The International Society of Data Scientists
Journal of Data Science and Artificial Intelligence
Vol. 3, No. 1. Online ISSN: 2831-4794
Available at https://www.isods.org/publication/

Evaluation on Classifier Performance Across Varied Data Types in Public Datasets Analysis

Nguyen Hoang Thong[a], Vo Van Tu[a], Nguyen Duy Tan[a], Vu Ngoc Thanh Sang[*,a]

[a]Faculty of Information and Technology, Saigon University, Ho Chi Minh City, Vietnam, 700000

ABSTRACT

In the rapidly evolving field of machine learning, the performance of classifiers is often influenced by the nature of the data they are applied to. This manuscript presents a comprehensive evaluation of the performance of various machine learning classifiers—Random Forest (RF), XGBoost (XGB), k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Artificial Neural Networks (ANN)—across three distinct data types: categorical, mixed, and numerical. Utilizing publicly available datasets and default classifier parameters, the study employs the F1-score as the key performance metric. The findings reveal that ensemble methods like RF and XGB excel in handling both categorical and mixed data types, while XGB stands out for numerical data. On the other hand, classifiers like KNN and SVM face challenges with categorical and mixed data types, respectively. Given that the study was conducted using default parameters, future work should focus on hyperparameter tuning to optimize classifier performance across different data types. This research serves as a valuable resource for machine learning practitioners, offering insights into effective model selection based on data characteristics.

Keywords: machine learning, data type, classifier performance

1. INTRODUCTION

In the ever-expanding landscape of data-driven decision-making, machine learning classifiers have emerged as indispensable tools for a multitude of applications, from medical diagnosis [5] and fraud detection [7] to natural language processing [13] and image recognition[17]. The effectiveness of these classifiers hinges on their ability to accurately categorize data points into predefined classes or categories. Hence, the evaluation of classifier performance is a critical endeavor that transcends disciplines and industries.

When designing a machine learning problem, one must carefully consider the nature of the data as it forms the hypothesis space [9]. In the context of tabular data, it can encompass numeric, categori- cal, or mixed variables. The selection of data types fundamentally influences every stage of the machine learning pipeline. Numeric data, such as sensor readings or financial metrics, often leads to regression or clustering tasks. Categorical data, like gender or prod- uct categories, is frequently associated with classification problems. The hypothesis space generation becomes more complex in scenar- ios where data types are mixed. The selection of an appropriate data collection method, whether it's through surveys, sensors, or web scraping, must align with the desired data types, ensuring that data is accurately captured and represented. Data preprocessing, including techniques like one-hot encoding for categorical data or normalization for numeric data, is essential to harmonize data types before feeding them into machine learning models. The effective utilization of data types across all phases of a machine learning project is pivotal for its overall success.

Comprehending the influence of different data types on the per- formance of machine learning classifiers is especially valuable. It provides us with essential insights into which types of data have the most impact on model predictions. This knowledge makes model decisions more transparent and builds trust in the model's reliabil- ity. It allows for a more straightforward interpretation of why the model makes specific predictions, which is vital to comprehend the explanation behind these outcomes. Additionally, this awareness of data types offers another substantial benefit by improving data quality. Identifying potential data quality issues early in the data preprocessing phase, such as missing values or outliers, enables corrective actions to be taken promptly. This issue results in cleaner more accurate, and reliable datasets. A well-

understood relation- ship between data types and machine learning classifiers makes the models more reliable and streamlines the process of using and interpreting their performance.

The study [18] is the most relevant reference. It focuses on as- sessing the influence of three distinct data types (numeric, category, and mixed) by estimating the efficacy of three classification algo- rithms: Random Forest, k-Nearest Neighbor, and Naïve Bayes. To facilitate their investigation, they leverage eight datasets from the UCI repository to train these algorithms. It is worth noting, how- ever, that the metrics employed in their study are based primarily on accuracy, which may not be ideal for evaluating datasets charac- terized by imbalances or for providing a comprehensive assessment of the classifiers' performance in general. Another related work, as described in [6], focuses on assessing the performance of the k-Nearest Neighbor algorithm when applied to datasets featuring a mix of numerical and binary data. In this research, multiple sim- ilarity measures are introduced, leveraging established distance metrics for the heterogeneous data to enhance the classification process. The study conducts experiments using six datasets and conducts comparisons between the their proposed measures and the conventional Euclidean distance metric.

In this paper, we experiment on 12 public datasets to estimate the impact of different data types on classifier performance models. The foundation of our study lies in the understanding that data is not uniform and manifests in various forms. We focus on three primary data types: numeric, categorical, and mixed. We contribute to a deeper understanding of the nuances inherent in the machine learning domain. We used the F1-score as a performance metric to comprehensively evaluate the model's performance, encompassing scenarios with imbalanced data. Our findings aim to empower machine learning practitioners and data scientists with valuable knowledge, facilitating more informed decisions when dealing with diverse data types.

2. RESEARCH METHODOLOGY

2.1 Dataset Description

Our methodology is built upon the use of publicly available datasets gathered from the internet in their raw, unprocessed form. Further- more, keeping the data unaltered reflects the real-world state of web-published sources. This authenticity is valuable for research aiming to mirror the conditions under which data is originally en- countered. Additionally, by establishing a foundational reference point with unprocessed data, we create a benchmark for future re- search. Consequently, researchers can use this unprocessed dataset as a starting point for comparison, allowing them to assess the impact of various preprocessing techniques more effectively. More- over, the raw data provides flexibility for exploration. It enables a wide range of preprocessing strategies to be applied in future studies, tailoring the data to specific research questions without bias from our initial preprocessing choices. Lastly, researchers can replicate our experiments and verify our findings with the same raw data, thus promoting transparency and ensuring that the entire research process is reproducible.

Table 1: Dataset Information

Dataset ID	Dataset Name	Dataset Type	No. of Dataset Instances	No. of Dataset Attributes
C1	Car Evaluation	Categorical	1,728	6
C2	Nursery	Categorical	12,960	8
C3	Congressional Voting Records	Categorical	435	16
C4	Lung cancer	Categorical	32	56
C5	Tic-Tac-Toe Endgame	Categorical	958	9
M1	Breast Cancer Wisconsin	Mixed	699	9
M2	Bank Marketing	Mixed	45,211	17
M3	Statlog (Heart)	Mixed	270	13
N1	Iris	Numeric	150	4
N2	Wine	Numeric	178	13
N3	Dry Bean Dataset	Numeric	13,611	16
N4	Glass Identification	Numeric	214	9

The Table 1 provides an insightful overview of datasets, each uniquely identified by a code that signifies its data type and order. The datasets are categorized into three distinct groups: Categori- cal (C), Mixed (M), and Numeric (N). Within each group, specific datasets are listed by their unique code, and their corresponding names span a wide range of applications. For example, in the Cate- gorical group, we find datasets such as "Car Evaluation" [8], "Nurs- ery" [16], "Congressional Voting Records" [1], "Lung cancer" [14], and "Tic-Tac-Toe Endgame" [4]. In the Mixed category, "Breast Can- cer Wisconsin (Original)" [19], "Bank Marketing" [15], and "Statlog (Heart)" [10] datasets are included. Finally, the Numeric group com- prises datasets like "Iris" [11], "Wine" [3], "Dry Bean Dataset" [2], and "Glass Identification" [12]. This structured presentation not only facilitates quick reference but also ensures that dataset sources can be cited accurately for future research and analysis within the domains of machine learning and data science.

2.2 Classifiers

The selected classifiers for our analysis encompass a diverse set of machine learning algorithms, including Logistic Regression (LR), Naive Bayes (NB), k-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Trees (DT), Random Forest (RF), XGBoost, and Artificial Neural Networks (ANN). By employing this array of classifiers, we aim to comprehensively assess how different data types interact with a variety of machine learning algorithms.

- **Logistic Regression (LR)**: At its core, Logistic Regression is a versatile algorithm widely used for binary and multi-class classi- fication. It operates on the principle of modeling the probability of a data point belonging to a specific class, employing a sigmoid function to constrain predictions between 0 and 1. The simplicity of its working principle, ease of interpretation, and efficiency in handling large datasets make it a popular choice. However, LR assumes linear relationships between features and may not perform optimally when data is highly non-linear.

- **Naive Bayes (NB)**: Naive Bayes is a probabilistic algorithm grounded in the Bayes' theorem. It is particularly adept at text classification and spam filtering. NB operates on the assumption of feature independence, simplifying computations by consider- ing each feature's impact in isolation. Its strengths include rapid training and prediction times, even with substantial feature sets. Nevertheless, the "naive" assumption of feature independence can be overly simplistic in practice, potentially impacting its perfor- mance when strong correlations exist between features.

- **k-Nearest Neighbors (KNN)**: K-Nearest Neighbors is an instance- based algorithm that classifies data points based on their proximity to their k-nearest neighbors in feature space. Its working principle is intuitive: if a majority of nearby data points belong to a specific class, the query point is assigned to that class. KNN excels in cap- turing complex decision boundaries and is robust to noisy data. However, it can be computationally expensive, particularly with large datasets, and requires careful consideration of the choice of k and distance metrics.

- **Support Vector Machine (SVM)**: SVM is a powerful classifi- cation algorithm based on the concept of finding an optimal hy- perplane that maximizes the margin between classes. Its working principle centers on identifying a decision boundary that not only classifies data points accurately but also generalizes well to unseen data. SVM is particularly effective in high-dimensional spaces and is versatile in handling linear and non-linear classification tasks. However, it can be sensitive to the choice of hyperparameters and computationally expensive, especially with large datasets.

- **Decision Trees (DT)**: Decision Trees, on the other hand, adopt a tree-like structure to make decisions by recursively partitioning the feature space based on attribute values. The working principle is intuitive, as each internal node represents a feature, and each branch represents a decision based on that feature. DTs are highly interpretable, making them valuable for understanding the logic behind classification. They are also robust to outliers and can handle both categorical and numerical data. Nevertheless, DTs are prone to overfitting, and their decision boundaries can be overly complex.

- **Random Forest (RF)**: Random Forest takes a different approach by combining multiple decision trees to achieve more robust and accurate classification. Its working principle involves creating an ensemble of decision trees, where each tree is trained on a random subset of the data and features. RF mitigates the overfitting tendency of individual decision trees while maintaining their interpretability. It excels in handling high-dimensional data and can provide feature importance rankings. However, it can be challenging to fine-tune, and the resulting model may be less interpretable than a single decision tree.

- **XGBoost (Extreme Gradient Boosting)**: XGBoost is a gradi- ent boosting framework renowned for its extraordinary predictive performance. Its working principle is rooted in ensemble learning, where multiple decision trees are trained sequentially to correct errors made by preceding trees. XGBoost employs gradient descent optimization to minimize a predefined loss function, allowing for a flexible and adaptive model. It excels in capturing complex relation- ships in data, handles missing values gracefully, and is robust to overfitting. However, fine-tuning XGBoost models can be intricate, and they may not be as interpretable as some other algorithms.
- **Artificial Neural Networks (ANNs)**: In contrast, Artificial Neural Networks mimic the intricate neural structure of the human brain. ANNs consist of interconnected nodes, or artificial neurons, organized in layers. The working principle involves training the network through backpropagation, adjusting weights and biases to minimize prediction errors. ANNs exhibit remarkable versatility and adaptability, capable of learning complex patterns in both struc- tured and unstructured data. They have achieved groundbreaking success in image recognition, natural language processing, and reinforcement learning. However, ANNs often require substantial computational resources, large datasets, and extensive hyperparam- eter tuning. Additionally, they may lack interpretability due to their complex architecture.

2.3 Evaluation Metrics

In machine learning, evaluating a model's performance is crucial, especially when dealing with imbalanced datasets, where one class significantly outnumbers another. Four key performance metrics commonly used in such scenarios are True Negative (TN), True Positive (TP), False Negative (FN), and False Positive (FP). TN and TP denote accurate predictions for the negative and positive classes, respectively, whereas FN and FP represent incorrect predictions where FN overlooks the presence of a positive class, and FP incorrectly identifies a negative class as positive. These metrics are essential for assessing how effectively a model distinguishes between classes, particularly in environments where class distribution is uneven.

- **Accuracy (ACC)**: Accuracy measures the ratio of correctly predicted instances to the total instances. It is a straightfor- ward metric but can be misleading in imbalanced datasets. It tends to be high even if the minority class is poorly predicted due to the dominance of the majority class.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision (PRC)**: Precision quantifies the ratio of true positive predictions to all positive predictions made by the model. It is especially useful when the cost of false positives is high.

$$PRC = \frac{TP}{TP + FP}$$

- **Recall (RC)**: Recall, also known as sensitivity or true posi- tive rate, measures the ratio of true positive predictions to all actual positive instances. It is particularly valuable when the cost of false negatives is high.

$$RC = \frac{TP}{TP + FN}$$

- **F1-Score**: The F1-Score is the harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives. It is suitable for imbalanced datasets as it addresses the trade-off between precision and recall.

$$F1 = \frac{2 \cdot PRC \cdot RC}{PRC + RC}$$

In summary, accuracy is not always the most informative metric for imbalanced datasets. Precision, recall, and the F1-Score offer a more nuanced evaluation of a model's performance, especially when one class is significantly smaller than the other.

2.4 Evaluation Metrics

In our research investigating the impact of data types on machine learning classifiers, we employ a robust evaluation methodology known as k-fold cross-validation, with k = 10. This technique is integral to our experimental design as it provides a comprehensive assessment of classifier performance across diverse data types. With k = 10, our dataset is divided into ten equally sized subsets or "folds." Each fold is used as the testing set in turn, while the remaining nine folds are used for training the classifier. This process is repeated ten times, ensuring that every data point is used for testing once. The results from each fold are then aggregated, offering a more reliable estimate of classifier performance. K-fold cross-validation is particularly advantageous when dealing with diverse data types, as it accounts for variations in the data splits, ultimately yielding a robust evaluation of how different data types impact classifier performance.

3. EXPERIMENT RESULTS

We conducted a manual review of the column definitions within each dataset to verify their data types. For categorical variables, we utilized the 'labelencoder()' function for encoding, and for data preprocessing, we applied 'minmaxscaler()' to standardize all data types. Subsequently, we trained our models using classifiers with default parameters from Scikit-learn library version 1.2.1. To assess model performance robustly, we employed k-fold cross-validation with k=10.

3.1 Classification Performance With Respect To Datasets

The Table 2 showcases the F1-score performance of chosen clas- sifiers on publicly available datasets. Indeed, the F1-score range within the presented table is quite extensive, spanning from a mini- mum value of 0.3752 to a maximum value of 1.000. This wide range of scores highlights the significant variability in classifier perfor- mance across the diverse dataset types and selected classifiers.

Table 2: F1-score Performance of Selected Classifier on Several Public Datasets

Datasets	Logistic Regression	Naive Bayes	KNN	SVM	Decision Tree	Random Forest	XGBoost	ANN
C1	0.9163	0.8656	0.8649	0.9703	0.9575	0.9550	0.9818	0.9863
C2	0.7557	0.8960	0.9326	0.9406	0.9946	0.9801	0.9995	0.9981
C3	0.6538	0.7178	0.5860	0.6854	0.6833	0.6695	0.6999	0.6695
C4	0.9677	0.9677	0.9562	0.9519	0.9677	0.9677	0.9812	0.9519
C5	0.6293	0.7135	0.7954	0.8728	0.8741	0.9359	0.9948	0.9200
M1	0.9558	0.9735	0.9555	0.9465	0.9474	0.9647	0.9560	0.9373
M2	0.8074	0.7483	0.7470	0.7268	0.7819	0.8419	0.8428	0.7850
M3	0.9252	0.9059	0.6515	0.6142	0.6890	0.7929	0.8129	0.8129
N1	1.0000	0.9668	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
N2	0.9791	0.9687	0.9454	0.9261	0.9831	0.9938	0.9931	0.9800
N3	0.8618	0.7974	0.7174	0.5979	0.8929	0.9264	0.9253	0.3752
N4	0.5290	0.5317	0.6420	0.6485	0.7012	0.8330	0.9812	0.7649

Across all classifiers, the C1 dataset consistently exhibits strong performance. ANN and XGB achieve the highest values, registering 0.9863 and 0.9818, respectively. Transitioning to the C2 dataset, it presents a mixed challenge for classifiers. While Random Forest and XGBoost excel with scores of 0.9995 and 0.9981, respectively, other classifiers such as LR and KNN face difficulties, scoring lower. In contrast, C3 showcases a challenging scenario for classifiers, with notably lower scores. However, SVM and DT perform relatively better, although all classifiers struggle to achieve high scores. Turn- ing to the C4 dataset, it witnesses consistently strong performance from most classifiers, with LR, NB, and DT leading the way. In an overall landscape, the C5 dataset presents a diverse challenge. Specifically, RF and XGB excel, while ANN and SVM face more sig- nificant difficulties, vividly demonstrating the dataset's complexity. These categorical datasets exhibit diverse characteristics, thereby challenging classifiers with varying degrees of complexity.

The M1 dataset consistently performs well, with ANN achieving the highest score of 0.9373, and most classifiers do a good job here, indicating that this dataset is a good fit for these models. Moving on to M2, it presents a mixed picture. DT and RF perform admirably, but NB struggles, resulting in a lower score. When we focus on M3, it's a challenging dataset for classifiers, as shown by their lower scores. In contrast, RF and ANN adapt well to this dataset's characteristics. As we explore these mixed datasets, we encounter various challenges, each with its level of difficulty. ANN stands out on the M1 dataset with the highest score, while M3 proves challenging for all classifiers. DT and RF consistently deliver steady performance across mixed datasets, highlighting their adaptability. M2 emerges as a dataset where RF and XGB are promising classifiers, producing outstanding results.

N1 demonstrates exceptional performance, with all classifiers achieving perfect scores of 1.0000, emphasizing how well-suited this dataset is for the models. Shifting to N2, it presents a favor- able environment for classifiers, resulting in high scores. RF and XGB particularly shine, showcasing their robustness. However, N3 proves to be quite demanding for classifiers, with lower scores over- all. In this challenging scenario, LR and NB are relatively better performers. Finally, the N4 dataset presents a mix of challenges. Here, RF, XGB, and ANN secure the highest scores, while SVM and DT face some difficulties with this dataset. These numeric datasets encompass various characteristics, posing unique challenges for classifiers due to their varying data types. N1 stands out as an ideal dataset, while RF and XGB consistently perform better across the numeric datasets, with a remarkable performance on N2. N3. On the other hand, it presents a formidable challenge for classifiers, evident from the lower scores it elicits across the board, underlining its inherent complexity.

3.2 Classification Performance With Respect To Data Types

The Figure 1 presented here represents a comprehensive evaluation of classifier performance across diverse data types, providing a valuable insight into the interaction between data characteristics and machine learning algorithms.

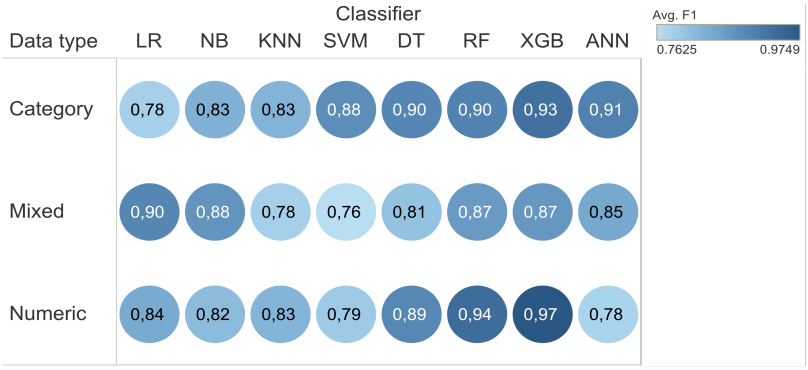


Figure 1: Classifiers with respect to data types

According to Figure 1, RF exhibits the highest average F1-score of 0.9016 when applied to the category data type. It is closely fol- lowed by XGB, with an average score of 0.9315. This result suggests that ensemble methods like RF and XGB perform exceptionally well when dealing with categorical data. These algorithms are capa- ble of capturing complex relationships within categorical features, making them suitable for this data type. On the other hand, the KNN algorithm has the lowest average F1-score of 0.8271 among all classifiers for the Category data type. Its performance is weaker because of distance metric dependence, which might not be as ef- fective for categorical data where the notion of distance can be less meaningful.

For mixture data types, RF and XGB both demonstrate strong per- formance, with average F1-scores of 0.8665 and 0.8706, respectively. These ensemble methods excel in handling mixed data, including categorical and numerical features. They can effectively capture patterns in both types of data. In contrast, SVM performs rela- tively poorly on the mixed data type, with an average F1-score of 0.7625. SVM inherently operates on numerical data and relies on distance calculations. Categorical features consisting of discrete labels do not naturally fit into this distance-based framework. Ad- ditionally, SVM relies on kernel functions to transform data into a higher-dimensional space where it can be more easily separated by a hyperplane. Selecting an appropriate kernel for mixed data can be non-trivial. Different features may require different kernel functions, and finding the right combination can be complex and require domain knowledge.

When dealing with numeric data types, XGB achieves the highest average F1-score of 0.9749 when applied to the numeric data type. XGBoost's ability to handle numerical data with complex relation- ships and robustness contributes to its outstanding performance. ANN records the lowest F1-score of 0.7800 among all classifiers for the numeric data type. ANNs consist of multiple layers of inter- connected neurons, and their performance heavily depends on the architecture and hyperparameters. ANN's performance is highly sensitive to its hyperparameters, including the number of layers, the number of neurons in each layer, the learning rate, activation functions, and more. Default hyperparameters may not be optimal for the selected numeric dataset, and hyperparameter tuning could significantly improve performance. In this case, other classifiers are more efficient.

4. CONCLUSIONS AND FUTURE WORKS

When it comes to handling categorical data, ensemble methods like RF and XGB emerge as the frontrunners. These classifiers ex- cel at capturing complex relationships within categorical features, making them highly suitable for this type of data. Their strength lies in their ability to aggregate multiple decision trees, improving the model's predictive accuracy and robustness. Conversely, the primary reason for KNN's weaker performance is its reliance on distance metrics, which may be less meaningful when dealing with categorical data. In essence, the notion of 'distance' becomes less effective when the data is not numerical, affecting the classifier's ability to make accurate predictions. For mixed data types, which include both numerical and categorical features, RF and XGB con- tinue to show strong performance. These classifiers are versatile and can effectively capture patterns in both data types. Their en- semble nature makes them less prone to overfitting, making them reliable choices for mixed data types. In contrast, SVM faces chal- lenges with mixed data. SVM inherently operates on numerical data and relies on distance calculations, making it less effective for handling categorical features. Moreover, the complexity of selecting an appropriate kernel function for mixed data types adds another layer of difficulty, often requiring specific domain knowledge for optimal performance. When dealing with purely numerical data, XGB stands out for its exceptional performance. Its strength lies in its ability to handle complex relationships within numerical data, making it a robust choice for such datasets. ANNs, however, record the lowest performance among all classifiers for numerical data. The performance of ANNs is highly sensitive to their architecture and hyperparameters. While they have the potential for high ac- curacy, their default settings may not always be optimal for every numerical dataset, indicating that hyperparameter tuning could sig- nificantly improve their performance. In conclusion, this evaluation is a valuable resource for machine learning practitioners, offering a nuanced understanding of how different classifiers perform across various data types. It highlights the strengths and weaknesses of each classifier, providing a roadmap for effective model selection based on the nature of the data at hand. Considering that the per- formance in this research was achieved using default parameters, future work should prioritize hyperparameter tuning to enhance the performance of classifiers across various data types.

REFERENCES

[1] "Congressional Voting Records," UCI Machine Learning Repository, 1987. [Online]. Available: https://doi.org/10.24432/C5C01P.

[2] "Dry Bean Dataset," UCI Machine Learning Repository, 2020. [Online]. Available: https://doi.org/10.24432/C50S4B.

[3] S. Aeberhard and M. Forina, "Wine," UCI Machine Learning Repository, 1991. [Online]. Available: https://doi.org/10.24432/C5PC7J.

[4] D. Aha, "Tic-Tac-Toe Endgame," UCI Machine Learning Repository, 1991. [Online]. Available: https://doi.org/10.24432/C5688J.

[5] A. Alanazi, "Using machine learning for healthcare challenges and opportunities," Informatics in Medicine Unlocked, vol. 30, 2022, Art. no. 100924.

[6] N. Ali, D. Neagu, and P. Trundle, "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets," SN Applied Sciences, vol. 1, 2019, pp. 1–15.

[7] R. B. Sulaiman, V. Schetinin, and P. Sant, "Review of machine learning approach on credit card fraud detection," Human-Centric Intelligent Systems, vol. 2, nos. 1-2, 2022, pp. 55–68.

[8] M. Bohanec, "Car Evaluation," UCI Machine Learning Repository, 1997. [Online]. Available: https://doi.org/10.24432/C5JP48.

[9] P. Domingos, "A Few Useful Things to Know about Machine Learning," Commun. ACM, vol. 55, no. 10, Oct. 2012, pp. 78–87, doi: 10.1145/2347736.2347755.

[10] D. Dua and C. Graff, "Statlog (Heart)," UCI Machine Learning Repository. [Online]. Available: https://doi.org/10.24432/C57303.

[11] R. A. Fisher, "Iris," UCI Machine Learning Repository, 1988. [Online]. Available: https://doi.org/10.24432/C56C76.

[12] B. German, "Glass Identification," UCI Machine Learning Repository, 1987. [Online]. Available: https://doi.org/10.24432/C5WW2P.

[13] A. Hodorog, I. Petri, and Y. Rezgui, "Machine learning and Natural Language Processing of social media data for event detection in smart cities," Sustainable Cities and Society, vol. 85, 2022, Art. no. 104026.

[14] Z. Q. Hong and J. Y. Yang, "Lung Cancer," UCI Machine Learning Repository, 1992. [Online]. Available: https://doi.org/10.24432/C57596.

[15] R. P. Moro, S., and P. Cortez, "Bank Marketing," UCI Machine Learning Repository, 2012. [Online]. Available: https://doi.org/10.24432/C5K306.

[16] V. Rajkovic, "Nursery," UCI Machine Learning Repository, 1997. [Online]. Available: https://doi.org/10.24432/C5P88W.

[17] P. Rani, S. Kotwal, J. Manhas, V. Sharma, and S. Sharma, "Machine learning and deep learning based computational approaches in automatic microorganisms image recognition: methodologies, challenges, and developments," Archives of Computational Methods in Engineering, vol. 29, no. 3, 2022, pp. 1801–1837.

[18] A. Singh, M. N. Halgamuge, and R. Lakshmiganthan, "Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms," Int. J. Adv. Comput. Sci. Appl., vol. 8, no. 12, 2017.

[19] W. Wolberg, "Breast Cancer Wisconsin (Original)," UCI Machine Learning Repository, 1992. [Online]. Available: https://doi.org/10.24432/C5HP4Z.