A Proposed Model for Multi-Dimensional Data Quality

Mohamed Ibrahim Marie [1]
Rabab Yehia Oraby [2]
Ahmed Mohamed Abd-Elwahab [3]

Abstract

In recent years, business intelligence has emerged as a critical field leveraging data analysis to generate actionable insights for informed decision-making. This paper emphasizes the significance of data quality and the choice of the most appropriate model to enhance the accuracy of predictions, which contributes to improving marketing strategies and banking decision-making. The aim of this paper is to evaluate the ability of machine learning models to predict the outcomes of direct marketing campaigns for banks using accurate and unbiased data. The five models tested were K-Nearest Neighbour's (KNN), Random Forest, Decision Tree, Gradient Boosting, and Multi-Layer Perceptron (MLP). The data showed that the Gradient Boosting model performed better than others in marketing applications, with an accuracy of 91.45%. Random Forest showed similar performance with an accuracy of 91.24% despite the longer prediction time, (MLP) achieved 90.24%, (KNN) achieved 89.55%, and Decision Tree was the fastest, but its accuracy was somewhat lower at 88.58%.

Keywords: Data Quality, Data Quality Dimensions, Quality Metrics, Model Evaluation, Machine Learning.

1. Associate Professor of Information Systems Department, Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt.
2. Data Analyst and Officer Administrator, International Ranking Unit, Helwan University, Cairo, Egypt.
3. Assistant Professor of Information Systems, Information Systems Department, Faculty of Commerce and Business Administration, Helwan University, Cairo, Egypt.

نموذج مقترح لجودة البيانات متعدد الأبعاد

الملخص

في السنوات الأخيرة، برزت استخبارات الأعمال كمجال حاسم يعتمد على تحليل البيانات لتوليد رؤى قابلة للتطبيق لدعم اتخاذ القرارات المدروسة. تؤكد هذه الورقة البحثية على أهمية جودة البيانات واختيار النموذج الأمثل لتعزيز دقة التنبؤات، مما يسهم في تحسين استراتيجيات التسويق واتخاذ القرارات المصرفية. يهدف هذا البحث إلى تقييم قدرة نماذج تعلم الآلة على التنبؤ بنتائج حملات التسويق المباشر للبنوك باستخدام بيانات دقيقة وغير متحيزة. النماذج الخمسة التي تم اختبارها هيKNN، RF، DT، GB، MLP. أظهرت البيانات أن نموذج تعزيز التدرج (GB) كان أداؤه أفضل من النماذج الأخرى في تطبيقات التسويق، بدقة بلغت 91.45%. أظهر نموذج الغابة العشوائية أداءً مشابهًا بدقة 91.24% على الرغم من وقت التنبؤ الأطول، وحقق نموذج الشبكة العصبية متعددة الطبقات (MLP) دقة 90.24%، وحقق نموذج الجار الأقرب (KNN) دقة 89.55%، وكان نموذج شجرة القرار الأسرع، لكن دقته كانت أقل قليلاً عند 88.58%.أظهر نموذج الغابة العشوائية أداءً مشابهًا بدقة بلغت 91.24% على الرغم من وقت التنبؤ الأطول، بينما حقق نموذج الشبكة العصبية متعددة الطبقات(MLP) دقة بلغت 90.24%، ونموذج الجار الأقرب (KNN) دقة بلغت 89.55%، وكان نموذج شجرة القرار الأسرع، لكن دقته كانت أقل قليلاً عند 88.58%.

الكلمات المفتاحية: جودة البيانات، أبعاد جودة البيانات، مقاييس الجودة، تقييم النماذج، تعلم الآلة.

1. Introduction

In today's world, access to vast amounts of data can lead to more power and informed decisions. Unlike the end of the last century, where data was generated daily, today's data is everywhere, created by individuals, groups, companies, and even things that depend on the internet, especially for companies, public and private companies, and service providers [1] and [2].

Information quality is the relevance, precision, utility, contextual appropriateness, understandability, and timeliness of data. It aligns with user needs and is considered high quality. Information quality can be divided into subcategories and dimensions for better understanding [3], [4] and [5].

Data analysis is extremely important, especially for businesses who use these analytics to make informed decisions about their strategies, including recruiting, marketing, and branding. In general, this analysis can be used to predict unknown or what we call extrapolation, making the concept of big data even more important than the concept of artificial intelligence. We specifically mention machine learning; with its advantages like speed, automation, no initial costs, and labor savings, that increases the competitive advantage of companies [3] and [6].

Improving information quality is challenging due to data's complexity and unstructured nature. Domain specialists should be involved in refining and mistake detection, using automatic or semi-automatic approaches and data mining techniques [7].

In fact, one of the most important steps in data mining is considered to be the data preparation step, which is the process of ensuring the quality of information by changing the original data into a suitable format for the analysis process [8].

Measuring data quality and setting improvement targets is a vital activity. It is critical to know that depending on a single information quality metric is insufficient. Information quality has numerous aspects, hence any dataset should be examined against these dimensions to ensure a full information quality assessment [9].

The use of poor-quality data with missing and incorrect values can lead to inaccurate and unreasonable results, rendering the entire data collection and analysis process useless to users. Therefore, it is crucial to have an effective data preprocessing framework in order to deal with inaccurate and missing values. Data preprocessing and purification play a crucial role in ensuring correct, first-class information [10] and [11].

This paper is organized as follows: Section 2, examines the related work of data quality. In section 3, Data quality dimensions and measures are discussed. In addition, a model for evaluating Data quality is presented in section 4. In section 5, outlines the results. Finally, conclusion and directions for future work are reported in section 6.

2. Related Work

In recent years, data quality has garnered significant attention from experts and scholars, with much of the research focusing on developing models and techniques to enhance data accuracy. Previous studies in this field can be categorized into several key areas. These include investigations into the dimensions of information quality, such as completeness, duplication, correctness, and consistency, as well as the development of models and machine learning techniques aimed at improving data quality accuracy. The following sections provide an overview of these studies.

2.1 Completeness / Missing Values Dimension

It is important to address missing data before proceeding with any data processing steps. By addressing the challenges posed by missing data, it provides valuable insights for improving the accuracy of predictions and decision-making processes.

In [12], Palanivinayagam and Damaševičius, introduced a unique missing data imputation method that outperformed conventional approaches and offered a more efficient approach, reaching an accuracy of 94.89%. Using SVM regression and two-level classification for machine-learning models.

In [13], Widyananda, et al., showed improved data categorization efficiency with missing values. Even with 35% missing data, the C5.0 method showed better data categorization performance when paired with k-Nearest Neighbor Imputation. It achieved a high classification accuracy of 93.40% and a test accuracy of 92% as well as faster processing times.

2.2 Duplication Dimension

Duplicates negatively impact finances, sales, and business automation. Identifying and managing them is crucial. Various tools are introduced to identify and eliminate these duplicates. The duplication dimension addresses unwanted duplicates in specific fields or datasets using specific measures.

In [14], Zakeri-Nasrabadi, et al., conducted a systematic literature review and meta-analysis of code similarity assessment and rating methodologies and identified issues like restricted datasets, empirical evaluations, hybrid methodologies, and specific programming languages.

In [15], Opdenplatz, et al., proposed a service-based duplicate detection solution that is simple, competitive, and has been applied in an industrial context, addressing the issue of professional knowledge required for this task.

2.3 Correctness / Incorrect Dimension

The model's accuracy dimension measures the accuracy and correctness of information, crucial for software applications to produce accurate results. Errors can lead to incorrect decisions and actions, making accuracy a critical aspect.

In [1], Javid, et al., highlighted how crucial data preparation methods are to maintaining data quality. also highlights popular strategies, including data reduction and purification, as well as the limitations of applying hybrid models in real-world situations.

In [16], Mohseni, et al., provided a fresh approach to finding mistakes and anomalies in test data sets. By eliminating outlier data and utilizing ideas like standard deviation, mean value, and the Euclidean distance metric, this technique improves classification accuracy.

2.4 Consistency / Inconsistent Dimension

Data quality consistency is crucial for maintaining orderly data structures and formats across multiple datasets. This ensures uniformity in organization, naming conventions, and measurement units.

In [17], Nassreddine, et al., introduced a k-means clustering distance-based technique for anomaly detection, which is effective in identifying outlier data in real data.

In [18], Alzahrani, et al., created an energy-efficient data consistency protocol for Internet of Things systems that used machine learning to categorize nodes as event or continuous monitoring nodes. The method performed better in terms of data consistency than previously used methods.

2.5 Models and Frameworks for Data Quality Accuracy and Machine Learning Techniques

To deal with the multidimensional nature of data quality and its impact on the accuracy of the application used, several frameworks and models have been proposed that refer to the dimensions of data quality and machine learning techniques. They are briefly summarized in Table 1.

In [19], Huang, investigated how to forecast the effectiveness of telemarketing efforts in the banking industry using machine learning (ML) techniques. It is concluded that the most efficient algorithm is XGBoost (XGB), which has a high accuracy (90.14%), F1-score (90.85%). In comparison to XGBoost, gradient boosting (GB) also demonstrated good performance, obtaining a slightly higher accuracy (90.41%) but a lower F1 score (89.44%).

In [20], Kaisar, et al., proposed a solution for predicting telemarketing outcomes for CRM systems that integrates preprocessing, data collection, and machine learning models such as gradient boosting, AdaBoost, and random forest. The random forest model performs better than other models with accuracy, precision, recall, AUC, and F1-score of 98.6%, 93.4%, 98.1%, 0.975, and 0.954, respectively.

In [21], Huynh, et al., applied a variety of machine learning techniques, such as logistic regression, K-nearest neighbor, supported linear vector machines, and steep gradient boosting with response encoding, to successfully solve problems with imbalanced datasets. KNN is the best prediction model, achieving an accuracy of 91.07% and an AUC score of 0.9324.

In [22], Lam, et al., presented the classification and regression tree (CART) algorithm, which effectively improves telemarketing strategies by achieving high values of 92% for accuracy (AS), 99% for recall, 86% for precision, and 92% for F1 score, thus improving telemarketing strategies.

In [23], Vongchalerm, utilized decision trees and logistic regression, two machine learning algorithms, to examine a marketing strategy. The logistic regression model outperformed the others, according to the data, with an AUC of 0.934, accuracy of 0.870, and dependability of 0.869.

In [24], Tékouabou, et al., presented a class membership-based classifier designed to handle heterogeneous data well. They used nominal variables in the decision-making process, avoiding the common pitfalls of arbitrary encoding, which can lead to overfitting or poor performance. The proposed CMB classifier had an accuracy of 97.3% and an AUC of 95.9%.

In [25], Masturoh, et al., revealed how the Multilayer Perceptron (MLP) algorithm with resampling techniques, which had an accuracy rate of 94.27%, may be used in conjunction with machine learning and data preparation to improve telemarketing success in the banking industry.

In [26], Borugadda, et al., applied a number of machine learning algorithms. The logistic regression model outperformed other machine learning algorithms in predicting consumer interest in long-term deposits via telemarketing, with an accuracy of 92.48%.

In [27], Mylavarapu, presented a thorough approach to evaluating the quality of data that incorporates context extraction, accuracy assessment, and consistency assessment, highlighting the importance of identifying discrepancies in outcomes.

TABLE 1. Summary of Related Work.

Reference	Year	Dimensions	ML Models	Findings
Huang, [19]	2024	Completeness, Consistency, Accuracy and Validity.	SVM, NB, RF, KNN, LR, DT, XGB, and GB.	XGB has been proven to have the most outstanding performance, achieving a high F1 score of 90.85% and accuracy of 90.14%.
Kaisar, et al. [20]	2024	Completeness, Consistency, Accuracy and Validity.	RF, AdaBoost, GB and XGB.	RF achieved 98.6%, 93.4%, 98.1%, 0.975, and 0.954 in accuracy, precision, recall, AUC, and F1-score, respectively.
Huynh, et al. [21]	2023	Completeness, Consistency, Accuracy and Duplication.	KNN, LR, SVM and XGB.	KNN is the best prediction model achieved an accuracy of 91.07% and an AUC score of 0.9324.
Lam, et al. [22]	2023	Completeness, Consistency and Accuracy.	DT	Achieved 92% for accuracy, 99% for recall, 86% for precision, and 92% for f1-score.
Vongchalerm [23]	2022	Missing Data, Accuracy and Consistency.	LR and DT	The LR model performed better, with an AUC of 0.934, accuracy of 0.870, and dependability of 0.869.
Tékouabou, et al. [24]	2022	Missing Data, Accuracy and Consistency.	Class membership-based (CMB) classifier	The proposed CMB classifier had an accuracy of 97.3% and an AUC of 95.9%.
Masturoh, et al. [25]	2021	Accuracy, Completeness and Consistency.	MLP	Achieved an accuracy rate of 94.27%.
Borugadda, et al. [26]	2021	Accuracy, Relevance, Completeness and Consistency.	LR, DT, SVM, NB and RF.	LR, the best performing model, achieved the highest accuracy of 92.72%.
Mylavarapu [27]	2020	Completeness, Validity, Accuracy, and Consistency	Natural language processing and Deep neural networks.	underscored the significance of data context and multiple dimensions in quality assessment.

3. Data Quality Dimensions and Measures

Advances in information technology significantly impact society's organizational structure, governance model, decision-making, business strategy, and individual lifestyle. The quality of data is influenced by the nature of the data and analytical processes, with new technologies enabling organizations, governments, foundations, and private agencies to support and make analytical decisions [9], [28] and [29].

Accuracy and rapid processing are crucial for data quality in business operations. Traditional systems struggle to handle massive data, necessitating advanced technologies for faster processing. Data quality determines dataset quality, requiring specific requirements for analysis dependability [29], [30] and [31].

3.1 Data Quality

Data quality is a subjective concept influenced by its use case and domain, determining its suitability for use and meeting user requirements. It measures the reliability of data in terms of accuracy, completeness, consistency, relevance, and timeliness. It is primarily assessed in terms of data types, sources, and applications. Furthermore, the chosen dimensions significantly affect the information's accuracy and quality [29], [30], [32] and [33].

Organizations rely on data analysis for decision-making and activities. Governments use data analysis to improve citizen service and address challenges. Acquiring high-quality data is a priority, as lack can lead to devastating consequences [34], [5], [30] and [35].

3.2 Measuring Multi-Dimensional Data Quality

This section explores the dimensions of data quality, which serve as metrics for measuring and managing data quality.

data quality is a multifaceted concept that is closely related to its intended use. Assessing data quality is complex, as different users may have different criteria for what constitutes high-quality data [5], [36], [37] and [38]. The conceptual framework of data quality has been expanded, and the dimensions have been classified into intrinsic, contextual, representative, and accessibility characteristics, each of which has specific measures that help identify gaps and opportunities for improvement [39] and [40]. In line with this viewpoint, our research extends the conceptual model of information quality, as shown in Figure 1.

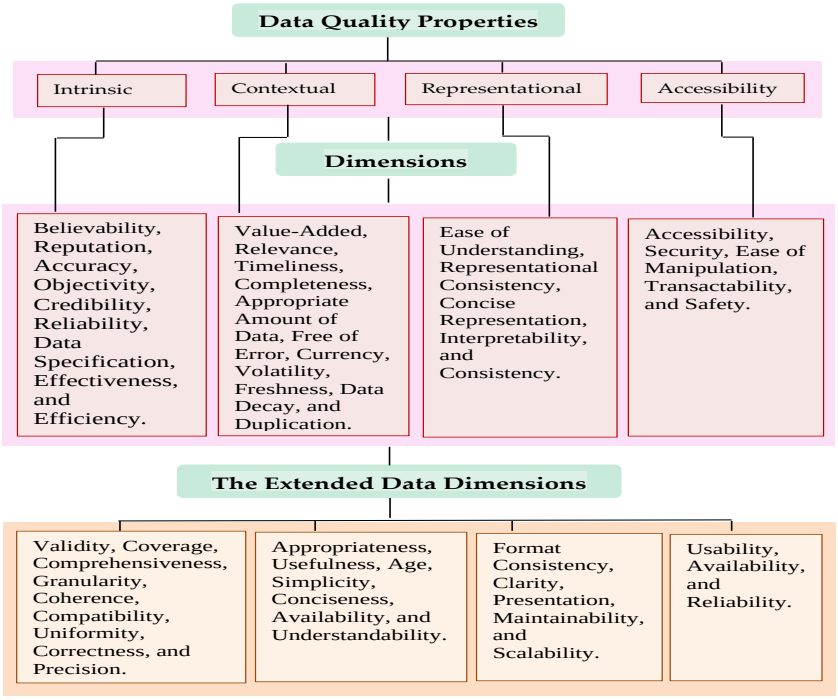

FIGURE 1. The Extended Conceptual Model of Data Quality Properties and Their Corresponding Data Quality Dimensions.

Understanding these dimensions is critical to enhancing data quality and supporting decision-making processes [29] and [41].

4. A Proposed Model for Multi-Dimensional Data Quality

This section details the model's components and methodologies, including data analysis, preprocessing, and machine learning algorithms, and the four dimensions of the proposed model: completeness, duplication, correctness, and consistency, and their use to measure information accuracy.

Understanding data nature and quality dimensions is crucial for evaluating them. A systematic approach is needed, with data cleansing being a primary task in business intelligence. Figure 2. provides a model that serves this purpose. This involves identifying and removing duplicates and errors and filling in missing values to improve data quality [11] and [42].

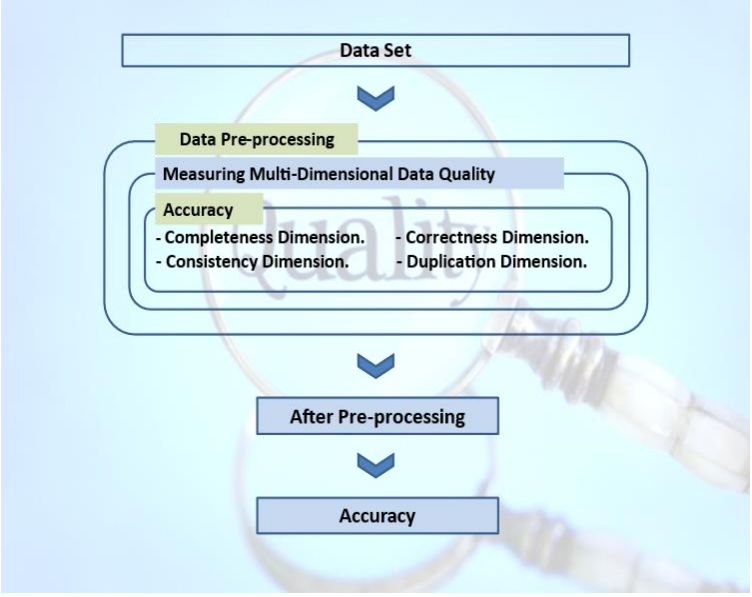

FIGURE 2. A Proposed Model for Multi-Dimensional Data Quality.

The quality evaluation process involves four main steps: dataset understanding, analysis, pre-processing, and assessment. Understanding customer characteristics is crucial for saving costs and time. The analysis process identifies quality issues like duplicate, inconsistent, incomplete, and incorrect data sets. Recognizing these issues helps take appropriate action. The treatment phase involves cleaning of data. Then assess and compare various strategies to find the best approach for correcting the identified issues.

4.1 Dataset Definition and Materials

This paper considers the real-world Portuguese bank dataset during telemarketing efforts. These marketing strategies involved direct phone conversations with bank consumers to encourage them to subscribe to a term deposit. The dataset was downloaded from the Machine Learning Repository at the University of California, Irvine (UCI). The dataset includes 41,188 instances and 21 features [19], [25], [43] and [44]. Whether or not customers signed up for the offer after the call is indicated by the "target Y", which is the target value represented by the binary (yes/no) results. The ratio of "yes" to "no" values is 88.8% of "no" responses and 11.2% of "yes", as shown in Figure 3.

This paper used Python to analyze and pre-process data, evaluate models, and compare their performance. It provides several libraries such as pandas, NumPy, Matplotlib, seaborn, and SciPy. It is an interactive, open source, object-oriented programming language.

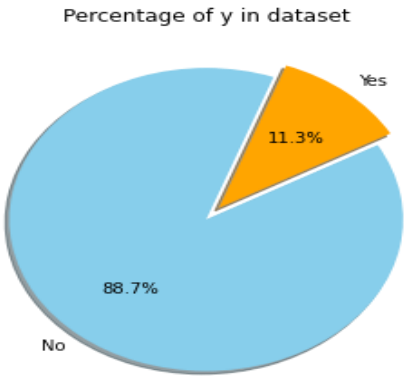

FIGURE 3. A Distribution of The Target "Y".

4.2 Data Preprocessing

The stage of preprocessing is critical to the prediction process because performance depends heavily on data cleaning operations that include processing and removing noise and inconsistent data, replacing missing values, and transforming and standardizing data to make it suitable for training machine learning algorithms on this data.

The initial data analysis involved checking for missing values, duplicates, and data distribution. The dataset was cleaned by removing duplicates and replacing “unknown” values using the mode method. Feature engineering involved converting categorical variables to numerical values. Various techniques, including “label encoding” and “one-hot encoding" are available to convert categorical data to numerical data. Since the latter technique has a much larger number of feature dimensions, the former strategy was used in this paper. The target variable, which indicates whether the customer subscribes to a term deposit, was encoded as binary values (1 for “yes” and 0 for “no”). To address the imbalance in data, the synthetic minority oversampling technique (SMOTE) was applied, as shown in Figure 4.

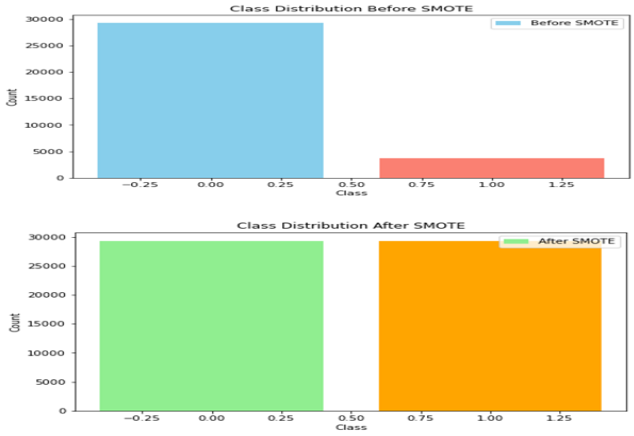

FIGURE 4. A Class Distribution Before and After Smote.

After the preprocessing of data, predictive five machine learning algorithm's such as Random Forest, Decision Tree, KNN, MLP, and Gradient Boosting. The dataset was split into training (70%), testing (20%), and validation (10%) sets, as shown in Figure 5.

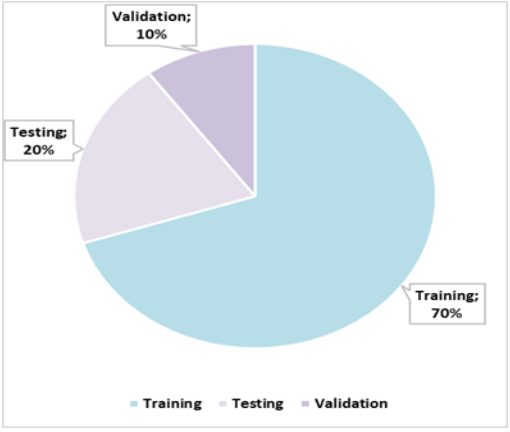

FIGURE 5. Training, Testing, and Validation for The Dataset.

5. Experimental and Results

In this section, the performance of Random Forest, Decision Tree, KNN, MLP and Gradient Boosting is analyzed and compared.

5.1 Performance Metrics

The performance of classification models Evaluated through a confusion matrix, accuracy, precision, recall, and F1 score.

- **Confusion Matrices**: The elements of the confusion matrix in this paper can be defined as shown in Table 2, where
  - **True Negative (TN):** When a class is expected to have no deposit but really does not.
  - **True Positive (TP):** When a class predicts a yes deposit, it actually does.
  - **False Positive (FP):** A class that is expected to deposit but not actually does.
  - **False Negative (FN):** When a class is expected to have no deposit but actually has one.

TABLE 2. Four Different Possible Outcomes That are Given by Confusion Matrices.

	Predicted Negative	Predicted Positive
Actual Negative	**TN**	**FP**
Actual Positive	**FN**	**TP**

- **Accuracy:** The model accuracy is a performance metric that measures the ratio of correctly predicted observations to total observations, referring to the number of accurately predicted samples. It is calculated by the following formula **(1)**:

$$Acc. = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

- **Precision:** Precision is the percentage of positive predictions that truly belong to a positive class, calculated by the following formula **(2)**:

$$\mathbf{PREC.} = \frac{TP}{TP+FP} \qquad (2)$$

- **Recall:** The proportion of accurately predicted positive observations out of all positive observations is calculated using formula **(3)**:

$$\mathbf{REC.} = \frac{TP}{TP+FN} \qquad (3)$$

- **F1-Score:** The F1 score evaluates a model's performance impartially, focusing on the precision and recall harmonic mean. A higher F1 score indicates an ideal balance between precision and recall. It is calculated by the following formula **(4)**:

$$F1. = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \qquad (4)$$

5.2 Analysis of Results

Confusion matrices of five classification algorithms are shown in Figure 6, in which the confusion matrix of gradient boosting has fewer wrong predictions as FP=198 and FN=418.

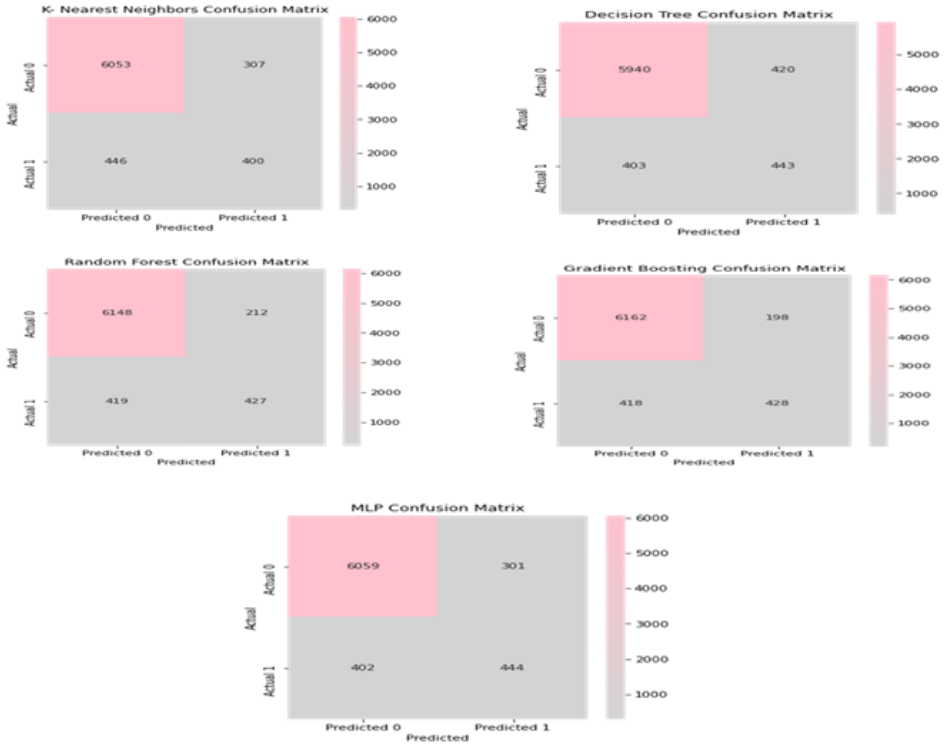

FIGURE 6. The Confusion Metrics of The Five Classification Algorithms.

The performance results of five classification models are displayed with four evaluation metrics in Table 3, also Training time (s), prediction time (s) and total time (s). Among the five models, the Gradient Boosting model gave better results with 91.45%, 91.45%, 90.68%, and 90.89% for accuracy, recall, precision, and F1-score, respectively.

TABLE 3. Comparison of Model Performance.

	MODEL	ACCURACY (%)	RECALL (%)	PRECISION (%)	F1-SCORE (%)	TRAINING TIME (S)	PREDICTION TIME (S)	TOTAL TIME (S)
0	**K-Neighbors Classifier**	89.56%	89.55%	88.85%	89.14%	0.01	1.09	1.10
1	**Decision Tree Classifier**	88.61%	88.58%	88.68%	88.63%	0.35	0.00	0.35
2	**Random Forest Classifier**	91.24%	91.24%	90.47%	90.70%	6.04	0.16	6.20
3	**Gradient Boosting Classifier**	91.45%	91.45%	90.68%	90.89%	6.16	0.02	6.17
4	**MLP Classifier**	91.13%	90.24%	89.77%	89.97%	2.30	0.00	2.30

5.3 Comparison with Previous Works

The results of our model were compared with the results in Ref. [19] in terms of accuracy as a basis for comparison as shown in Table 4. Since accuracy is the main metric in the proposed model, the gradient boosting classifier achieved superior performance compared to other classifiers. Among the existing works, Huang, [19] achieved an accuracy of 90.41% for the gradient boosting classifier, but the best performance overall was for the XGB classifier.

TABLE 4. Comparison with Previous Work.

REFERENCE	ACCURACY OF THE GRADIENT BOOSTING CLASSIFIER
HUANG, [19]	90.41%
PROPOSED MODEL	91.45%

6. Conclusion and Future Work

In this paper, the experimental results show that the graded boosting model has the best prediction effect. Among the five machine learning models, the graded boosting model has the highest accuracy value. The graded boosting model can predict deposit subscription to determine the success of the bank's direct marketing with an accuracy of 91.45%. KNN, decision tree, random forest, and multilayer classifier (MLP) achieved accuracy scores of 89.56%, 88.61%, 91.24%, and 91.13%, respectively. The decision tree model is actually the model that has the weakest performance, and based on the confusion matrix results, the boosting model has the least number of negative predictions (616) and the highest number of positive predictions (6590) compared to the other models.

There are limitations for future work. Outliers are detected and dealt with. There is a certain time when keeping outliers is important because they contain valuable information that forms part of the paper. Therefore, it is recommended to compare the results with and without outliers to assess the need to remove outliers. In addition, feature selection to filter the relevant data for prediction and remove irrelevant features to increase and improve the prediction accuracy of the models. Performance testing using new classifiers with additional dimensions.

References

1. Javid, I., et al., *Data pre-processing for cardiovascular disease classification: A systematic literature review.* Journal of Intelligent & Fuzzy Systems, 2023(Preprint): p. 1-21.
2. Maharana, K., S. Mondal, and B. Nemade, *A review: Data pre-processing and data augmentation techniques.* Global Transitions Proceedings, 2022. **3**(1): p. 91-99.
3. Djafri, L. and Y. Gafour. *Machine learning algorithms for big data mining processing: A review*. in *International Conference on Artificial Intelligence and its Applications*. 2021. Springer.
4. Anusha, Y., R. Visalakshi, and K. Srinivas, *Big Data Quality-A Survey paper to attain Data quality*. 2021.
5. Chernov, Y., *Data Quality Measurement Based on Domain-Specific Information*, in *Data Integrity and Data Governance*. 2022, IntechOpen.
6. Shirazi, F., et al., *New product success through big data analytics: an empirical evidence from Iran.* Information Technology & People, 2022. **35**(5): p. 1513-1539.
7. Ibrahim, A., M. Ibrahim, and N.S.M. Satar, *Factors influencing master data quality: A systematic review.* International Journal of Advanced Computer Science and Applications, 2021. **12**(2).
8. Yang, J., et al., *Social media data analytics for business decision making system to competitive analysis.* Information Processing & Management, 2022. **59**(1): p. 102751.
9. Ridzuan, F., W.M.N. Wan Zainon, and M. Zairul. *A Thematic Review on Data Quality Challenges and Dimension in the Era of Big Data*. in *Proceedings of the 12th National Technical Seminar on Unmanned System Technology 2020: NUSYS'20*. 2022. Springer.
10. Hasan, M.K., et al., *Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021).* Informatics in Medicine Unlocked, 2021. **27**: p. 100799.
11. Han, J., J. Pei, and H. Tong, *Data mining: concepts and techniques*. 2022: Morgan kaufmann.
12. Palanivinayagam, A. and R. Damaševičius, *Effective Handling of Missing Values in Datasets for Classification Using Machine Learning Methods.* Information, 2023. **14**(2): p. 92.
13. Widyananda, W., et al., *Application of Data Mining and Imputation Algorithms for Missing Value Handling: A Study Case Car Evaluation Dataset.* Iraqi Journal of Science, 2023: p. 2481-2491.

14. Zakeri-Nasrabadi, M., et al., *A systematic literature review on source code similarity measurement and clone detection: Techniques, applications, and challenges.* Journal of Systems and Software, 2023: p. 111796.
15. Opdenplatz, J., U. Şimşek, and D. Fensel, *Duplicate Detection as a Service.* arXiv preprint arXiv:2207.09672, 2022.
16. Mohseni, N., et al., *Outlier Detection in Test Samples using Standard Deviation and Unsupervised Training Set Selection.* International Journal of Engineering, 2023. **36**(1): p. 119-129.
17. Nassreddine, G., J. Younis, and T. Falahi, *Detecting Data Outliers with Machine Learning.* Al-Salam Journal for Engineering and Technology, 2023. **2**(2): p. 152-164.
18. Alzahrani, Y., J. Shen, and J. Yan. *Energy-Efficient Data Consistency based Sampling Rate Optimization and Aggregation Method for IoT*. in *2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. 2023. IEEE.
19. Huang, S., *A Comparative Analysis of Machine Learning Algorithms for Predicting the Telemarketing Campaigns of Portuguese Banking Institutions.* Advances in Economics, Management and Political Sciences, 2024. **94**: p. 44-53.
20. Kaisar, S., et al., *Enhancing Telemarketing Success Using Ensemble-Based Online Machine Learning.* Big Data Mining and Analytics, 2024. **7**(2): p. 294-314.
21. Huynh, L.D., et al. *Potential Customers Prediction in Bank Telemarketing*. in *Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 2*. 2023. Springer.
22. Lam, N.N.M., N.H. Tran, and D.H. Dinh. *Application of Decision Tree Algorithm for the Classification Problem in Bank Telemarketing*. in *International Conference on Intelligent Computing & Optimization*. 2023. Springer.
23. Vongchalerm, L., *Analysis of predicting the success of the banking telemarketing campaigns by using machine learning techniques*. 2022, Dublin, National College of Ireland.
24. Tékouabou, S.C.K., et al., *A machine learning framework towards bank telemarketing prediction.* Journal of Risk and Financial Management, 2022. **15**(6): p. 269.
25. Masturoh, S., et al., *TELEMARKETING BANK SUCCESS PREDICTION USING MULTILAYER PERCEPTRON (MLP) ALGORITHM WITH RESAMPLING.* Jurnal Pilar Nusa Mandiri, 2021. **17**(1): p. 19-24.

26. Borugadda, P., P. Nandru, and C. Madhavaiah, *Predicting the success of bank telemarketing for selling long-term deposits: An application of machine learning algorithms.* St. Theresa Journal of Humanities and Social Sciences, 2021. **7**(1): p. 91-108.
27. Mylavarapu, S.S.G.S., *Context-aware quality assessment of structured and unstructured data*. 2020, Oklahoma State University.
28. Moreno-Mateos, M.A. and D. Carou, *A note on big data and value creation*, in *Machine Learning and Artificial Intelligence with Industrial Applications: From Big Data to Small Data*. 2022, Springer. p. 1-18.
29. Wang, J., et al., *Overview of data quality: Examining the dimensions, antecedents, and impacts of data quality.* Journal of the Knowledge Economy, 2024. **15**(1): p. 1159-1178.
30. Soni, S. and A. Singh. *Improving Data Quality using Big Data Framework: A Proposed Approach*. in *IOP Conference Series: Materials Science and Engineering*. 2021. IOP Publishing.
31. Cichy, C. and S. Rass, *An overview of data quality frameworks.* Ieee Access, 2019. **7**: p. 24634-24648.
32. Bhatt, A., *Data quality–The foundation of real-world studies.* Perspectives in Clinical Research, 2023. **14**(2): p. 92-94.
33. Amina, O.G., and Ta'a, A. , *Big data and data quality dimensions: a survey.* JOURNAL OF SOFTWARE ENGINEERING & INTELLIGENT SYSTEMS, 2018. **Volume 3**(Issue 1).
34. Shi, P., et al., *Data consistency theory and case study for scientific big data.* Information, 2019. **10**(4): p. 137.
35. Altendeitering, M., *Design principles for data quality tools*. 2023.
36. Wook, M., et al., *Exploring big data traits and data quality dimensions for big data analytics application using partial least squares structural equation modelling.* Journal of Big Data, 2021. **8**: p. 1-15.
37. Ridzuan, F. and W.M.N.W. Zainon, *A Review on Data Quality Dimensions for Big Data.* Procedia Computer Science, 2024. **234**: p. 341-348.
38. Wang, J., et al., *Overview of data quality: Examining the dimensions, antecedents, and impacts of data quality.* Journal of the Knowledge Economy, 2023: p. 1-20.
39. Nasr, M., E. Shaaban, and M.I. Gabr. *Data quality dimensions*. in *Internet of Things—Applications and Future: Proceedings of ITAF 2019*. 2020. Springer.
40. Wang, R.Y. and D.M. Strong, *Beyond accuracy: What data quality means to data consumers.* Journal of management information systems, 1996. **12**(4): p. 5-33.

41. Gabr, M., Y. Helmy, and D. Elzanfaly, *Data quality dimensions, metrics, and improvement techniques.* Future Comput. Inf. J, 2021. **6**: p. 25-44.
42. Saraswat, P. and S. Raj, *Data pre-processing techniques in data mining: A Review.* International Journal of Innovative Research in Computer Science & Technology, 2022. **10**(1): p. 122-125.
43. Ghatasheh, N., I. Altaharwa, and K. Aldebei, *Modeling the Telemarketing Process using Genetic Algorithms and Extreme Boosting: Feature Selection and Cost-Sensitive Analytical Approach.* IEEE Access, 2023.
44. Xie, C., et al., *How to improve the success of bank telemarketing? Prediction and interpretability analysis based on machine learning.* Computers & Industrial Engineering, 2023. **175**: p. 108874.