JURNAL INFORMATIKA, Vol.12 No.1 April 2025, Page 57-67
ISSN (Print) 2355-6579 | ISSN (Online) 2528-2247
DOI: https://doi.org/10.31294/inf.v12i1.22459

Digital Marketing Strategy Optimization Using Support Vector Machine Algorithm

Ihsan AlFauzi[1], Budiman[2*], Nur Alamsyah[3]

[1,2,3] Universitas Informatika dan Bisnis Indonesia
Jalan Soekarno Hatta No. 634 Bandung, Indonesia

Correspondence e-mail: budiman@unibi.ac.id

Submission:	Revision:	Acceptance:	Available Online:
16-06-2024	27-02-2025	20-03-2025	30-04-2025

Abstract

Information and communication technology (ICT) is essential in rapidly disseminating information. This research discusses the influence of ICT use in marketing promotions through TV, radio, and social media and compares the performance of several classification algorithms in processing the promotion data. The dataset is from Kaggle, with promotional attributes on TV, radio, and social media. The Cross-Industry Standard Process for Data Mining (CRISP-DM) is used. Algorithms tested include Naive Bayes, K-Nearest Neighbor, Support Vector Machine (SVM), Random Forest, and XGBoost. The results showed that SVM had the best performance with 80% accuracy, followed by KNN (79%), Naive Bayes (77%), XGBoost (77%), and Random Forest (76%). SVM provided the most accurate and consistent predictions in marketing promotion classification. This research concludes that the optimal utilisation of ICT and the application of appropriate classification algorithms can increase the effectiveness of marketing promotions in the digital era.

Keywords: Marketing Promotion, Classification Algorithms, Information and communication technology

1. Introduction

In this current tumultuous era of globalisation, information and communication technology make dissemination of information possible at a speed and effectiveness that was unfathomable in the past. In today's society, information technology is an indispensable tool that must not be ignored, and it ensures continued technology relevance among people in the digital age. Nowadays, with the rapid growth of advancing business information and IT technologies, people can retrieve information on various platforms and media (Budiman, 2021).

Information technology has advanced tremendously in data collection and storage, and the need for information generation arises from this data. The information gathered will, therefore, be critical for making every decision in a particular situation (Alamsyah et al., 2023).

Indonesia itself is experiencing significant growth in Internet penetration; the Indonesian Internet Service Providers Association (APJII; noted that internet penetration in Indonesia has reached 78.19% in 2 Internet penetrated 215.626.156 people from a total population of 275.773.901 people (APJII, 2023), thus providing broader access to information sharing. This certainly impacts all aspects, one of which is in the marketing field, where innovative technology has become popular because it can improve conceptual business, including the company's main strategic objectives. (Al-Ababneh, 2024)

Marketing is the most crucial thing; marketing is an activity, a series of organisations and a process of communicating, distributing and exchanging value offers to customers, partners and society (Wawolumaya et al., 2022). In the digital era, where marketing strategies continue to develop, many entrepreneurs are starting to use media to carry out their marketing strategies; one of the marketing strategies commonly used is promotion. (Yoesoep, 2022)

Promotion is one of the elements that play an essential role in a business or business. With the implementation of promotions, information can be conveyed and influence potential customers to buy the products offered. (Malik et al., 2023). Online promotional media is becoming popular and is supported by growing internet users. Various promotional methods, including television and radio, can be applied, but social media is currently a common choice. (Puspitarini & Nuraeni, 2019)

Media such as television, radio, and social media certainly have their advantages and disadvantages. Media selection depends on the target audience, budget, and promotional objectives that need to be achieved. Television has wide coverage and strong visual capabilities, but the advertising costs can be high. With its

more auditive nature, radio remains a practical option for local promotional campaigns at a more affordable price. With its rapid growth, social media provides direct audience interaction and in-depth analytics opportunities. (Yoedtadi, 2019)

In this context, data mining is a relevant approach. Data mining is a process of extracting previously unknown information from data. (Budiman et al., 2020)

Data mining is a tool and application that uses statistical analysis of data by extracting or extracting previously unknown data and information (Indrayana & Solikhin, 2020). Data mining is simply the process of extracting data that leads to discovering new information by looking for patterns or rules from a large amount of data. (Ardilla et al., 2021)

The method to extract information or patterns from data is through applying techniques in data mining, including classification algorithms (Alamsyah, Budiman, et al., 2024). Classification itself is an algorithm in data mining that classifies or determines a criterion from data based on previous data that has been studied (yunial, 2020)

Based on the literature review, the data to be processed in this study is a marketing promotion dataset (Marketing Promotion TV vs Radio vs Social Media) obtained from the Kaggle.com site or platform. This dataset includes information about a business's various marketing promotions to increase sales. Each row in the dataset represents an independent marketing promotion using promotional budgets for TV, social media and radio, where the value of each attribute is numerical and categorical.

In addition, Naïve Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest, and XGBoost algorithms are popular tools in promotional marketing analysis due to their effectiveness in data pattern classification, clustering, and prediction. Based on the probability of success of a campaign, the Naive Bayes algorithm can be employed for classification. In contrast, K-NN can be utilised to compare marketing techniques with comparable data patterns. SVM helps determine how best to classify the effectiveness of promotional strategies. For the classification of large volumes of complex data, Random Forest helps decide which features are most critical to the outcome of promotions. In addition to features like budget, platform, and audience, the XGBoost works effectively to determine promotion success, for instance, other factors. Through the algorithm's application, the marketing analysis's effectiveness can be improved, making it easier to formulate data-backed decisions for digital promotions.

Research on data mining by comparing classification algorithms has been widely published. In this research, references from previous studies are needed so that they can find out the methods used.

An investigation by Saud Abd et al., Kurdi et al., and Givari et al. emphasise the importance of using digital marketing channels and classification algorithms due to their enhancement of consumer purchase decisions and data analysis and, at the same time, testing a wide array of machine learning algorithms including Naive Bayes, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest, and XGBoost, Saud Abd et al. found in their research work that the SVM algorithm provided a reasonable accuracy rate of 80%, showing its capability to predict digital marketing campaigns' results consistently and accurately (Saud Abd et al., 2024). Kurdi et al. examined the role of digital marketing channels, such as online advertising, social media, email marketing, and search engines, on purchasing decisions from consumers, and it was established that eWOM is critical in moderating the relationship between marketing channels and buying decisions (Kurdi et al., 2022). Givari et al., in their work, compare SVM, Random Forest, and XGBoost algorithms for credit application approval and showed that the XGBoost model attained the highest accuracy of 82% with a 70% recall and 92% precision, thereby demonstrating the high accuracy and reliability advantage of XGBoost across a breadth of classification problems. Thus, it would seem that where prediction occurs, irrespective of digital marketing or credit approval, the adoption of classification algorithms like SVM, Random Forest, and XGBoost can be rigorously predictive; hence, they make for quality improvements in data-informed decisions (Givari et al., 2022).

2. Research Methods

The research method used is the Cross-Industry Standard Process for Data Mining (CRISP-DM), one of the data mining process models or frameworks. This model was initially developed in 1996 by five companies: Integral Solutions Ltd (ISL), Teradata, Daimler AG, NCR Corporation, and OHRA. CRISP-DM has six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. (Amalia Yunia Rahmawati, 2020).

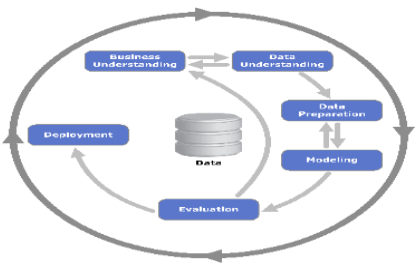


Figure 1. Stages of CRISP-DM

Each is explained as follows (Alamsyah, Yoga, et al., 2024) :

a. Business Understanding

This stage requires knowledge of business objects, building or obtaining data, and matching modelling objectives to business objectives to create the best model. The activities include clearly defining the overall goals and requirements, translating these goals, determining the restrictions in formulating the data mining problem, and preparing an initial strategy to achieve these goals.

b. Data Understanding

To examine the data to identify problems, as in data visualisation, a summary of the data can be helpful to confirm whether the data is distributed as expected or reveal unexpected deviations that need to be addressed in the data.

c. Data Preparation

This stage is often revisited when problems are found during model building. Activities carried out include selecting cases and parameters to be analysed (Select Data), transforming specific parameters (Transformation) and cleaning data so that the data is ready for the modelling stage (Cleaning).

d. Modelling

This stage uses statistical and machine learning methods to determine the data mining techniques, tools, and algorithms to be applied. Then, the data mining techniques and algorithms are applied to the data with the help of tools. If it is necessary to adjust the data to a particular data mining technique, you can return to the data preparation stage. Some modelling methods are classification, scoring, ranking, clustering, finding relations, and characterisation.

e. Evaluation

They are interpreting the data mining results generated in the modelling process at the previous stage. Evaluation is carried out on the model applied in the last stage with the aim that the model determined can match the objectives achieved in the first stage.

f. Deployment

The deployment stage is the most valued stage of the CRISP-DM process. Planning for deployment begins during business understanding and should incorporate not only how to generate model values but also how to convert decision scores and incorporate decisions in operational systems.

2.1 Naïve Bayes

The naive Bayes algorithm is one of the most effective and efficient inductive learning algorithms in machine learning and data mining (Syarli & Muin, 2016). Naive Bayes is an algorithm that can classify a particular variable using probability and statistical methods.

$$P(Ci \mid X) = \frac{P(X \mid Ci)P(Ci)}{P(X)} \quad (1)$$

Description :
X: Criteria for a case based on input
Ci: The solution class of the i-th pattern, where i is the number of class labels
P(Ci | X): Probability of class label Ci with input criteria X
P(X | C): Probability of input criteria label X with class label Ci
P(Ci): Probability of class label Ci

2.2 K-Nearest Neighbors

The K-nearest neighbour's method is a non-parametric method that can be used for classification based on k-nearest neighbours and regression. The K-Nearest Neighbor algorithm is a method to classify the object that is closest to the object based on the learning data (Harun et al., 2020).

$$Dist(X,Y) = \sqrt{\sum_{i=1}^{D}(X_i - Y_i)^2} \quad (2)$$

Description :
Dist(X,Y): Distance between objects
X_i: Sample data
Y_i: Test data
D: Data dimension
i: Variable data

2.3 Support Vector Machine

Support Vector Machine (SVM) works based on the Structural Risk Minimization (SRM) principle, which aims to find the best hyperplane to separate two classes in the input space. The level of accuracy in the model that the SVM switching process will generate is highly dependent on the kernel function and parameters used (Monika & Furqon, 2018).

The data in a dataset is given the variable xi, while the class in the dataset is given the variable yi. The SVM method divides the dataset into two classes. The first class, separated by the hyperplane, is 1, while the other is -1 (Monika & Furqon, 2018).

$$Xi * W + b \geq 1 \; untuk \; yi = 1 \quad (3)$$

$$Xi * W + b \geq -1 \; untuk \; yi = 1 \quad (4)$$

Description :
Xi: i-th data
W: perpendicular SVM weight value
b: bias value
Yi: i-th data class

2.4 Random Forest

Random Forest is a supervised classification method. As the name implies, this method forms a forest of several trees (Alamsyah et al., 2023). The more trees there are in a forest, the more the power or predictive ability of the forest increases. In other words, the more trees there are, the more accurate the predictions tend to be (Polamuri et al., 2019).

$$Regression:\ T(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(X) \quad (5)$$

$$Classification:\ C_{rf}^{B}(x) = majority\ vote\ \{C_b(X)\}_1^B \quad (6)$$

2.5 XGBoost

XGBoost is used for supervised learning problems where it uses training data with multiple features $\mathcal{X}i$ to predict the target variable $\mathcal{Y}i$ Before studying the tree specifically. Specifically by reviewing the essential elements of supervised learning. Model and model parameters in supervised learning usually refer to the mathematical structure in which the prediction $\mathcal{Y}i$ is made from the input $\mathcal{X}i$. A typical example is a linear model, where the prediction is given as $\mathcal{Y}\,i$ = $\sum j\ \theta j\ \mathcal{X}ij$, a linear combination of input feature weights. Predicted values can have different interpretations depending on the task, i.e. regression or classification. Parameters are the unspecified parts that we need to learn from the data. In a linear regression problem, the parameter is the coefficient θ. Usually, θ is used to denote the parameter Chen & Guestrin (Shafila, 2020).

2.6 Model Evaluation

Confusion matrix is a tool for predictive analysis in machine learning to check the performance of classification-based models (Ting, 2017). The structure of the confusion matrix is represented through rows and columns, where rows are the actual classes of the instances and columns are the predicted classes. Confusion matrix is represented as a 2 x 2 matrix with four terms, namely 'true positive' (TP), 'true negative' (TN), 'false positive' (FP) and 'false negative' (FN) (Hasnain et al., 2020).

The 'true negative' (TN) value is the amount of harmful data correctly detected, while 'false positive' (FP) is harmful data detected as positive. Meanwhile, 'true positive' (TP) is positive data that is detected correctly. 'false negative' (FN) is the opposite of 'true positive', so it is positive data but detected as harmful data (Alfi, 2019)

Table 1. Confusion Matrix

Class		True Class: Positive	True Class: Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

(Source: Zohreh Karimi, 2021)

1. True Positive (TP): actual data in the positive class, and the model predicts positive.
2. True Negative (TN): actual data in the negative class, and the model predicts negative.
3. False Positive (FP): actual data in the negative or neutral class, but the model predicts positive.
4. False Negative (FN): actual data in the positive or neutral class, but the model predicts negative.

3. Results and Discussion

3.1 Business Understanding

This research aims to provide an in-depth understanding of the impact of classification on sales results through promotional media such as television, radio and social media in a dataset related to marketing promotions. In addition, this research aims to identify and determine the most effective algorithm for classifying sales results, hoping to provide more accurate and targeted performance results for strategic decision-making in the marketing domain.

The purpose of Data Mining from this research is to explore practical and previously unknown information from the 'Marketing Promotion TV vs Radio vs Social Media' dataset, which then looks for performance results and assesses and compares the performance of previously determined classification algorithms such as Naive Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest and XGBoost.

3.2 Data Understanding

a. Data Collection

The dataset used for this research is 'Marketing Promotion TV vs Radio vs Social Media', accessed from the Kaggle platform. This dataset presents information related to marketing promotions through television, radio, and social media, along with attributes that include sales results.

b. Data Description

Table 2. Attributes in the Dataset

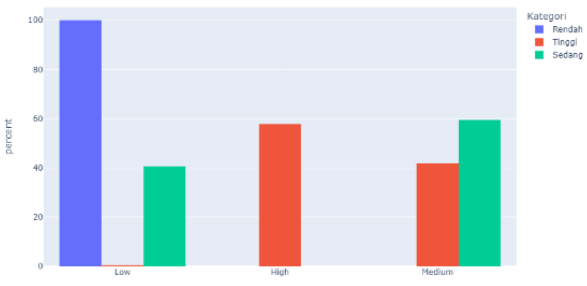
No	Attributes	Description
1	TV	Indicates categories on television media, with values such as 'Low', 'Medium', and 'High'
2	Radio	Expresses the value of spending on radio media in numerical form
3	Social Media	Expresses the value of the expenditure on social media in numerical form
4	Influencer	Indicates the category of influencer type used, with values such as 'Micro', 'Mega', 'Nano', and 'Macro'
5	Sales	States the value of sales generated

(Source: https://www.kaggle.com/datasets/yakhyojon/marketing-promotion)

In Table 2, one attribute is added based on the value called label, where this label is the category or level of the sales value to be classified.

c. Data Exploration

In the context of your sales data, you can explore the distribution of numerical variables such as 'TV, 'Radio', 'Social Media', and 'Sales' using graphs or histograms.



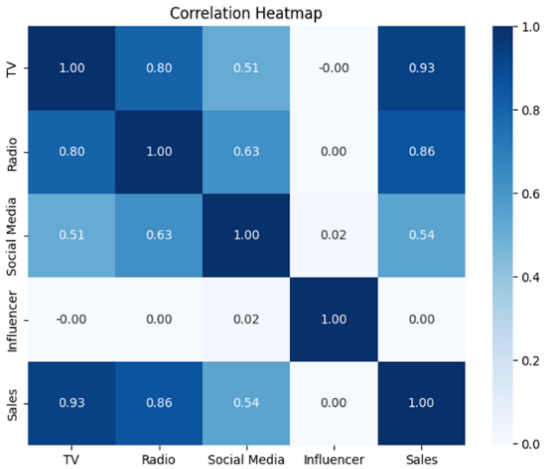
(Source: Research Process, 2024)

Figure 2. TV Histogram

In Figure 2, the context of TV histogram visualisation, there are three main categories: 'Low', 'High' and 'Medium', which represent the level of 'Label'. It can be seen that the 'Low' category labelled as low reaches a high of 100%, the 'High' category labelled as high reaches 57%, and the 'Medium' category labelled as medium reaches 59%. Shows that the dataset has relatively low 'TV' values. This indicates that most 'TV' values in the dataset are close to low.

d. Data Quality Verification

Data quality verification aims to ensure that the data used in the analysis is of good quality, free from errors, and reliable. It can be seen in the relationship between variables and other variables, one of which uses a Correlation heat map.



(Source: Research Process, 2024)

Figure 3. Correlation Heatmap

The correlation in Figure 3 between 'TV' and 'Sales' is 0.93. A high correlation value close to 1 indicates a strong positive linear relationship between TV and Sales. This means that the higher the TV, the higher the Sales.

The correlation between 'Radio' and 'Sales' shows a correlation value of 0.86, and the correlation between 'Social Media' and 'Sales' shows a correlation value of 0.54. In contrast, the correlation between 'Influencers' and 'Sales' shows a correlation value of 0.00. The low correlation value indicates no significant linear relationship between influencers and sales.

3.3 Data Preparation

This stage is used to refine the data included in the modelling process.

a. Selecting Data

The marketing promotion data obtained as raw data can be reviewed in the following table :

Table 3. Marketing Promotion Dataset

TV	Radio	Social Media	Influencer	Sales	Label
Low	3.518070	2.293790	Micro	55.261284	Low
Low	7.756876	2.572287	Mega	67.574904	Low
High	20.348988	1.227180	Micro	272.250108	High
Medium	20.108487	2.728374	Mega	195.102176	High
High	31.653200	7.776978	Nano	273.960377	High
…	…	…	…	…	…
Medium	14.656633	3.817980	Micro	191.521266	High
High	28.110171	7.358169	Mega	297.626731	High
Medium	11.401084	5.818697	Nano	145.416851	Medium
Medium	21.119991	5.703028	Macro	209.326830	High
Low	13.221237	3.660566	Micro	135.773151	Medium

(Source: https://www.kaggle.com/datasets/yakhyojon/marketing-promotion)

Table 4 shows the selection of attributes and the reduction of one attribute based on the previous data quality verification results, namely the attributes 'Influencer' and 'Sales'.

Table 4. Table Selection

TV	Radio	Social Media	Label
Low	3.518070	2.293790	Low
Low	7.756876	2.572287	Low
High	20.348988	1.227180	High
Medium	20.108487	2.728374	High
High	31.653200	7.776978	High
…	…	…	…
Medium	14.656633	3.817980	High
High	28.110171	7.358169	High
Medium	11.401084	5.818697	Medium
Medium	21.119991	5.703028	High
Low	13.221237	3.660566	Medium

(Source: Research Process, 2024)

b. Independent and Dependent Variables

Independent variables are variables that are assumed to have an influence on or cause changes in other variables. The independent variables in this analysis refer to 'TV', 'Radio', and 'Social Media'. While the dependent variable is the variable that is assumed to receive the influence or effect of the independent variable, the dependent variable in this analysis is 'Label'.

c. Label Encoding

Label encoding is a coding technique used to convert categorical values in a variable into numerical values. In this process, the dependent variable 'Label' is label encoding with conditioning, which aims to transform the category value of 'Label' into a numerical representation, namely 'High' is mapped to value 2, 'Medium' is mapped to value one and 'Low' is mapped to value 0.

d. Normalisation

Normalisation transforms the values in a dataset into a specific range or scale, usually 0 to 1 or -1 to 1. The results of normalisation using the Min-Max Scaling method on the features 'TV', 'Radio', and 'Social Media'. Each row represents one data point in the dataset, while each column represents the value of each feature after normalisation.

Attribute 'TV': Each value has been normalised to be between 0 and 1. For example, a value of 0 in the 'TV' feature indicates that the data has the lowest spending, while a value of 1 indicates spending—these are the other attributes.

3.4 Modelling

In this phase, the data mining process uses the Naïve Bayes, Random Forest, K-Nearest Neighbors, Support Vector Machine and XGBoost algorithm models. Data processing tools, namely Python with Google collabs, support this process. Before modelling, the dataset will be divided into training data by 70% and testing data by 30%, where the testing data consists of 400 records and testing 172 records.

a. Naïve Bayes Classification Algorithm

The stage of reading the Naïve Bayes algorithm model and conducting training data, where the results represent the Gaussian model object that has been initialised.

b. SVM Classificainitialisedthm

SVC(kernel='rbf', C=10, gamma=1), the output is a representation of a Support Vector Machine (SVM) model object that has been initialised with a Radial Basis Function (RBF) kernel used to handle non-linear data. 'C=10' refers to the balance between misclassification in the training data, which aims to minimise misclassification in the training data and Model Complexity. Meanwhile, 'gamma=1' creates an SVM model that adapts and is sensitive to small details in the data.

c. KNN Classification Algorithm

The output of KNeighborsClassifier (n_neighbors=13) shows that the K-Nearest Neighbors (KNN) model has been successfully

created with a particular configuration, where 'KNeighborsClassifier' is the class of the KNN model used for classification and n_neighbors=13 is the best K.

d. Random Forest Classification Algorithm

RandomForestClassifier (max_depth = 10, n_estimators = 150, random_state = 42) indicates a successful Random Forest Classifier model creation. This Random Forest Classifier is the class of Random Forest models used for classification tasks.

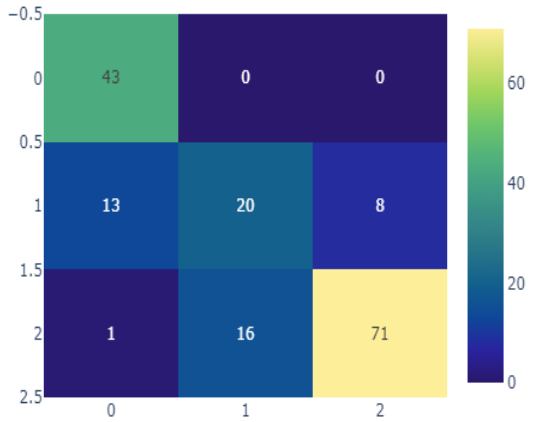
e. XGBoost Classification Algorithm

XGBoost output results This information provides an overview of how the XGBoost model is configured and can be used to make predictions on new data or evaluate test data.

3.5 Evaluation

The test results carried out at the modelling stage will obtain a model, which will then be evaluated to measure the performance of each algorithm. In general, the performance of the classification is described using a confusion matrix.

a. Confusion Matrix of Naïve Bayes Algorithm



(Source: Research Process, 2024)

Figure 4. Confusion Matrix Naïve Bayes

The results are obtained in the figure:4 Class 'Low' (0), TP(True Positive) 43 predicted data records labelled 'Low' and correct. Class 'Medium' (1), TP(True Positive) 20 predicted data records labelled 'Low' and correct. Class 'High' (2) True Positive (TP): The model correctly predicted 71 data records as High. The numbers from the matrix can be used to see the model's performance, such as accuracy, precision, recall and f1-score.

Table 5. Classification Report Naive Bayes

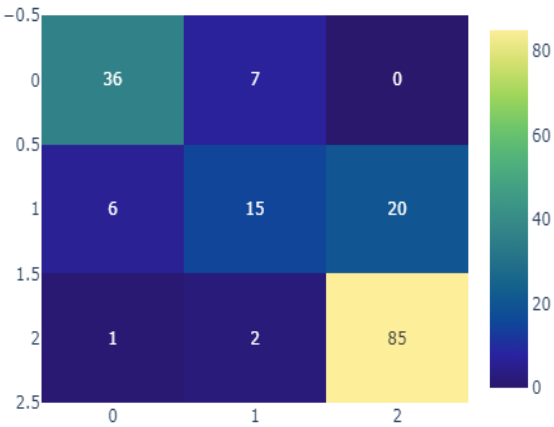
	precision	recall	F1-score	support
0	0.75	1.00	0.85	43
1	0.56	0.49	0.52	41
2	0.90	0.81	0.85	88
accuracy			0.78	172
macro avg	0.74	0.76	0.74	172

(Source: Research Process, 2024)

As seen in Table 5, For class 0, the model shows a precision level of 75%, which means that of all predictions categorised as class 0, 75% are classified as actual class 0. Recall for class 0 is 100%, indicating that the model can identify all instances of class 0.

Class 1 is seen to be inferior, with 56% precision and 49% recall. Of the class 1 predictions, only 56% predicted class 1, and the model could only identify 49% of the actual instances of class 1. Class 2 showed good performance with 90% precision and 81% recall. The F1 score, as a balance between precision and recall, reached 85% for classes 0 and 2, while class 1 had an F1 score of 52%. The total accuracy of the model reached 78%. Thus, the overall performance evaluation of the model showed a reasonably high success rate.

b. Confusion Matrix of KNN Algorithm



(Source: Research Process, 2024)

Figure 5. Confusion Matrix KNN

The image's in Figure 5 result is Class 'Low' (0), True Positive (TP): The model correctly classified 36 data as Low. Class 'Medium' (1), True Positive (TP): The model correctly classified 15 data as a medium. Class 'High' (2), True Positive (TP): The model correctly classified 85 data as High.

Table 6. Classification Report KNN

	precision	recall	F1-score	support
0	0.83	0.81	0.82	43
1	0.63	0.41	0.50	41
2	0.83	0.97	0.89	88
accuracy			0.80	172
macro avg	0.76	0.73	0.74	172

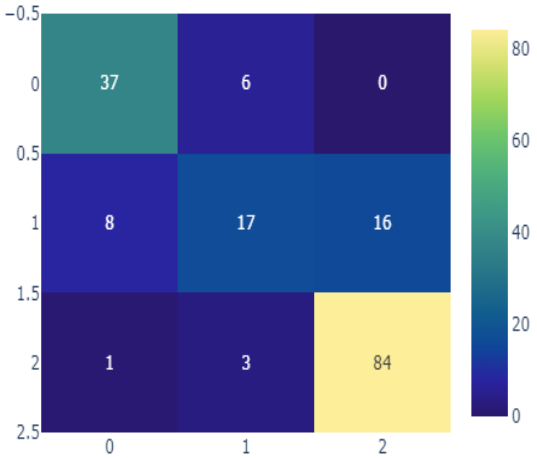
(Source: Research Process, 2024)

In Table 6, class 0, for which the model has a high precision (83%), meaning all predictions categorised as class 0, 83% of categorised correct class 0 and a recall of about 81%, indicating the model's ability to detect most Class 0 instances. The F1 score of about 82% reflects the balance between precision and recall for this class.

Class 1, where the Precision for Class 1 is 63%, indicates the model's ability to identify instances labelled as Class 1 with moderate accuracy. The recall is about 41%, and the F1 score is about 50%, reflecting a balance between lower precision and recall compared to Class 0 and 2.

Class 2, where the model is very good at identifying instances of Class 2, with a precision of 83%, recall of 97% and an F1 score of around 89%. This shows that the model gives exact predictions for Class 2.

c. Confusion Maxtrix of SVM Algorithm



(Source: Research Process, 2024)

Figure 6. Confusion Matrix SVM

The figure 6, result is Class 'Low' (0), True Positive (TP): The model correctly classified 37 data as Low. Class 'Medium' (1), True Positive (TP): The model correctly classified 17 data as medium. Class 'High' (2), True Positive (TP): The model correctly classified 84 data as High.

Table 7. Classification Report SVM

	precision	recall	F1-score	support
0	0.80	0.86	0.82	43
1	0.65	0.41	0.51	41
2	0.84	0.95	0.89	88
accuracy			0.80	172
macro avg	0.77	0.74	0.74	172

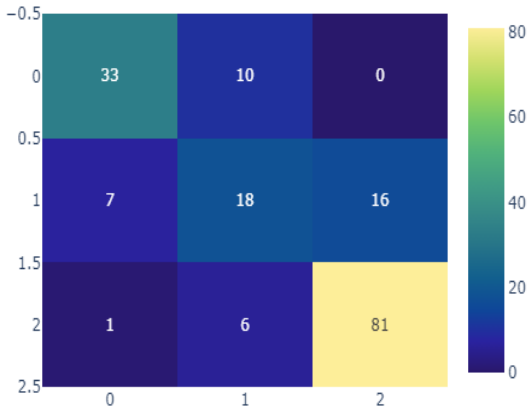
(Source: Research Process, 2024)

In Table 7, class 0 has a precision of about 80% in identifying Class 0 and a recall of about 86%, indicating the model's ability to detect most of Class 0. The F1-Score of about 82% reflects the balance between precision and recall for this class.

Class 1, Precision for Class 1 is 65%, indicating the model's ability to identify with the Class 1 label with moderate accuracy. The recall is about 41%, indicating that the model may overlook many Class 1 instances. The F1 score is about 51%, reflecting the balance between lower precision and recall compared to Class 0 and 2.

Class 2, the model is very good at identifying from Class 2, with a precision of 84%, recall of 95%, and F1-Score of about 89%. This indicates that the model tends to provide precise predictions for Class 2. The model's overall accuracy is 80%, reflecting the percentage of correct predictions from the entire dataset.

d. Confusion Matrix of Random Forest Algorithm



(Source: Research Process, 2024)

Figure 7. Confusion Matrix Random Forest

As can be seen in the figure 7, the result is Class 'Low' (0), True Positive (TP): The model correctly classified 33 data as Low. Class 'Medium' (1), True Positive (TP): The model correctly classified 18 data as a medium. Class 'High' (2), True Positive (TP): The model correctly classified 81 data as High.

Table 8. Classification Report Random Forest

	precision	recall	F1-score	support
0	0.80	0.77	0.79	43
1	0.53	0.44	0.48	41
2	0.84	0.92	0.88	88
accuracy			0.77	172
macro avg	0.72	0.71	0.71	172

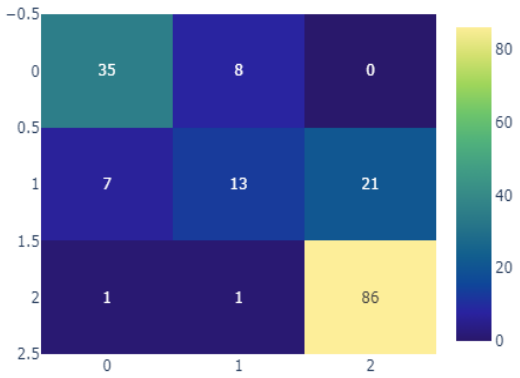
(Source: Research Process, 2024)

In Table 8, class 0, the model has a precision of about 80% in identifying those that belong to Class 0 and a recall of about 77%, indicating the model's ability to detect most of Class 0. The F1-Score of about 79% reflects the balance between precision and recall for this class.

Class 1, Precision for Class 1 is 53%, indicating the model's ability to identify with the Class 1 label with moderate accuracy. The recall was about 44%, indicating that the model probably detected most of Class 1. The F1 score was about 48%, reflecting the lower precision and recall balance compared to Class 0 and 2.

Class 2, the model is very good at identifying instances of Class 2, with a precision of 84%, recall of 92%, and F1-Score of about 88%. This shows that the model tends to give correct predictions for Class 2. Likewise, the model's overall accuracy is 77%, indicating the percentage of accurate predictions from the entire dataset.

e. Confusion Matrix of XGBoost Algorithm



(Source: Research Process, 2024)

Figure 8. Confusion Matrix XGBoost

As can be seen in the figure, the result is Class 'Low' (0), True Positive (TP): The model correctly classified 35 data as Low. Class 'Medium' (1), True Positive (TP): The model correctly classified 13 data as medium. Class 'High' (2), True Positive (TP): The model correctly classified 86 data as High. The numbers from the matrix can be used to see the model's performance, including accuracy, precision, recall, and f1 score.

Table 9. Classification Report XGBoost

	precision	recall	F1-score	support
0	0.81	0.81	0.81	43
1	0.59	0.32	0.41	41
2	0.80	0.98	0.88	88
accuracy			0.78	172
macro avg	0.74	0.70	0.70	172

(Source: Research Process, 2024)

In Table 9, class 0, the model has a precision of about 81% in identifying those that belong to Class 0 and a recall of about 81%, indicating the model's ability to detect most instances of Class 0. The F1-Score of about 81% reflects the balance between precision and recall for this class.

Class 1, Precision for Class 1 is 59%, indicating the model's ability to identify with the Class 1 label with moderate accuracy. The recall is about 32%, and F1 Score is about 41%, reflecting a balance between lower precision and recall compared to Class 0 and 2.

Class 2: The model is very good at identifying from Class 2, with a precision of 80%, recall of 98%, and F1-Score of 88%. This indicates that the model tends to provide exact predictions for Class 2, and the overall accuracy of the model is 78%, reflecting the percentage of correct predictions from the entire dataset.

3.6 Deployment

This stage will make a report to complete the entire activity; in this research, the K-Nearest Neighbors, Naïve Bayes, Support Vector Machine, Random Forest and XGBoost algorithms in the classification comparison for marketing promotion. The analysis results obtained the accuracy value of each algorithm, where the KNN (K-Nearest Neighbors) algorithm successfully classified the data with an accuracy rate of 79.65%. This means that about 79.65% of the predictions are correct according to the actual class, Naive Bayes achieved an accuracy of 77.91%, SVM (Support Vector Machine) has an accuracy rate of 80.23%, Random Forest has an accuracy rate of 76.74% and XGBoost also has an accuracy of 77.91%.

In Table 10, the SVM model performed best with the highest accuracy and balanced precision, recall and F1-Score values. The Random Forest model showed low accuracy and slightly lower precision performance than the other models. We

will determine how well each model can classify data about these marketing promotions based on their accuracy levels.

Table 10. Algorithm Comparison Results

Model	Accuracy %	Precision %	Recall %	F1 %
KNN	79.65	78.06	79.65	78.04
Naïve Bayes	77.91	78.08	77.91	77.39
SVM	80.23	78.67	80.23	78.60
Random Forest	76.74	75.47	76.74	75.89
XGBoost	77.91	75.56	77.91	75.31

(Source: Research Process, 2024)

4. Conclusion

Based on the results of the research conducted, it can be concluded that by applying the Cross Industry Standard for Data Mining (CRISP-DM) methodology, the research begins with an understanding of the business, followed by the preparation and knowledge of data, which will later be used in the modelling stage. Several classification algorithms such as Naive Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest and XGBoost are used in the modelling stage. Based on the analysis, the Random Forest Algorithm provides classification accuracy results of 0.76 (76%). The SVM algorithm provides classification accuracy results of 0.80 (80%). KNN algorithm provides classification accuracy results of 0.79 (79%). The Naive Bayes algorithm provides classification accuracy results of 0.77 (77%), and the XGBoost algorithm provides 0.77 (77%) classification accuracy results. Based on the results of testing classification models that have been carried out using datasets related to marketing promotions obtained from Kaggle under the name 'Marketing Promotion TV vs Radio vs Social Media', the Support Vector Machine (SVM) shows the highest performance among all models with precision, recall, F1-score and overall accuracy model values of 0.80 (80%). This indicates that SVM can provide accurate and consistent predictions for each class.

Reference

Al-Ababneh, H. A. (2024). Information technologies and their impact on electronic marketing. *E3S Web of Conferences*, *474*, 02010. https://doi.org/10.1051/e3sconf/202447402010

Alamsyah, N., Budiman, B., Parama Yoga, T., & Rakhman Alamsyah, R. Y. (2024). A stacking ensemble model with SMOTE for improved imbalanced classification on credit data. *Telkomnika (Telecommunication Computing Electronics and Control)*, *22*(3), 657. https://doi.org/10.12928/telkomnika.v22i3.25921

Alamsyah, N., Budiman, Danestiara, V. R., Akbar, I., & Setiana, E. (2023). Optimising Computational Efficiency in Feature Selection for Machine Learning Models: A Study Crime Detection Based on Criminal Data. *2023 Eighth International Conference on Informatics and Computing (ICIC)*, 1–6. https://doi.org/10.1109/ICIC60109.2023.10382057

Alamsyah, N., Yoga, T. P., & Budiman, B. (2024). Improving Traffic Density Prediction Using Lstm With Parametric Relu (Prelu) Activation. *JITK (Jurnal Ilmu Pengetahuan Dan Teknologi Komputer)*, *9*(2), 154–160. https://doi.org/10.33480/jitk.v9i2.5046

Alfi, M. (2019). *Analisis Sentimen Berdasarkan Knowledge Pattern dan Learning Vector Quantization*. 9–18.

Amalia Yunia Rahmawati. (2020). *Pengenalan Data Mining* (Issue July).

APJII. (2023). *Survei APJII Pengguna Internet di Indonesia Tembus 215 Juta*. https://apjii.or.id/berita/d/survei-apjii-pengguna-internet-di-indonesia-tembus-215-juta-orang

Ardilla, Y., Manuhutu, A., Ahmad, N., Hasbi, I., Manuhutu, M. A., Ridwan, M., Wardhani, A. K., & others. (2021). *Data Mining Dan Aplikasinya*. Penerbit Widina.

Budiman, B. (2021). Perbandingan Algoritma Klasifikasi Data Mining untuk Penelusuran Minat Calon Mahasiswa Baru. *Nuansa Informatika*, *15*(2), 37–52. https://doi.org/10.25134/nuansa.v15i2.4162

Budiman, B., Nursyanti, R., Alamsyah, R. Y. R., & Akbar, I. (2020). *Data Mining Implementation Using Naïve Bayes Algorithm and Decision Tree J48 In Determining Concentration Selection*. *1*(3).

Givari, M. R., Sulaeman, M. R., & Umaidah, Y. (2022). Perbandingan Algoritma SVM, Random Forest Dan XGBoost Untuk Penentuan Persetujuan Pengajuan Kredit. *Nuansa Informatika*, *16*(1), 141–149. https://doi.org/10.25134/nuansa.v16i1.5406

Harun, R., Pelangi, K. C., & Lasena, Y. (2020). Penerapan Data Mining untuk Menentukan Potensi Hujan Harian dengan Menggunakan Algoritma K Nearest Neighbor (KNN). *Jurnal Manajemen Informatika Dan Sistem Informasi*, *3*(1), 8–15.

Hasnain, M., Pasha, M. F., Ghani, I., Imran, M., Alzahrani, M. Y., & Budiarto, R. (2020). Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking. *IEEE Access*, *8*, 90847–90861. https://doi.org/10.1109/ACCESS.2020.2994222

Indrayana, R., & Solikhin, M. (2020). Analisis Sentimen Pada Media Sosial Twitter. *Seminar Nasional Pendidikanipa Danmatematika*, *12*(1), 79–86.

Kurdi, B., Alshurideh, M., Akour, I., Alzoubi, H., Obeidat, B., & AlHamad, A. (2022). The role of digital marketing channels on consumer buying decisions through eWOM in the Jordanian markets. *International Journal of Data and Network Science*, *6*(4), 1175–1186.

Malik, R., Kusumadinata, A. A., & Hasbiyah, D. (2023). Keragaman Media Sosial Instagram Sebagai Media Promosi. *Karimah Tauhid*, *2*(1), 26–35.

Monika, I. P., & Furqon, M. T. (2018). Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(10), 3165–3166.

Polamuri, S. R., Srinivas, K., & Mohan, A. K. (2019). Stock market prices prediction using random forest and extra tree regression. *Int. J. Recent Technol. Eng*, *8*(1), 1224–1228.

Puspitarini, D. S., & Nuraeni, R. (2019). Pemanfaatan media sosial sebagai media promosi. *Jurnal Common*, *3*(1), 71–80.

Saud Abd, N., Salim Atiyah, O., Taher Ahmed, M., & Bakhit, A. (2024). Digital Marketing Data Classification by Using Machine Learning Algorithms. *Iraqi Journal for Electrical and Electronic Engineering*, *20*(1), 245–256. https://doi.org/10.37917/ijeee.20.1.23

Shafila, G. A. (2020). Implementasi Metode Extreme Gradient Boosting (Xgboost) untuk Klasifikasi pada Data Bioinformatika (Studi Kasus: Penyakit Ebola , GSE 122692). *Dspace.Uii.Ac.Id*, 1–77.

Syarli, & Muin, A. A. (2016). Metode Naive Bayes Untuk Prediksi Kelulusan. *Jurnal Ilmiah Ilmu Komputer*, *2*(1), 22–26.

Ting, K. M. (2017). Confusion Matrix. In *Encyclopedia of Machine Learning and Data Mining* (Issue October, pp. 260–260). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_50

Wawolumaya, E. D., Tampi, D. L., Rogahang, J. J., & ... (2022). Analisis Strategi Pemasaran Dalam Upaya Meningkatkan Volume Penjualan Pada Rose Collection. *… : Journal Of …*, 134–141.

Yoedtadi, M. G. (2019). Tv Sosial: Televisi dan Media Sosial. *Konferensi Nasional Komunikasi Humanis*, *November 2019*, 2.

Yoesoep, R. E. (2022). Manajemen Pemasaran. In *Eureka Media Aksara*.

Yunial, agus heri. (2020). Analisa Perbandingan Algoritma Klasifikasi Support Vector Machine, Decession Tree Dan Naive Bayes. *Prosiding Seminar Informatika Dan Sistem Informasi*, *5*(2), 138–156.