Vol. 8, No. 4, August 2026

ISI Journal of Information Systems and Informatics

Published By Asosiasi Doktor Sistem Informasi Indonesia

Comparative Classification of Promotional Sources in Higher Education Admissions Using K-Nearest Neighbor and Naive Bayes

Elly Yanuarti[1], Sujono[2], Djoko Soetarno[3], Rahmat Sulaiman[4]

[1]Information System Department, [2]Information Technology Education Department, [4]Informatics Department, ISB Atma Luhur, Pangkalpinang, Indonesia

[3]Informatics Department, Binus University, Jakarta, Indonesia

Received:
February 21, 2026
Revised:
June 3, 2026
Accepted:
August 5, 2026
Published:
August 30, 2026

Corresponding Author:
Author Name*:
Elly Yanuarti
Email*:
elly@atmaluhur.ac.id

DOI:
10.63158/journalisi.v8i4.1689

Abstract. This study evaluates and compares the performance of Naive Bayes and K-Nearest Neighbor (KNN) algorithms for classifying promotional-source categories in higher education admissions based on ten years of historical admission records. The objective is to analyze the capability of machine learning approaches in identifying patterns of applicant acquisition sources and to provide insights for institutional data-driven evaluation. After data preprocessing and quality filtering, 2,618 out of 4,901 records with complete target-variable information were retained and classified into seven promotional-source categories. Both algorithms were assessed using 5-fold cross-validation with multiple evaluation measures, including accuracy, macro-averaged recall, and comparison against a majority-class baseline to address the effect of severe class imbalance. Experimental results indicate that KNN achieved substantially higher overall accuracy (82.24%) than Naive Bayes (43.74%). However, neither model surpassed the majority-class baseline, demonstrating that accuracy alone can lead to misleading conclusions in highly imbalanced classification problems. In contrast, Naive Bayes obtained higher macro recall (36.32% compared with 18.14% for KNN), indicating a broader capability in recognizing minority promotional-source categories. The findings emphasize the importance of imbalance-aware evaluation and provide analytical insights into historical promotional-source distributions to support strategic admission planning and future institutional decision-making.

Keywords: promotional-source classification, higher education admission, K-Nearest Neighbor, Naive Bayes, class imbalance.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

1. INTRODUCTION

The rapid growth of information technology and digital transformation has significantly changed the competitive landscape in higher education, particularly in the new student admission process. Universities are no longer competing solely based on academic quality, but also on their ability to reach prospective students through multiple promotional channels, including printed brochures, referrals from relatives, mobile and digital platforms, and mass media [1]. The expansion of digital media has encouraged higher education institutions to utilize various online platforms as promotional tools for student recruitment [2]. However, the increasing number of promotional channels is often not accompanied by a systematic way of classifying, from historical admission records, which channel a given applicant actually reports as their source of information [3]. In this context, this study distinguishes two analytically distinct aspects of promotional-source analysis: (a) media frequency, how often a channel is cited by applicants as their source of information; (b) applicant volume, the number of historical records associated with each channel. This study is scoped specifically to the classification of promotional-source categories from historical admission records; conversion outcomes, promotional cost, and return-on-investment, which would be required to evaluate promotional effectiveness, are outside the scope of the data used here.

Most universities still evaluate their historical promotional-source records through manual recapitulation rather than systematic, data-driven methods [3]. Advances in artificial intelligence and machine learning have opened opportunities for analytical decision-support approaches that use historical institutional data to assist classification and interpretation in higher education [5], [6]. Several studies illustrate this potential: Saleem et al. used ensemble machine learning to predict students' e-learning performance with high accuracy [6]; Latif et al. showed that multi-level machine learning approaches can improve decision-support effectiveness in higher education [5]; and Nauman et al. demonstrated that machine learning can support faster, more consistent institutional decision-making than manual approaches [4].

Among classification methods, Naive Bayes and K-Nearest Neighbor (KNN) are widely used in educational data classification for their computational efficiency, interpretability, and effectiveness on categorical data [7], [2], [8]. Prior studies have applied both

algorithms to student-related classification tasks with mixed results: Riadi et al. found that Naive Bayes and KNN offer different precision–recall trade-offs when predicting graduation timeliness, rather than one method being uniformly superior [9]; Zayed et al. applied KNN in a higher-education recommendation system [2] and Zhou et al. showed that a hybrid Naive Bayes–KNN model can outperform either algorithm alone [8]. These two algorithms were selected for the present study because they are computationally tractable on a large, largely categorical admission dataset; they represent contrasting classification philosophies probabilistic independence versus instance-based proximity enabling a diagnostic comparison under severe class imbalance; and they remain interpretable enough for adoption by non-technical admissions staff, unlike more complex ensemble methods such as Random Forest or gradient boosting [10].

Research on admission promotion using machine learning has grown in recent years but remains largely single-algorithm and descriptive: Toro and Lestari compared classification algorithms to identify promotion locations [11]; Malik and Kamayani found digital promotional media significantly associated with enrollment using KNN [12]; and Barus applied Naive Bayes to classify new students [7]. A broader marketing and recruitment literature likewise underscores the need for systematic, evidence-based promotional analysis: universities increasingly rely on social media and multi-channel strategies for student acquisition [13]; strategic channel prioritization is seen as critical amid enrollment decline [14]; AI-powered digital marketing measurably shapes enrollment decisions [15]; integrated analytics dashboards can improve admission decision-making [16]; and branding research argues for channel-level evidence over reach-based metrics alone [17]. Methodologically, Bayesian attribution modeling illustrates the broader challenge of tracing outcomes to marketing channels that motivates this study's narrower, classification-based scope [18], while historical admission data has been shown to meaningfully forecast enrollment outcomes, supporting its use here [19].

Despite this body of work, most prior studies predict academic performance or overall enrollment status rather than classifying promotional-source categories from historical records [11], [1], and admission-focused studies that do exist typically apply a single algorithm without accounting for severe class imbalance across categories [7], [12]. The broader imbalance literature shows that raw accuracy can be misleading when one class dominates a dataset; resampling techniques such as SMOTE address this at the cost of

synthetic-data assumptions [20], while macro-averaged and balanced metrics offer a resampling-free alternative [21], [22], directly relevant here, where the majority promotional-source category accounts for over 80% of usable records. This motivates two gaps this study addresses : (a) the absence of a comparative Naive Bayes–KNN evaluation on multi-category, severely imbalanced promotional-source data using macro-averaged and majority-baseline-adjusted metrics rather than raw accuracy alone; and (b) the limited use of historical, institution-level admission records rather than survey or aggregated data for promotional-source classification [13], [14].

Based on these conditions, there is a clear gap regarding the limited use of machine learning-based classification, evaluated with imbalance-aware metrics, for historical promotional-source records in higher education admissions. The primary objective of this study is therefore to classify promotional-source categories from historical student-admission data using KNN and Naive Bayes, and to compare their classification performance using accuracy alongside macro-averaged and majority-baseline-adjusted metrics appropriate for severely imbalanced multi-class data. This study addresses the following research question: Which promotional-source categories are most accurately classified from historical admission records, and how do KNN and Naive Bayes compare once class imbalance is explicitly accounted for through macro-averaged metrics and a majority-class baseline? This study does not attempt to measure promotional effectiveness, conversion, or return-on-investment, and its findings should not be interpreted as a basis for promotional budget allocation without further analysis incorporating conversion and cost data.

This study applies a comparative classification approach using K-Nearest Neighbor and Naive Bayes to historical promotional-source records, allowing examination of how differently structured classifiers behave on the same severely imbalanced, largely categorical dataset [23]. The resulting classification patterns are intended to support, rather than substitute for, further institutional analysis; the study does not implement a functioning recommendation system, and any operational application would require additional development beyond the present scope. Its main contributions are as follows: (a) a classification-focused operationalisation of promotional-source analysis that separates historical media frequency from applicant volume, without conflating either with unmeasured conversion or effectiveness outcomes; (b) a comparative classification

experiment applying KNN (k = 5, MixedEuclideanDistance) and Naive Bayes to 4,901 admission records spanning ten years, of which 2,618 records with a complete promotional-source label were evaluated using accuracy, macro-averaged metrics, a majority-class baseline, and per-class confusion-matrix metrics; (c) a set of analytical implications, developed in RapidMiner Studio, that organize classification results for interpretation by university admissions staff; and (d) an explicit comparison of both algorithms against a majority-class baseline under extreme class imbalance, offering evidence-based guidance for selecting classifiers on similarly imbalanced educational-analytics tasks [16], [24].

2. METHODS

2.1. Research Design

This study employed a comparative classification design to analyze historical promotional-source categories in new student admission data. It focused on classifying the source of information through which prospective students became aware of the institution, as recorded in the admission database. Accordingly, the dependent variable in this study was Information Source, while the remaining admission-related attributes were treated as predictor variables. The analytical objective of the study was to compare the performance of Naive Bayes and K-Nearest Neighbor (KNN) in classifying promotional-source categories from historical admission data and to interpret the results for decision-support purposes. The overall research process undertaken in this study is illustrated in Figure 1, which outlines the eight sequential steps carried out using RapidMiner Studio, from raw data collection through to analytical interpretation.

As illustrated in Figure 1, the research process consisted of eight sequential steps. In Step 1 (Data Collection), ten years of historical new-student admission records were compiled from the institution's admission database, including applicant demographic information, school background, and the self-reported promotional source through which each applicant learned about the institution. In Step 2 (Data Import and Preparation), the compiled dataset was imported into RapidMiner Studio using the Read Excel operator, after which the Set Role operator designated Information Source as the target (label) attribute, and the Replace Missing Values operator was applied to the remaining predictor attributes. In Step 3 (Label-Completeness Filtering), records without a valid Information

Source value were identified; because RapidMiner's supervised-learning operators require a complete label for every training and test instance, these 2,283 records were excluded automatically during model training and testing, yielding a final classification dataset of 2,618 records distributed across seven promotional-source categories, as detailed in Table 1. In Step 4 (Model Training), the Multiply operator duplicated this dataset into two identical streams, which were used to train Naive Bayes with Laplace correction enabled (Step 4a) and K-Nearest Neighbor configured with k = 5, weighted vote, and MixedEuclideanDistance (Step 4b) in parallel. In Step 5 (Cross-Validation), both models were evaluated using 5-fold cross-validation with RapidMiner's automatic sampling type, which applies stratified sampling for a nominal target variable such as Information Source [25]. In Step 6 (Model Evaluation), performance was measured through accuracy, the confusion matrix, and macro- and weighted-averaged precision, recall, and F1-score, following standard multiclass evaluation practice [21], [22]. In Step 7 (Baseline Comparison), both models' accuracy was benchmarked against a majority-class baseline of 83.00%, obtained by always predicting the dominant Brochures category, to determine whether either classifier provided a genuine improvement over trivial prediction. Finally, in Step 8 (Analytical Interpretation), the resulting classification patterns, confusion matrices, and comparative metrics were interpreted to characterize historical promotional-source patterns and their implications for university admissions analytics, as presented in Section 3.

2.2. Dataset and Variables

The dataset used in this study was obtained from the new student admission records of over a period of 10 years. The dataset contains admission-related information collected during the student registration process and stored in the institutional admission database. After the preprocessing stage, the final dataset used in the experiment consisted of 4,901 usable records. Further inspection of the dataset revealed that 2,283 of these 4,901 records contained missing values on the target variable, Information Source, and could therefore not be used in the supervised classification process. These records were excluded prior to model training, resulting in a final classification dataset of 2,618 records, which was subsequently used for both Naive Bayes and KNN training, evaluation, and the confusion matrices. Table 1 presents the distribution of the seven promotional-source categories across the 2,618-record classification dataset used in this study.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

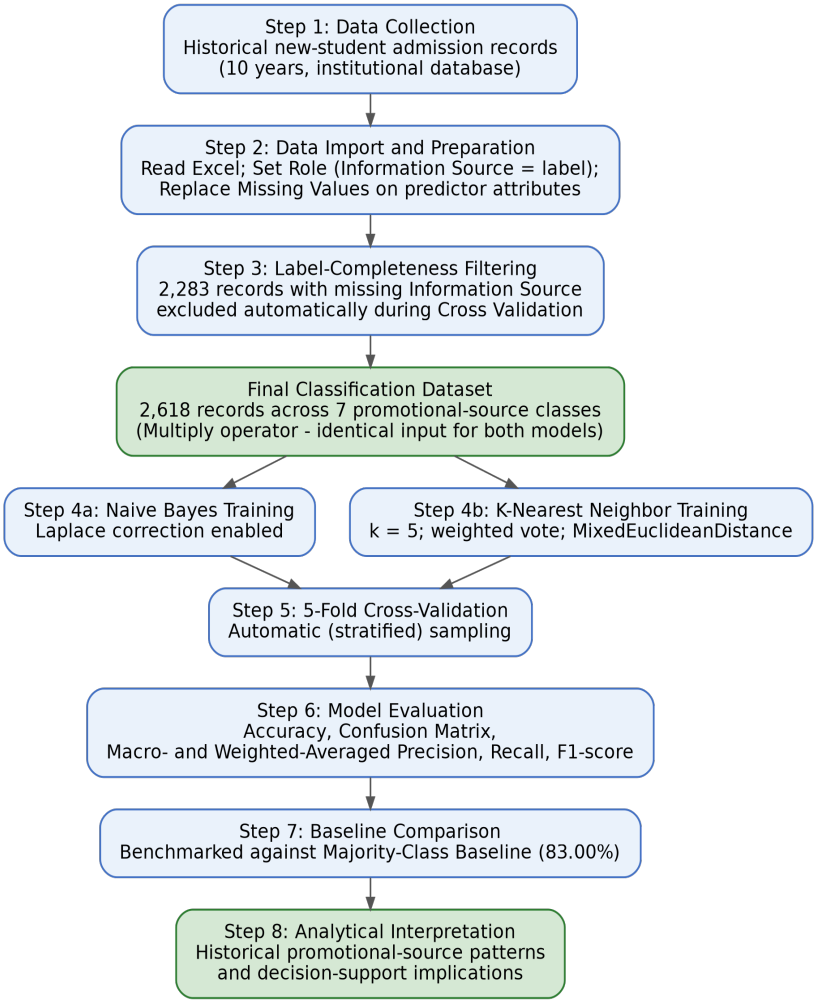

Figure 1. Research design and experimental workflow

Table 1. Class Distribution of the Classification Dataset

Promotional Source	Number of Records (Support)	Percentage
Brochures	2,173	83.00%
Relatives	171	6.53%
Android	141	5.39%
Others	112	4.28%
Internet	17	0.65%
Radio	2	0.08%
Newspaper	2	0.08%
Total	2,618	100.00%

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

The target variable in this study was Information Source, which represents the promotional-source category through which prospective students became aware of the institution. The categories in this variable include Brochures, Relatives, Android, Others, Internet, Radio, and Newspaper. Nine predictor variables from the admission database were used as input attributes for classification, excluding the applicant's name, which was retained only as a record identifier and excluded from the model. Table 2 lists these nine predictor variables together with their data type and a brief description, including the proportion of missing values within the 2,618-record classification dataset.

Table 2. Predictor Variables Used in the Classification Model

Promotional Source	Data Type	Description
Admission Year	Numerical	Year in which the prospective student registered for admission (8 distinct years in the classification dataset).
Province	Nominal	Applicant's province of origin (23 categories; complete for all 2,618 records).
Subdistrict	Nominal	Applicant's subdistrict of origin (91 categories; 41.1% missing in the classification dataset).
School Name	Nominal	Name of the applicant's secondary school (294 distinct categories; complete for all 2,618 records, but high-cardinality).
School Major	Nominal	Academic major/track completed at secondary school (146 categories; 33.2% missing)
Scholarship Type	Nominal	Type of scholarship associated with the applicant's registration (5 categories; 86.2% missing).
Registration Wave	Numerical	Admission registration batch/wave number (6 distinct values; complete for all 2,618 records).
Study Program	Nominal	Study program applied to: SI, TI, or BD (3 categories; complete for all 2,618 records).
Re-Registration Status	Nominal	Whether the applicant completed re-registration (Yes/No; complete for all 2,618 records).

Two predictor variables exhibited substantial missingness within the classification dataset: Scholarship Type (86.2% missing) and Subdistrict (41.1% missing). In addition, School Name, although fully populated, is a high cardinality nominal attribute with 294 distinct categories across 2,618 records close to one category per nine records. These characteristics are discussed as limitations of the predictor data, as they plausibly constrain the amount of usable signal available to both classifiers.

2.3. Data Processing

Data preprocessing was conducted in RapidMiner before the classification stage. Based on the final workflow (Read Excel → Set Role → Replace Missing Values → Multiply → Cross Validation), the preprocessing stage consisted of data import, role assignment, missing-value handling, followed by duplication of the cleaned dataset for comparative model training. First, the admission dataset was imported from an Excel file using the Read Excel operator. Second, the variable information source was assigned as the label attribute using the Set Role operator, indicating that it served as the dependent variable to be predicted by both classification algorithms. Third, missing values in the predictor attributes were handled using the Replace Missing Values operator, with the attribute filter type set to all. The workflow does not contain a separate Filter Examples step for the label attribute; instead, the 2,283 records lacking a valid value for Information Source were automatically excluded during the training and testing stages inside the Cross Validation operator, since RapidMiner's supervised learning operators require a complete label value for every training and test instance. This resulted in the working classification dataset of 2,618 records.

Regarding categorical handling, the predictor attributes were retained in their original nominal or numerical form rather than being manually re-encoded. No Nominal-to-Numerical or comparable encoding operator is present in the workflow; instead, the KNN operator was configured with mixed measures and MixedEuclideanDistance, which computes similarity for nominal attributes (via a matching-based nominal distance) and numerical attributes (via Euclidean distance) jointly within a single similarity function. This configuration allowed both classification algorithms to process the mixed-type admission attributes without requiring a separate manual encoding step.

Two predictor attributes retained substantial missingness even after this stage — Scholarship Type (86.2%) and Subdistrict (41.1%) because the Replace Missing Values operator imputes a default value rather than removing incomplete records; the practical effect of this imputation on model performance is discussed as a limitation. After missing-value treatment, the cleaned dataset was passed through the Multiply operator to generate identical input streams for the two classification models. This ensured that Naive Bayes and KNN were trained and evaluated using the same preprocessed dataset.

2.4. Classification Models

1) Naive Bayes

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem and assumes conditional independence among predictor variables given the target class [7]. The algorithm estimates the probability that a data instance belongs to a particular class using patterns learned from previously observed data. The Naive Bayes operator in RapidMiner estimates class-conditional probabilities using frequency counts for nominal attributes and a distributional (Gaussian) assumption for numerical attributes [7]. In the final configuration used in this study, the Laplace correction option was enabled, ensuring that categories with zero observed frequency in the training folds were assigned a small non-zero probability rather than a probability of exactly zero.

2) K-Nearest Neighbor (KNN)

The second classification model used in this study was K-Nearest Neighbor (KNN). KNN is an instance-based classification algorithm that assigns the class of a new observation based on the labels of its nearest neighbors in the feature space [8], [9]. In the final experiment, KNN was configured with k = 5. The weighted vote option was enabled so that the contribution of neighboring observations was weighted during the classification process. The measure type was set to mixed measures, and the similarity calculation used MixedEuclideanDistance.

2.5. Experimental Environment and Evaluation Procedure

All preprocessing, model training, and performance evaluation procedures were conducted using RapidMiner. After preprocessing, the same cleaned, label-complete dataset of 2,618 records was used as input for both Naive Bayes and KNN to ensure a fair comparison between the two classification models.

The performance of each model was evaluated using 5-fold cross-validation, a widely used resampling procedure for estimating classifier accuracy and comparing models [25]. Based on the final RapidMiner configuration, the number of folds was set to 5, the sampling type was set to automatic, use local random seed was disabled, leave-one-out and split-on-batch-attribute were disabled, and parallel execution was enabled. For a nominal target variable such as Information Source, RapidMiner's automatic sampling type applies stratified sampling by default, allocating examples to folds so that the proportion of each class is preserved across folds as closely as possible. Given the severe imbalance of the classification dataset (Table 1), this default stratification is appropriate for the results reported here. consistent with prior findings that stratified cross-validation improves accuracy estimation over non-stratified alternatives [25]. Because the classification dataset is highly imbalanced (see Table 1), a majority-class baseline was included alongside the two trained models, following standard practice for evaluating classifiers on imbalanced data [21], [22]. This baseline predicts the majority class (Brochures) for every record and achieves 83.00% accuracy without using any input attributes; it is used throughout Section 3 to contextualize the accuracy achieved by Naive Bayes and KNN.

The evaluation metrics selected in RapidMiner were accuracy, weighted mean precision, weighted mean recall, and confusion matrix. Because the class distribution of Information Source was highly imbalanced, model performance was not interpreted solely from overall accuracy. The confusion matrix was used to examine class-level prediction behavior, especially for minority categories such as Internet, Radio, and Newspaper, which had substantially fewer records than the dominant Brochures category. To provide a more balanced interpretation of model performance, Section 3 also reports macro precision, macro recall, macro F1-score, and support-weighted metrics, recalculated directly from the corrected confusion matrices using standard definitions.

3. RESULTS AND DISCUSSION

3.1. Experimental Setup and Evaluation Scheme

This study evaluated the classification performance of Naive Bayes and K-Nearest Neighbor (KNN) on the 2,618-record classification dataset, using 5-fold cross-validation. Because this dataset was partitioned into 5 non-overlapping folds, each record was used

as a test instance exactly once across the five cross-validation iterations. Consequently, the confusion matrices reported in this section aggregate predictions across all five folds and contain exactly 2,618 cases in total. In the confusion matrices presented below (Tables 4 and 6), each row represents the class predicted by the model, and each column represents the true (actual) class of the record, consistent with RapidMiner's standard Performance output. The right-most column reports class precision (correct predictions divided by all predictions made for that class), and the bottom row reports class recall (correct predictions divided by all true instances of that class). Because the promotional-source dataset was highly imbalanced across categories (Table 1), evaluation in this study combines four complementary perspectives: (a) overall accuracy, (b) the majority-class baseline, (c) macro-averaged precision/recall/F1-score, which weight all seven classes equally regardless of size, and (d) support-weighted precision/recall/F1-score, which weight classes by their sample size.

3.2. Naive Bayes Testing Results

Based on data processing using the Naive Bayes algorithm, the following performance results were obtained as shown in Table 3.

Table 3. Naive Bayes Testing

Parameter	Value
Accuracy	43.74%
Weighted Mean Precision	20.84%
Weighted Mean Recall	32.02%

The first experiment was conducted using the Naive Bayes algorithm to classify historical promotional-source categories in new student admissions. As shown in Table 3, based on the 5-fold cross-validation results generated by RapidMiner, the Naive Bayes model achieved an accuracy of 43.74%. Compared with the 83.00% majority-class baseline established in Section 2.5, this accuracy is substantially lower. This does not indicate that Naive Bayes performed poorly in an absolute sense; rather, it reflects that the model actively distributed its predictions across multiple promotional-source categories instead of defaulting to the dominant Brochures class. This behavior is examined in more depth through the confusion matrix and the macro-averaged metrics in Sections 3.3 and 3.6.

3.3. Confusion Matrix of Naive Bayes Classification Result

Table 4 shows that the strongest classification performance occurred in the Brochures class, with 942 correctly classified records out of 1,099 predicted as Brochures (85.71% precision). However, the Brochures class also received substantial numbers of misclassified records that truly belonged to other categories: 61 from Relatives, 44 from Android, 43 from Others, 8 from Internet, 1 from Radio.

Table 4. Confusion Matrix of Naive Bayes Classification Results

Predicted\ True	Brochures	Relatives	Android	Others	Internet	Radio	Newspaper	Class Precision
Brochures	942	61	44	43	8	1	0	85.71%
Relatives	525	79	1	21	0	0	1	12.60%
Android	155	7	82	8	5	0	1	31.78%
Others	407	23	2	37	0	0	0	7.89%
Internet	77	0	11	2	4	0	0	4.26%
Radio	34	0	0	0	0	1	0	2.86%
Newspaper	33	1	1	1	0	0	0	0.00%
Class Recall	43.35%	46.20%	58.16%	33.04%	23.53%	50.00%	0.00%	

Read column-wise, the confusion matrix also shows that Naive Bayes correctly recovered a meaningful share of minority-class instances: 46.20% recall for Relatives, 58.16% for Android, 33.04% for Others, and 23.53% for Internet. figures that are considerably higher than those achieved by KNN (Section 3.5). This indicates that, despite its lower overall accuracy, Naive Bayes was comparatively more capable of identifying non-dominant promotional-source categories.

3.4. KNN Testing Results

The second experiment was conducted using the K-Nearest Neighbor (KNN) algorithm to classify promotional-source categories based on the same dataset and evaluation procedure. As shown in Table 5, according to the RapidMiner output, the KNN model achieved an accuracy of 82.24%. At first glance, this accuracy appears strong. However, compared against the 83.00% majority-class baseline (Section 2.5), KNN's accuracy is in fact 0.76 percentage points below the baseline. In other words, a classifier that always predicts "Brochures" — without using any information about the applicant would outperform the KNN model on overall accuracy.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 5. KNN Testing Result

Parameter	Value
Accuracy	82.24%
Weighted Precission	27.83%
Weighted Recall	18.18%

3.5. Confusion Matrix of KNN Classification

Table 6 shows that KNN performed strongly only on the Brochures class, correctly classifying 2,117 of 2,173 true Brochures records (97.42% recall). For every other class, KNN's recall was extremely low: 1.75% for Relatives, 6.38% for Android, 21.43% for Others, and 0.00% for Internet, Radio, and Newspaper, meaning KNN failed to correctly identify a single true instance of these three minority categories.

Table 6. Confusion Matrix of KNN Classification Result

Predicted\ True	Brochures	Relatives	Android	Others	Internet	Radio	Newspaper	Class Precision
Brochures	2,117	164	131	88	17	2	2	83.97%
Relatives	15	3	0	0	0	0	0	16.67%
Android	5	1	9	0	0	0	0	60.00%
Others	35	3	1	24	0	0	0	38.10%
Internet	1	0	0	0	0	0	0	0.00%
Radio	0	0	0	0	0	0	0	0.00%
Newspaper	0	0	0	0	0	0	0	0.00%
Class Recall	97.42%	1.75%	6.38%	21.43%	0.00%	0.00%	0.00%	

This pattern shows that KNN's high accuracy is driven almost entirely by the dominance of the Brochures class in the dataset, rather than by a genuine ability to distinguish among promotional-source categories. This is consistent with the near-baseline accuracy reported in Section 3.4 and motivates the use of macro-averaged and balanced-accuracy metrics, which are less sensitive to class imbalance, in Section 3.6.

3.6. Algorithm Performance Comparison

Table 7 places the accuracy of both classifiers in context. Neither Naive Bayes nor KNN exceeded the 83.00% majority-class baseline. Naive Bayes fell 39.26 percentage points

below the baseline, reflecting its more distributed predictions across categories. KNN, despite reporting the numerically higher accuracy of the two models, fell 0.76 percentage points below the baseline indicating that its apparent strength in Table 5 (82.24% accuracy) does not represent a genuine improvement over trivially predicting the majority class.

Table 7. Accuracy Compared Against the Majority-Class Baseline

Method	Accuracy	Gap vs. Majority-Class Baseline (83.00%)
Majority-Class Baseline	83.00%	-
Naive Bayes	43.74%	-39.26 pp
KNN	82.24%	-0.76 pp

3.7. Class-Balanced Evaluation Using Macro and Weighted Metrics

Because accuracy alone is misleading for imbalanced datasets, this section reports macro-averaged and support-weighted precision, recall, and F1-score, recalculated directly from the confusion matrices in Tables 4 and 6 using standard definitions. Macro recall is equivalent to balanced accuracy, since both compute the unweighted mean of per-class recall.

Table 8. Macro-Averaged and Weighted Performance Comparison

Algorithm	Macro Precision	Macro Recall (Balanced Accuracy)	Macro F1-score	Weighted Precision	Weighted Recall	Weighted F1-Score
Naive Bayes	20.73%	36.32%	20.55%	74.05%	43.74%	51.90%
KNN	28.39%	18.14%	18.91%	75.65%	82.24%	76.87%

Table 9. Per-Class Precision, Recall, F1-score and Support – Naive Bayes

Class	Support	Precision	Recall	F1-score
Brochures	2173	85.71%	43.35%	57.58%
Relatives	171	12.6%	46.2%	19.8%
Android	141	31.78%	58.16%	41.1%
Others	112	7.89%	33.04%	12.74%
Internet	17	4.26%	23.53%	7.21%
Radio	2	2.86%	50%	5.41%
Newspaper	2	0%	0%	0%

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Table 10. Per-Class Precision, Recall, F1-score and Support – KNN

Class	Support	Precision	Recall	F1-score
Brochures	2173	83.97%	97.42%	90.2%
Relatives	171	16.67%	1.75%	3.17%
Android	141	60%	6.38%	11.54%
Others	112	38.1%	21.43%	27.43%
Internet	17	0%	0%	0%
Radio	2	0%	0%	0%
Newspaper	2	0%	0%	0%

Table 8 shows that Naive Bayes achieved a higher macro recall / balanced accuracy (36.32%) than KNN (18.14%), confirming that it was more effective, on average, at identifying instances across all seven promotional-source categories rather than concentrating on the dominant class. KNN, in contrast, achieved a higher macro precision (28.39% vs. 20.73%) and much higher weighted recall (82.24%, identical to its accuracy), which reflects its tendency to predict the dominant Brochures class almost exclusively rather than a genuine advantage in class-level discrimination as confirmed by its 0% recall on three of the seven classes in Table 10.

Tables 9 and 10 also show that the Radio and Newspaper categories each have a true support of only 2 records in the 2,618-record classification dataset. With so few instances, per-class precision, recall, and F1-score for these two categories are highly unstable a single correct or incorrect prediction changes the reported recall by 50 percentage points. These per-class figures should therefore be interpreted as indicative rather than statistically reliable, and this limitation is noted explicitly in Section 3.9.

3.8. Discussion

The comparative results directly address this study's research question. Across the seven promotional-source categories, Brochures was the only category reliably classified by either algorithm, while Relatives, Android, Others, Internet, Radio, and Newspaper were identified with substantially lower, and for KNN often negligible, reliability. This pattern is consistent with the broader class-imbalance literature, which shows that classifiers trained on skewed distributions tend to default toward the majority class unless explicitly corrected [21], [22]. It also extends prior comparative work on Naive Bayes and KNN in

educational classification: whereas Riadi et al. [9] and Zhou et al. [8] reported that the relative strength of Naive Bayes and KNN depends on the prediction task and data characteristics, the present findings show that, under severe class imbalance, this trade-off manifests specifically as a precision-recall asymmetry, in which KNN attains higher precision on the dominant class at the cost of recall on minority classes, while Naive Bayes trades overall accuracy for more balanced class-level recall.

These findings also carry a methodological implication beyond this dataset. Because KNN's raw accuracy (82.24%) appeared strong in isolation but did not exceed a majority-class baseline (83.00%), evaluating classifiers on imbalanced educational data using accuracy alone risks overstating model quality, echoing similar cautions raised in the general imbalanced-learning literature [21], [22]. Reporting a majority-class baseline alongside macro-averaged metrics, as done in this study, provides a more defensible basis for algorithm selection in similarly imbalanced admissions-analytics tasks.

From a substantive standpoint, the results suggest that the dominance of Brochures in the historical record is better explained by data-collection patterns, such as how consistently admissions staff recorded this response during registration, than by superior reach or persuasiveness relative to other channels. This interpretation is consistent with the scope defined in Section 1, which explicitly separates historical media frequency from promotional effectiveness, and it cautions against using either model's output as direct evidence for reallocating promotional resources without complementary conversion data.

3.9. Decision Support Implications

Table 11 summarizes the decision-support implications of these findings for each promotional-source category. Overall, the findings indicate that the proposed analytical interpretation should not rely on accuracy or on either classifier in isolation. Because neither Naive Bayes nor KNN exceeded the majority-class baseline in overall accuracy, and because KNN in particular showed minimal ability to identify non-dominant categories, decision support conclusions drawn from this study are limited to describing historical patterns in the data rather than recommending specific promotional-budget actions. Any application of these results to promotional planning should be treated as a starting point for further, conversion-aware analysis rather than a direct decision rule.

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

Rather than serving as a direct measure of promotional effectiveness, the analytical framework presented in this study functions as a support tool that helps university management identify dominant historical promotional-source patterns and compare algorithmic classification behavior using historical admission data.

Table 11. Decision Support Implications Based on Historical Promotional-Source Pattern

Promotional Source	Interpretation from Historical Data	Decision Support Implication
Brochures	Dominant category in historical admission records (83.00% of usable data); neither classifier meaningfully outperformed a trivial majority-class prediction for this category.	Interpret brochure dominance as a reflection of historical data frequency, not evidence of promotional effectiveness. Any resource-allocation decision would require separate conversion or cost data, which this study does not include.
Relatives	Secondary source, frequently misclassified into the dominant class by KNN (1.75% recall); identified more often by Naive Bayes (46.20% recall).	Improve data capture for this channel; do not rely on KNN outputs alone to characterize this category.
Android	Emerging digital source with limited representation (141 records); moderately identified by Naive Bayes (58.16% recall) but weakly by KNN (6.38% recall).	Monitor digital engagement trends; treat classification results as exploratory rather than conclusive given the algorithm-dependent variation.
Others	Heterogeneous category with unstable classification pattern across both models.	Reassess and refine category definitions for future data collection before drawing further conclusions.
Internet	Low representation (17 records); moderate recall under Naive Bayes (23.53%) but zero recall under KNN.	Treat as a monitored channel; supplement with descriptive trend analysis rather than classifier output.
Radio	Extremely limited representation (2 records); recall estimates from either model are not statistically reliable at this sample size.	Do not use these classification results as a basis for reducing or reallocating this channel; collect more data before evaluation.

Newspaper	Extremely limited representation (2 records); recall estimates from either model are not statistically reliable at this sample size.	Same as Radio: reassess only after additional data collection; current results are inconclusive.

3.10. Limitations

This study has several limitations that should be considered when interpreting the results. First, the classification dataset is highly imbalanced: the Brochures category alone accounts for 83.00% of the 2,618 usable records, while two of the seven categories (Radio and Newspaper) are represented by only 2 records each, making per-class metrics for these categories statistically unstable. Second, the confusion-matrix orientation (predicted rows vs. true columns) required explicit verification against RapidMiner's raw Performance output; readers extending this line of research should verify metric orientation directly against software output rather than reformatted tables. Third, the dataset originates from a single higher-education institution over a 10-year period and does not include conversion, cost, or return-on-investment data for any promotional channel, which limits the extent to which classification results can inform budget-allocation decisions.

Fourth, the quality of the predictor attributes themselves plausibly limited both models' ability to discriminate among promotional-source categories. Of the ten predictor variables used (Table 2), three retained substantial missingness even after imputation: Scholarship Type (86.2% missing in the classification dataset) and Subdistrict (41.1% missing). Because these missing values were imputed with a default value via the Replace Missing Values operator rather than removed, a large share of the values used for these two attributes during training and testing were not genuine observations, which may have diluted their predictive contribution. In addition, School Name — although fully populated — is a high-cardinality nominal attribute with 294 distinct categories across the 2,618-record dataset, close to one distinct category per nine records; such high cardinality can cause a nominal-distance measure (as used by KNN's MixedEuclideanDistance) to behave close to a per-record identifier for many observations, adding noise rather than generalizable signal to the similarity calculation. These predictor-level limitations offer a concrete, data-grounded explanation for the low macro-averaged performance of both classifiers, complementing the class-imbalance

explanation above, and should be addressed in future work through predictor selection, cardinality reduction (e.g., grouping schools by region or type), and more principled missing-data handling than uniform default imputation.

4. CONCLUSION

This study developed an analytical interpretation of historical promotional-source categories in higher education admissions using Naive Bayes and K-Nearest Neighbor (KNN), evaluated on a final classification dataset of 2,618 records drawn from ten years of admission data, in which the Brochures category accounted for 83.00% of usable records. KNN achieved a higher overall accuracy (82.24%) than Naive Bayes (43.74%), but neither exceeded the 83.00% majority-class baseline established in this study, indicating that KNN's apparent strength was driven primarily by the dominance of the Brochures class rather than genuine discriminative ability, while Naive Bayes achieved a meaningfully higher macro recall / balanced accuracy (36.32% vs. 18.14%) and a comparable macro F1-score (20.55% vs. 18.91%), reflecting a broader, though still limited, ability to identify minority promotional-source categories; overall, neither algorithm can be described as unconditionally superior. This dominance of Brochures reflects historical recording frequency rather than evidence of promotional effectiveness, and the findings do not support claims regarding promotional effectiveness, conversion efficiency, or budget allocation, particularly given the severe class imbalance, single-institution scope, and predictor-level limitations discussed in Section 3.9. Future research should incorporate conversion-based variables and promotional cost data, adopt longitudinal or temporal validation across admission cohorts, and improve imbalance handling and predictor quality through resampling strategies (e.g., SMOTE or class weighting) and cardinality reduction, always benchmarked against a majority-class baseline.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the university and all parties who provided support, assistance, and research data that enabled this study to be completed successfully. Appreciation is also extended to everyone who contributed valuable suggestions and input throughout the research preparation process.

REFERENCES

[1] P. Kotler and K. Keller, *Marketing Management*, 16th ed. Pearson, 2022.

[2] Y. Zayed, Y. Salman, and A. Hasasneh, "A Recommendation System for Selecting the Appropriate Undergraduate Program at Higher Education Institutions Using Graduate Student Data," *Appl. Sci.*, vol. 12, no. 24, 2022.

[3] Z. Munawar et al., *BIG DATA ANALYTICS: Konsep, Implementasi, dan Aplikasi Terkini*. Kaizen Media Publishing, 2023.

[4] M. Nauman, N. Akhtar, A. Alhudhaif, and A. Alothaim, "Guaranteeing Correctness of Machine Learning Based Decision Making at Higher Educational Institutions," *IEEE Access*, vol. 9, 2021.

[5] S. Latif, F. XianWen, and L. Wang, "Intelligent Decision Support System Approach for Predicting the Performance of Students Based on Three-Level Machine Learning Technique," *J. Intell. Syst.*, vol. 30, no. 1, 2021.

[6] F. Saleem, Z. Ullah, B. Fakieh, and F. Kateb, "Intelligent Decision Support System for Predicting Student's E-Learning Performance Using Ensemble Machine Learning," *Mathematics*, vol. 9, no. 17, 2021.

[7] S. P. Barus, "Implementation of Naive Bayes Classifier-Based Machine Learning to Predict and Classify New Students at Matana University," *J. Phys. Conf. Ser.*, vol. 1842, 2021.

[8] X. Zhou, L. Guo, R. Li, L. Liu, and J. Pan, "Intelligent Teaching Recommendation Model Based on Naive Bayes and Improved KNN," *Information*, vol. 16, no. 6, 2025.

[9] I. Riadi, R. Umar, and R. Anggara, "Comparative Analysis of Naive Bayes and K-NN Approaches to Predict Timely Graduation using Academic History," *Int. J. Comput. Digit. Syst.*, vol. 16, no. 1, pp. 1163–1174, 2024.

[10] T. G. Dietterich, "Ensemble methods in machine learning," in *Lecture Notes in Computer Science*, vol. 1857, pp. 1–15, 2000.

[11] R. Toro and S. Lestari, "Perbandingan Algoritma Klasifikasi Untuk Penentuan Lokasi Promosi Penerimaan Mahasiswa Baru Pada IIB Darmajaya Lampung," *Techno.com*, 2023.

[12] L. A. R. Malik and M. Kamayani, "Faktor-Faktor Yang Mempengaruhi Minat Calon Mahasiswa Baru Mendaftar Menggunakan Algoritma K-Nearest Neighbor," *Infotech J. Technol. Inf.*, vol. 9, no. 2, 2023.

[13] H.-A. Tran, H. Evanschitzky, S. Ludwig, B. Nguyen, D. Grewal, and A. M. Farrell, "How

ISI Journal of Information Systems and Informatics
Published By Asosiasi Doktor Sistem Informasi Indonesia

universities can use social media for student acquisition," *J. Acad. Mark. Sci.*, 2025.

[14] C. H. Phillips and S. J. Jones, "Strategic and tactical marketing strategies for regional public universities to address the enrollment cliff," *J. High. Educ.*, vol. 96, no. 4, pp. 653–677, 2024.

[15] A. S. AbuDaabes, H. M. Selim, and R. Hijazi, "From clicks to campus: How AI-powered digital marketing shapes student enrollment decisions," *Asia-Pac. J. Bus. Adm.*, 2025.

[16] K. C. Li, B. T. Wong, and M. Liu, "Development of a multi-model analytics system to enhance decision-making in student admission," *Interact. Technol. Smart Educ.*, vol. 22, no. 3, pp. 506–523, 2025.

[17] S. K. Pawar, "University branding in the age of social media: a systematic literature review and integrative framework," *J. Appl. Res. High. Educ.*, 2025.

[18] R. Sinha, D. Arbour, and A. M. Puli, "Bayesian Modeling of Marketing Attribution," arXiv:2205.15965, 2022.

[19] L. Yang, L. Feng, L. Zhang, et al., "Predicting freshmen enrollment based on machine learning," *J. Supercomput.*, vol. 77, pp. 11853–11865, 2021.

[20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[21] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[22] S. Li, L. Song, X. Wu, Z. Hu, Y.-M. Cheung, and X. Yao, "Multiclass imbalance classification based on data distribution and adaptive weights," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 10, pp. 5265–5279, 2024.

[23] N. Nurwati and Y. Santoso, "Analisis Perbandingan K-Nearest Neighbors dan Naive Bayes Untuk Rekomendasi Pilihan Program Studi Bagi Mahasiswa," *IDEALIS*, vol. 8, no. 1, 2025.

[24] M. Munawirah and A. O. Arisha, "Implementasi Naive Bayes untuk Klasifikasi Peminatan Program Studi pada Penerimaan Mahasiswa Baru," *Bull. Inf. Technol.*, vol. 6, no. 1, 2025.

[25] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proc. 14th Int. Joint Conf. Artificial Intelligence (IJCAI)*, Montreal, Canada, 1995, pp. 1137–1143.