indicates whether customers are satisfied or unsatisfied with their shopping experience on Lazada. To better understand the datasets, various features are used to generate visualizations with the target variable, which is customer satisfaction level.

Figure 3: Customer Satisfaction Level by Membership Type



Based on Fig. 3, gold members are the most satisfied, followed by silver and bronze. Gold members, who are satisfied, may be able to earn bigger discounts than normal members. With these advantages, most gold members are satisfied with the Lazada platform. However, most silver and bronze members are disappointed with the Lazada platform, possibly owing to a lack of special bonuses and high shipping fees.

Figure 4: Scatter Plot of Total Spend by Item Purchased and Satisfaction Level

Based on Fig. 4, the total amount spent increases when the number of items purchased increases, resulting in better customer satisfaction. Customers are more likely to spend on more items that meet their expectations, indicating a pleasant shopping experience on Lazada. However, Fig. 4 above does not show a strong relationship between these two variables as the data points are dispersed and lack a distinct pattern; hence, more in-depth cleaning is required.

Figure 5: Number of Satisfied and Unsatisfied Customer by Average Rating

Fig. 5 illustrates a general trend in which the number of satisfied customers increases as the average rating increases. Customers who give an average rating of 3 or 4 are dissatisfied with their purchases made on Lazada. An average rating of 4 to 5 suggests that the customer is pleased with the items purchased. However, there is a noticeable spike in dissatisfied customers, with an average rating of 4.1, indicating that most customers in this dataset are dissatisfied with their purchases on Lazada.

3.3 Data Pre-processing

There are a total of 1995 rows with null values distributed across different columns. As a result, numerical columns with null values are replaced with the median, and categorical columns with null values are replaced with mean values for each column. Outliers in the numerical columns may cause the training data to overfit or underfit. Hence, outliers are removed by ensuring that they are within the column's lower and upper limits.

There are 97 duplicate rows in this dataset, which will be dropped since there is sufficient data to train the ML model. After removing the duplicate rows, there are 7602 rows remaining for analysis. Data inconsistencies might lead to misinterpretation of results; thus, inconsistent data in the numerical column is handled by specifying the data types and decimal numbers for each column. However, inconsistent data in categorical columns will be replaced. For instance, "KL" is replaced with "Kuala Lumpur" because they represent the same state. Next, correlation analysis is performed both before and after data cleaning, as shown in Fig. 6. After data cleaning, the numerical columns have a stronger correlation.

Figure 6 (a): Correlation Analysis before Data Cleaning

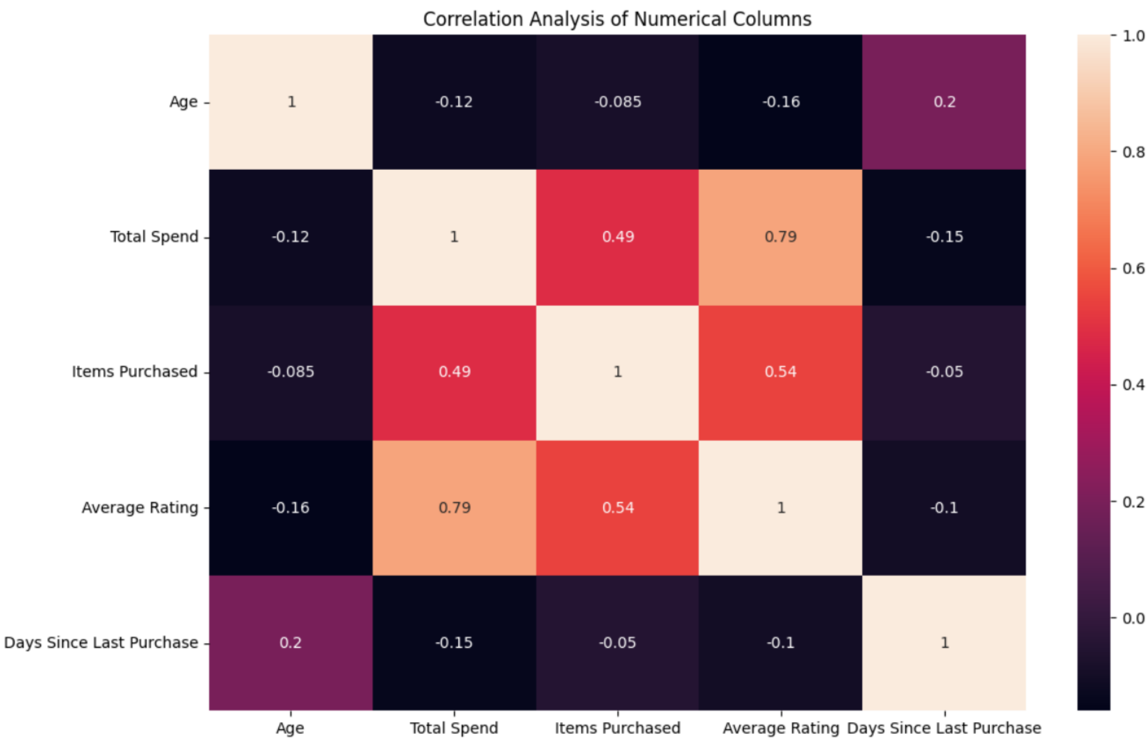
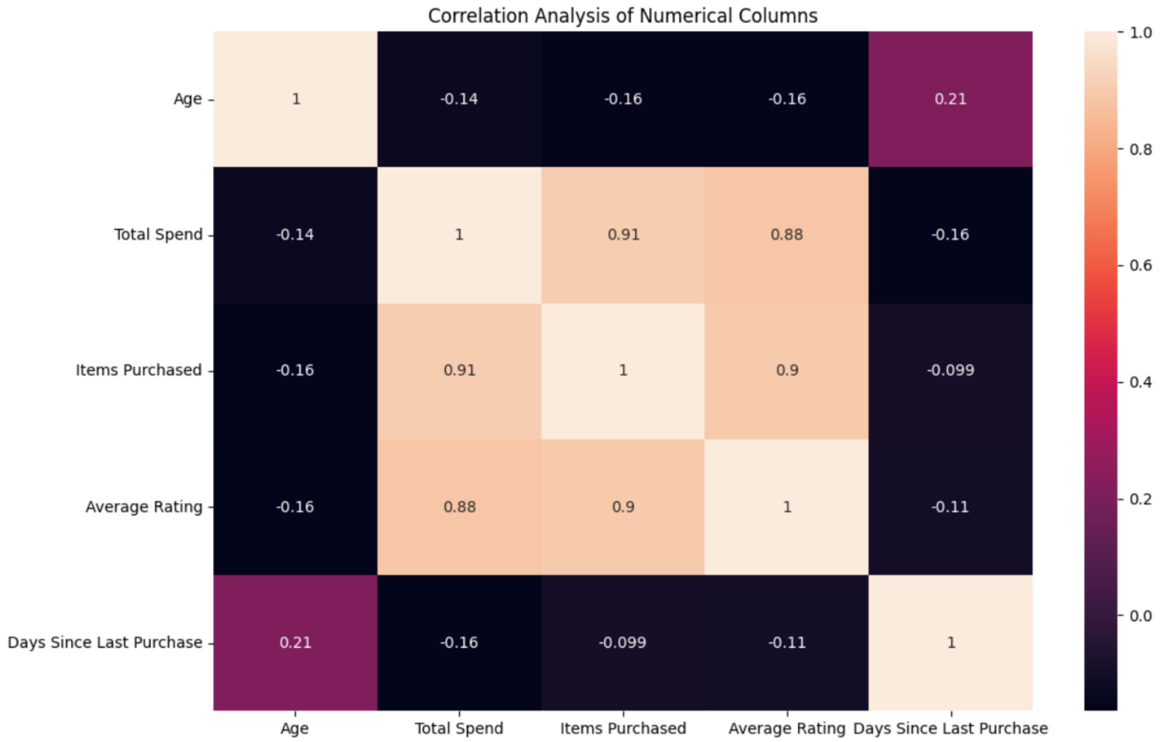


Figure 6 (b): Correlation Analysis after Data Cleaning



Moreover, categorical data are encoded using two different encoders. Label Encoder is used for variables with two unique classes, including gender, discount applied, and satisfaction level, whereas One Hot Encoder is used for variables with three or more unique classes, including state and membership type. This is because Label Encoding is commonly used for binary classification; however, One Hot Encoding works better with categories as it prevents the model from assuming an ordinal relationship between two categories, which can lead to incorrect predictions (Potdar et al., 2017).

Additionally, the satisfaction level distribution is unequal with 2615 rows satisfied and 4987 unsatisfied. When one class outnumbers another, the dataset becomes skewed. An uneven dataset may cause the ML model hard to learn pattern, resulting in poor performance. As seen in Fig. 7, using Synthetic Minority Over-sampling Technique (SMOTE) to produce synthetic samples for minority classes can help in balancing the target variable.

Figure 7: Result of Balancing Target Variable with SMOTE

```
[620] # Initialize SMOTE
      smote = SMOTE()

      # Oversampling the separated dataset with SMOTE
      new_features, new_target = smote.fit_resample(feature_smote, target_smote)

      # Check the new distribution after applying SMOTE
      Counter(new_target)

      Counter({0: 4987, 1: 4987})
```

Next, data is normalized using the MinMaxScaler to improve model performance by ensuring that larger scale features do not dominate the learning process, allowing the model to analyse all features equally (Pelletier, 2024). Furthermore, the dataset's most relevant features are selected using SelectKBest with mutual_info_classif. The top five features chosen, as illustrated in Fig. 8, are used to predict customer satisfaction.

Figure 8: Result of Feature Selection

```
# Show the shape of the array to check the number of features have been selected
print(f'Shape of the array: {X_selected.shape}')

# Show the top k features
print(f'\nTop {k} features selected: ')
feature_scores.head(k)
```

```
Shape of the array: (9974, 5)

Top 5 features selected:
```

	Feature	Score
2	Total Spend	0.603373
4	Average Rating	0.573289
3	Items Purchased	0.522875
24	Membership Type_Gold	0.451362
6	Days Since Last Purchase	0.334248

Following that, Principal Component Analysis (PCA) is used to reduce noise and overfitting by decreasing the number of features in the dataset while maintaining as much information as possible (Lever et al., 2017). This dataset was trimmed from 9974 to two rows. As shown in Fig. 9, Membership Type_Gold has the largest loading, suggesting that it has the most significant impact on Principal Component (PC) 1. However, Days Since Last Purchase has the largest positive loading on PC2, indicating that it has the most impact on PC2.

Figure 9: Result of Principal Component Loadings

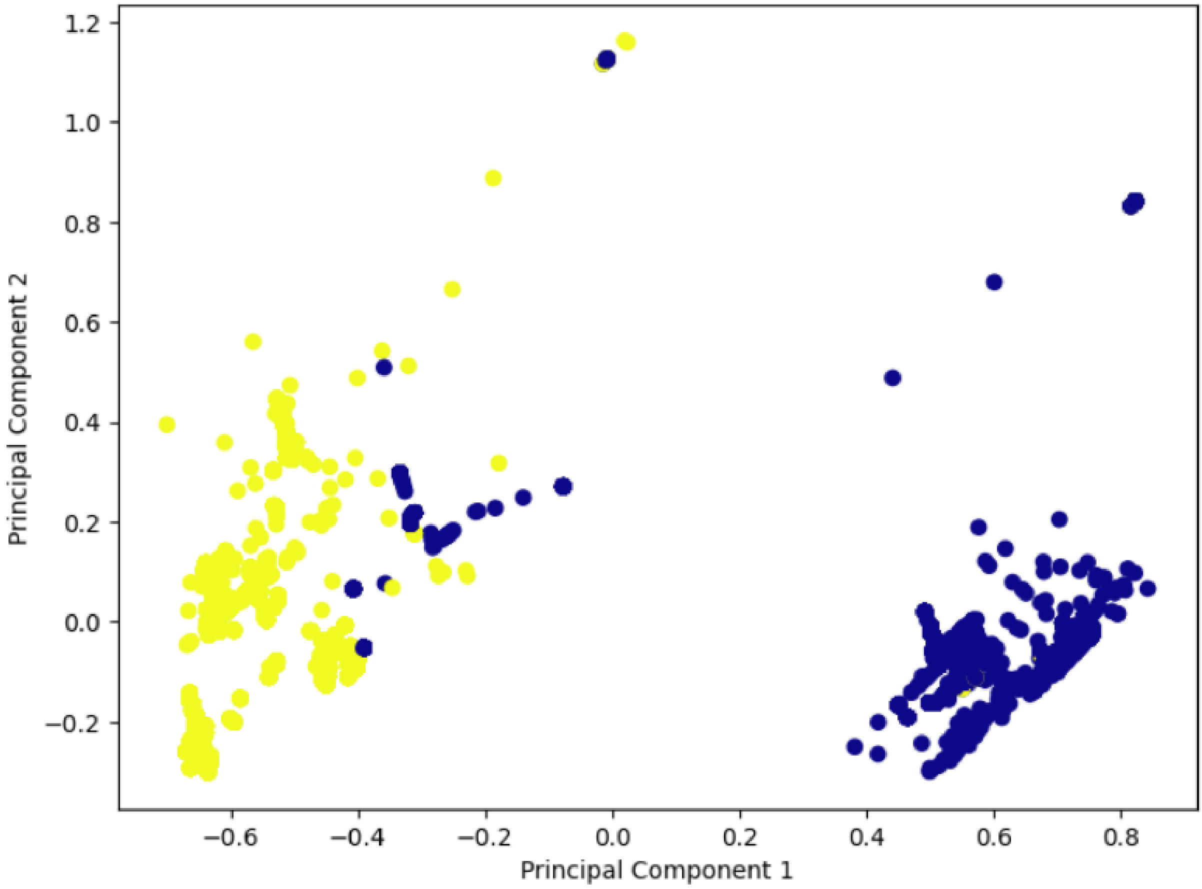
```
[ ] print("Principal Component Loadings:")

    # Verify the number of pca components
    pd.DataFrame(pca.components_.T, columns=['PC1', 'PC2'], index=feature_names)
```

```
Principal Component Loadings:
```

	PC1	PC2
Total Spend	0.279063	0.295559
Items Purchased	0.382431	0.473152
Average Rating	0.248349	0.329588
Days Since Last Purchase	-0.153491	0.706334
Membership Type_Gold	0.831044	-0.285020

According to Fig. 10, the data points are projected onto the two primary components, with two colours denoting two unique clusters. The right cluster has greater PC1 values than the left cluster, suggesting a group of satisfied customers with gold membership status, more purchases, and a higher total amount spent. However, the left cluster shows low or negative PC1 values, indicating that customers who made recent purchases are unsatisfied.

Figure 10: Scatter Plot of Principal Components



3.4 Classification Analysis and Technique

Classification analysis is a data mining approach that identifies patterns or groups between individual observations, helping researchers to better understand their dataset and make accurate predictions (Keita, 2024). This project employs classification algorithms, including RF, NB, KNN, and NN to predict if a customer is satisfied or unsatisfied based on various factors. After preprocessing the data, it is divided and trained with different algorithms.

Firstly, the RF model's findings are aggregated using an ensemble approach that produces several DTs (Donges, 2024). To avoid overfitting, each tree is trained with different sets of features. This model was chosen as it incorporates predictions from each tree, resulting in more robust findings (Donges, 2024). Next, Gaussian NB is a probabilistic classification approach that assumes that all continuous features follow a normal distribution (Martins, 2023). It can forecast the probability of dependent features being categorized in each group. As a result, this model was chosen as it can handle continuous features by assuming a Gaussian distribution for each one (Martins, 2023).

The KNN model computes both the number of neighbors and the distance between data points (Halder et al., 2024). It is a non-parametric approach to categorizing samples by using the majority of its KNN in the feature space. Thus, it was picked since it adapts to different patterns and forecasts based on data structure (Halder et al., 2024). The NN model is a complicated network comprised of many fundamental nodes known as neural cells (Stewart, 2021). Combining non-linear transfer functions produces input mappings and projected output responses. This method was selected as it can discover hidden patterns (Stewart, 2021).

In this project, the four models stated previously are trained by fitting them to training data. These models are trained in two ways: with and without hyperparameter tuning. Hyperparameter tuning fine-tunes model parameters using approaches, including grid search and random search, which can increase model performance by identifying the optimal hyperparameter combination (Yang & Shami, 2020). After training the models, key performance metrics are generated to evaluate its performance.

4.0 Results and Discussion

4.1 Comparing the Performance of ML Models without Hyperparameter Tuning

The RF model's accuracy across both classes is 0.9870, which means it accurately predicted 98.7 percent of all cases. This model's average accuracy, recall, and F1-score across classes are all 0.98, indicating all classes are treated equally regardless of their frequency. Hence, this model performs exceptionally well in both classes with high accuracy. The confusion matrix backs up this claim as it has high true positive and true negative values, indicating successful prediction of satisfied and unsatisfied customers. However, the fact that 19 dissatisfied customers were misclassified suggests that further fine-tuning of the algorithm may better capture unsatisfaction patterns.

The Gaussian NB algorithm has 0.9298 accuracy overall, which is slightly lower than the previous model. The precision for satisfied customers is higher than unsatisfied customers, indicating most of the instances predicted as satisfied were truly satisfied. However, the sensitivity will reduce to attain higher precision. Therefore, the sensitivity for dissatisfied customers is greater than for satisfied customers, implying nearly all the unsatisfied customers were captured. Furthermore, the confusion matrix shows a larger number of false positives than false negatives, indicating a tendency to overestimate dissatisfied customers. Thus, this model will not be used for hyperparameter tuning.

The KNN model's accuracy across both classes is 0.9840 with an average precision, recall, and F1-score of 0.98. This treats both satisfied and unsatisfied customers equally without regard to their occurrence in this dataset. This claim is supported further by the confusion matrix, which has only 48 misclassified instances out of 2993. The error distribution between false positives and false negatives is balanced with 30 for satisfied customers and 18 for unsatisfied customers, demonstrating that the model does not prefer one class over another. However, the overall accuracy score is slightly lower than the RF model, indicating further enhancement is needed.

Lastly, the NN model achieves an accuracy of 0.9576 while maintaining good precision and recall values. This model has an input layer, three hidden layers with ReLU activation function, and an output layer with sigmoid activation function. It compiles the model using the HeNormal initializer, Adam optimizer, and binary cross-entropy loss. This model is trained with fixed hyperparameters: $epoch = 50,$ $batch\ size = 32,$ $validation\ split = 0.2,$ $verbose = 1$. By setting the model in this manner, the model's average precision, recall, and F1-score across both classes are 0.95. Additionally, the confusion matrix shows a significant rate of misclassification, with 127 instances misclassified out of 2993. Hence, this model will not perform hyperparameter tuning due to high misclassifications and a long training time.

A summary of performance metrics and confusion matrix of each ML model mentioned previously is shown as follows:

Table 2: Classification Report of ML Models without Hyperparameter Tuning

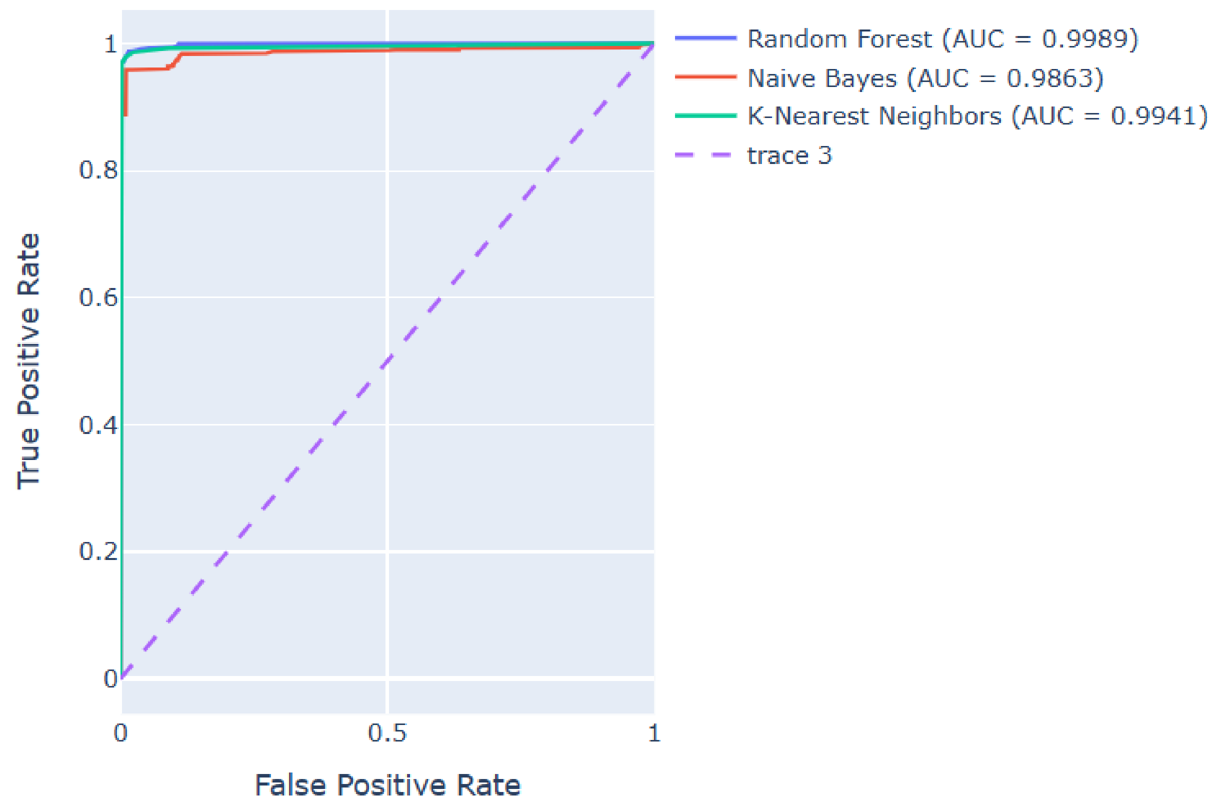
Model	**Class 0 (Satisfied)**			**Class 1 (Unsatisfied)**			**Accuracy**
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	
RF	0.9875	0.9869	0.9872	0.9864	0.9871	0.9867	0.9870
Gaussian NB	0.9831	0.8774	0.9272	0.8854	0.9843	0.9323	0.9298
KNN	0.9805	0.9882	0.9843	0.9876	0.9796	0.9836	0.9840
NN	0.9788	0.9370	0.9575	0.9374	0.9789	0.9577	0.9576

Table 3: Confusion Matrix of ML Models without Hyperparameter Tuning

Model	**True Negative**	**False Positive**	**False Negative**	**True Positive**
RF	1505	20	19	1449
Gaussian NB	1338	187	23	1445
KNN	1507	18	30	1438
NN	1429	96	31	1437

Fig. 11 depicts an AUC-ROC curve comparing the performance of three ML models, including RF, NB, and KNN, based on their AUC values. With an average AUC of 0.99, both the RF and KNN models perform excellently, almost reaching the upper left corner of the ROC space. This high AUC indicates that it has an outstanding ability to differentiate between positive and negative classes at various thresholds (Nahm, 2022). As a result, these two models are trained with hyperparameter tuning to determine which models are better.

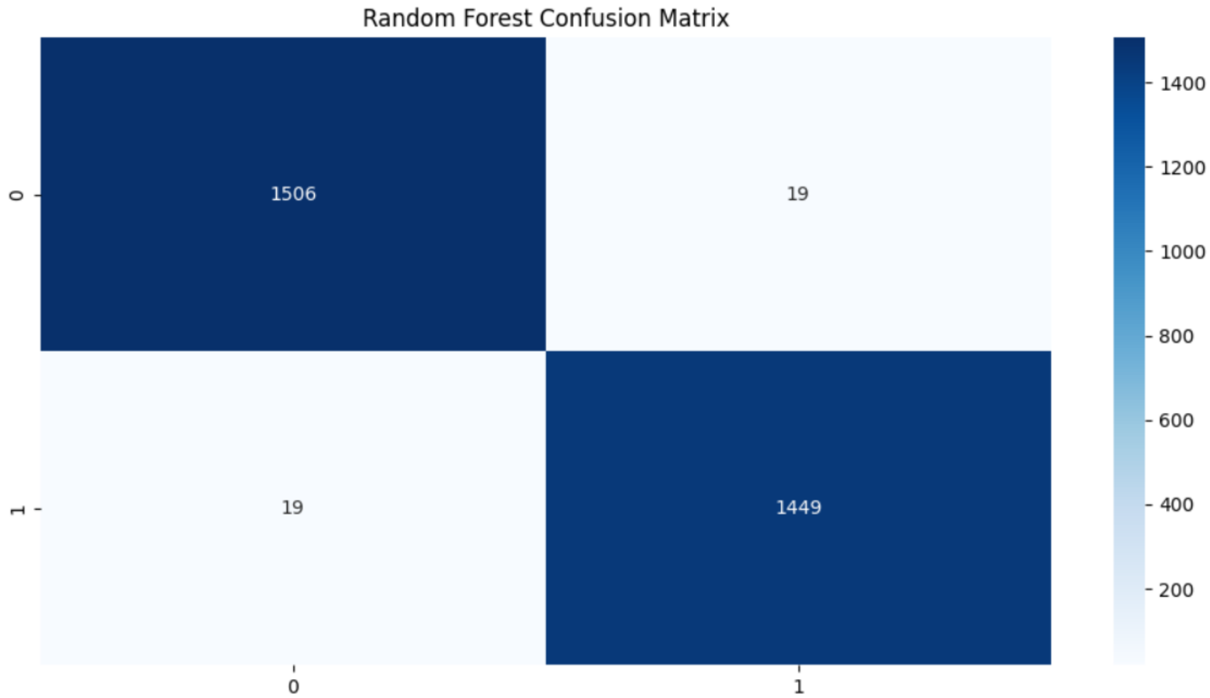
Figure 11: AUC-ROC Curve of ML models without Hyperparameter Tuning



4.2 Comparing the Performance of ML Models with Hyperparameter Tuning

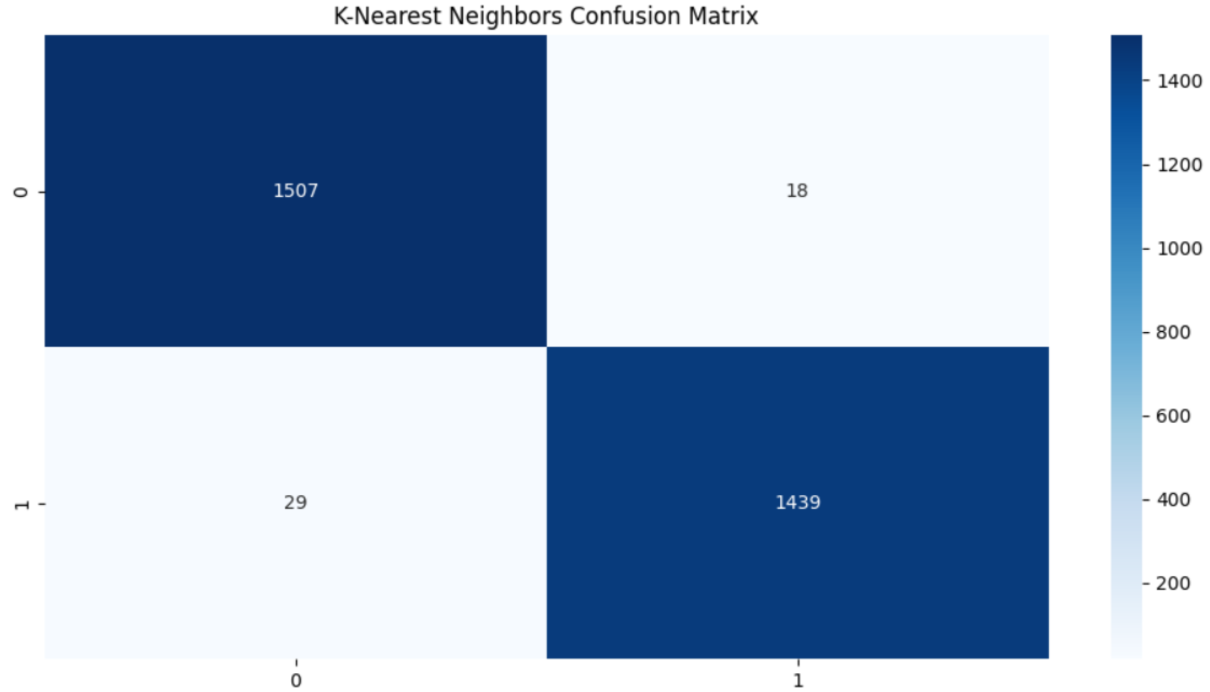
The two models with the best performance, which are RF and KNN models, are trained with hyperparameter tuning to achieve better results. After hyperparameter tuning, the accuracy of RF model has increased from 0.9870 to 0.9873, indicating the hyperparameters were effectively adjusted for better performance. This involves setting the number of estimators to 200, maximum features to “log2”, and maximum depth to 15, which results in increased accuracy for the RF model. Furthermore, the confusion matrix in Fig. 12 shows that the model is very accurate in predicting both classes with just 38 wrong predictions out of 2993 samples. The low number of misclassifications indicates that the model’s current setup is very effective for this project.

Figure 12: Confusion Matrix of the RF Model



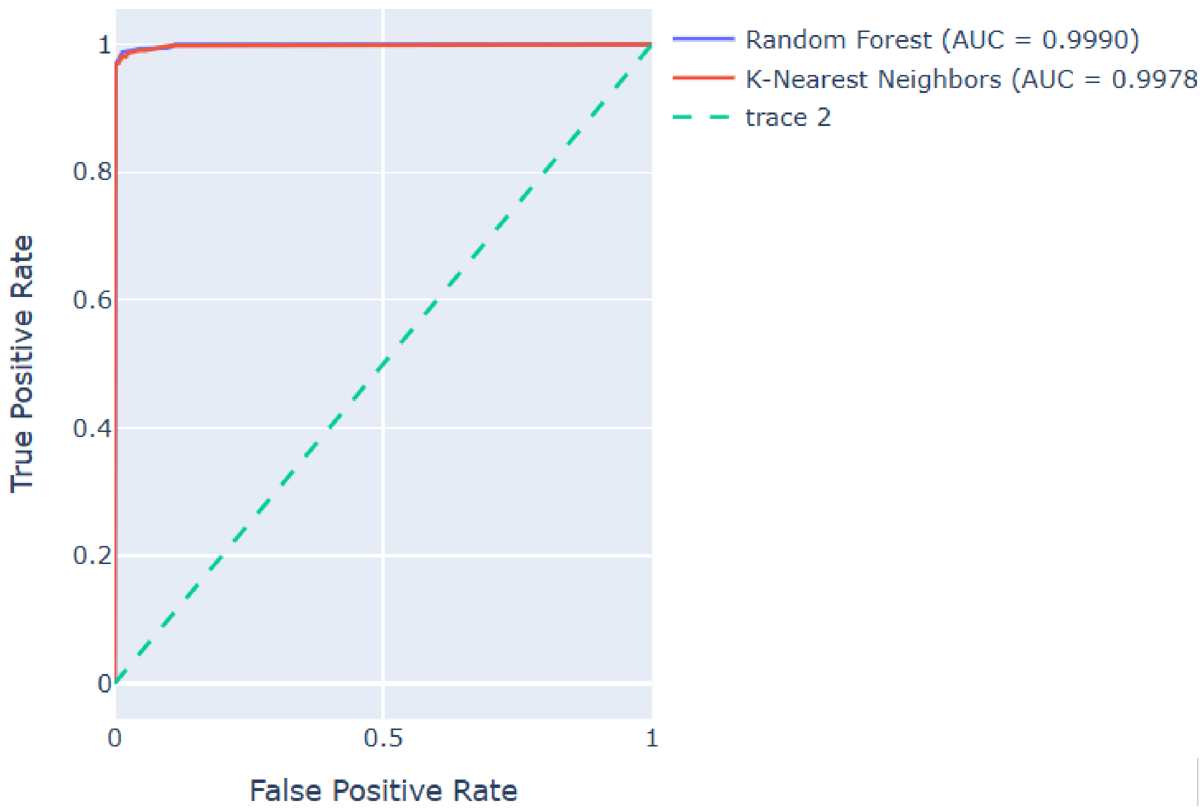
Moreover, the accuracy of the KNN model has increased from 0.9840 to 0.9843 after hyperparameter tuning. With hyperparameter modifications, including metric set to euclidean, number of neighbors set to 13, and weights set to distance, the KNN model slightly improves. The confusion matrix in Fig. 13 illustrates that there are few misclassifications, with 18 false positive and 29 false negatives, implying that the model is slightly underpredicting unsatisfied customers when compared to the RF model.

Figure 13: Confusion Matrix of KNN Model



Futhermore, the RF model has an AUC score of 0.9990, indicating that the model is nearly perfect in differentiating between satisfied and unsatisfied customers. However, the KNN model with the AUC score of 0.9978 still shows that the model can clearly distinguish between two classes, but there might be marginal differences in its ability to correctly classify certain instances compared to the RF model; hence, it is less powerful than the RF model.

Figure 14: AUC-ROC Curve of ML models with Hyperparameter Tuning



Overall, the RF model with hyperparameter tuning outperformed all of the other models in this project, with the greatest accuracy score and the fewest misclassifications. This model can accurately capture the factors influence customer satisfaction and provide reliable predictions. With high AUC scores, the RF model may accurately identify customers who are unsatisfied with their purchases and enhance decision making to mitigate the issue effectively. High performing RF models minimize misclassification, resulting in more effective interventions and reduced costs associated with inaccurate predictions. This optimization reduces operating expenses since resources are used more efficiently. Ultimately, a high-performance RF model allows Lazada to handle issues more proactively and operate in a more data-driven and competitive manner, resulting in a better reputation and a more resilitent market position.

5.0 Conclusion

In conclusion, a high-performance ML model is developed to predict customer satisfaction on Lazada. Using diverse data and classification techniques, the model has accurately classified customer satisfaction, generating significant insights into the factors impacting customer satisfaction. Although the RF model has performed excellently, several limitations must be addressed. The quality and quantity of training data have a significant impact on the model's performance. Therefore, only 7602 rows for model training are inadequate and might result in biased predictions. While the RF model is generally easy to understand, it can grow more complicated as depth increases. Overfitting can occur when the model becomes very complicated and fits the training data too closely, resulting in poor generalization performance. For future improvements, more diversified and comprehensive data should be collected to improve the model's generalizability. Advanced ML approaches, including ensemble methods and deep learning, have the potential to improve model performance. To react to changing customer preferences and market trends, the model's performance should be checked regularly with new data. All in all, a great customer satisfaction prediction model empowers businesses to make data-driven decisions and improve overall customer experience on Lazada.

6.0 References

Anirudh, V. K. (2020, September 3). *How the CRISP-DM method can help manage your next data science project* [Online Image]. Analytics India Magazine. https://analyticsindiamag.com/ai-origins-evolution/crisp-dm-data-science-project/

Data preprocessing for ML: options and recommendations [Online Image]. (2024, September 6). TensorFlow. https://www.tensorflow.org/tfx/guide/tft_bestpractices

Donges, N. (2024, March 8). *Random Forest: A complete guide for ML*. Built In. https://builtin.com/data-science/random-forest-algorithm

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-024-00973-y

Hotz, N. (2024, April 28). *What is CRISP DM?* Data Science Process Alliance. https://www.datascience-pm.com/crisp-dm-2/

Kathryn, R. (2023, August 28). *Top Southeast Asia ecommerce companies: Shopee and Lazada.* Kathryn Read. https://kathrynread.com/top-southeast-asia-ecommerce-companies-shopee-lazada-and-co/

Keita, Z. (2024, August 8). *Classification in ML: An Introduction*. https://www.datacamp.com/blog/classification-machine-learning

Le, H., Huynh, T., DO, Nguyen, M. H., Tran, H., Pham, T. T., Nguyen, N. T., & Nguyen, V. (2024). Predictive model for customer satisfaction analytics in E-commerce sector using ML and deep learning. *International Journal of Information Management Data Insights*, *4*(2), 100295. https://doi.org/10.1016/j.jjimei.2024.100295

Lever, J., Krzywinski, M., & Altman, N. (2017). Principal component analysis. *Nature Methods*, *14*(7), 641–642. https://doi.org/10.1038/nmeth.4346

Luna, Z. (2022, August 24). Understanding CRISP-DM and its importance in Data Science projects. Medium. https://medium.com/analytics-vidhya/understanding-crisp-dm-and-its-importance-in-data-science-projects-91c8742c9f9b

Martins, C. (2023, November 2). *Gaussian naive Bayes explained with Scikit-Learn*. Built In. https://builtin.com/artificial-intelligence/gaussian-naive-bayes

Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean Journal of Anesthesiology*, *75*(1), 25–36. https://doi.org/10.4097/kja.21209

Noori, B. (2021). Classification of customer reviews using ML algorithms. *Applied Artificial Intelligence*, *35*(8), 567–588. https://doi.org/10.1080/08839514.2021.1922843

Pelletier, H. (2024, November 9). Data Scaling 101: Standardization and Min-Max Scaling explained. *Medium*. https://towardsdatascience.com/data-scaling-101-standardization-and-min-max-scaling-explained-60789833e160

Potdar, K., S, T., & D, C. (2017). A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications*, *175*(4), 7–9. https://doi.org/10.5120/ijca2017915495

Rahib, M. a. A., Saha, N., Mia, R., & Sattar, A. (2024). Customer data prediction and analysis in e-commerce using ML. Bulletin of Electrical Engineering and Informatics, 13(4), 2624–2633. https://doi.org/10.11591/eei.v13i4.6420

Saputra, A., & Suharjito. (2019). Fraud Detection using ML in e-Commerce. *International Journal of Advanced Computer Science and Applications*, *10*(9). https://doi.org/10.14569/ijacsa.2019.0100943

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process model. Procedia Computer Science, 181, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Stewart, M., PhD. (2021, December 10). Introduction to Neural networks - towards data science. *Medium*. https://towardsdatascience.com/simple-introduction-to-neural-networks-ac1d7c3d7a2c

Uzunoglu, C. (2024, June 24). Global eCommerce Market 2024: market growth, top players & online share. *ECDB*. https://ecommercedb.com/insights/global-ecommerce-market-2024-size-market-growth-online-share/4784

Wong, A., & Marikannan, B. P. (2020). Optimising e-commerce customer satisfaction with ML. *Journal of Physics Conference Series*, *1712*(1), 012044. https://doi.org/10.1088/1742-6596/1712/1/012044