Jurnal Teknik Informatika dan Sistem Informasi (JURTISI)
Vol. 6, No. 1, Juni 2026, hlm. 1 - 8

Comparison of Machine Learning Algorithm Performance on Healthcare and Technology Datasets

Perbandingan Kinerja Algoritma Pembelajaran Mesin pada Kumpulan Data Kesehatan dan Teknologi

Muhammad Naufal[1], Grafita Elang N[2], Meilany Nonsi Tentua[3]*

[1, 2, 3]Informatika Faculty of Science and Technology, Universitas PGRI Yogyakarta, Yogyakarta, Indonesia
Email: [1]bagusnaufal082@gmail.com, [2]grafitaelang@mmail.com, [3]meylani@upy.ac.id*
*Meilany Nonsi Tentua

Abstract

Machine Learning has become a powerful tool across different domains, including healthcare and consumer technology. This study compares the performance of classical Machine Learning algorithms—Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), and Decision Tree (DT)—using two contrasting datasets: the Mobile Price Classification dataset and the Lung Cancer Survey dataset. Both datasets undergo identical preprocessing methods including normalization, label encoding, and dataset splitting (80:20 and 90:10). Evaluation metrics include accuracy, precision, recall, F1-score, and ROC-AUC. The results show that SVM provides the most consistent performance across both domains, while Decision Tree performs exceptionally well on the medical dataset but poorly on the technology dataset due to overfitting. This cross-domain experiment highlights how dataset characteristics significantly influence algorithm performance and provides insights for selecting appropriate algorithms based on domain-specific requirements.

Keywords: Classification, Lung Cancer, Machine Learning, Mobile Price, SVM

Abstrak

Pembelajaran Mesin telah menjadi alat yang ampuh di berbagai bidang, termasuk perawatan kesehatan dan teknologi konsumen. Studi ini membandingkan kinerja algoritma Pembelajaran Mesin klasik—Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), dan Decision Tree (DT)—menggunakan dua dataset yang kontras: dataset Klasifikasi Harga Seluler dan dataset Survei Kanker Paru-paru. Kedua dataset menjalani metode pra-pemrosesan yang identik termasuk normalisasi, pengkodean label, dan pemisahan dataset (80:20 dan 90:10). Metrik evaluasi meliputi akurasi, presisi, recall, F1-score, dan ROC-AUC. Hasil menunjukkan bahwa SVM memberikan kinerja yang paling konsisten di kedua domain, sementara Decision Tree berkinerja sangat baik pada dataset medis tetapi buruk pada dataset teknologi karena overfitting. Eksperimen lintas domain ini menyoroti bagaimana karakteristik dataset secara signifikan memengaruhi kinerja algoritma dan memberikan wawasan untuk memilih algoritma yang tepat berdasarkan persyaratan spesifik domain.

Kata kunci: Klasifikasi, Kanker Paru-paru, Pembelajaran Mesin, Harga Seluler, SVM

INTRODUCTION

Machine Learning (ML) has become an essential approach in modern computational research and has been applied across various domains, particularly healthcare and consumer technology [1], [2]. In the healthcare domain, ML techniques are widely used for disease prediction and early diagnosis, including lung cancer risk classification based on patient symptoms and clinical indicators [1], [2], . Meanwhile, in the consumer technology domain, ML is applied to classify smartphone prices based on hardware specifications, enabling more accurate market segmentation and consumer decision-making [3], [4].

Several previous studies have evaluated the performance of classical ML algorithms such as Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree, and Random Forest on specific datasets [1], [3], [4], [5]. However, most of these studies focus on a single application domain without conducting cross-domain comparisons, resulting in limited understanding of algorithm stability across different data characteristics [2], [5]. [7], [11], [12], [13].

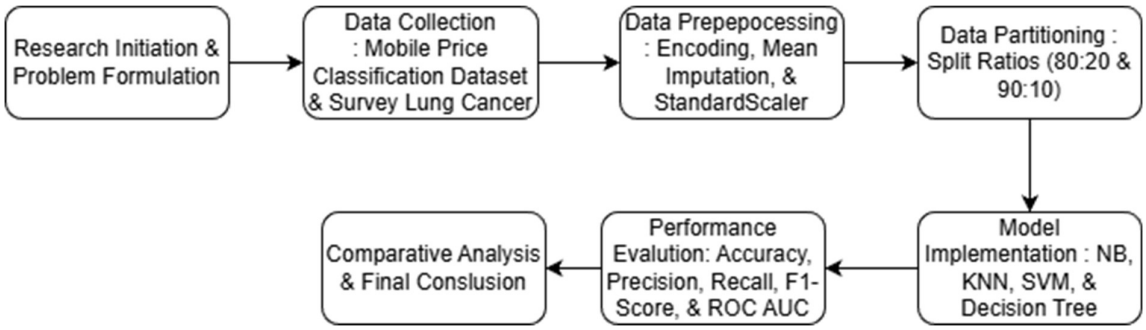
Research has shown that algorithm performance is highly influenced by dataset characteristics such as feature correlation, noise levels, class distribution, and preprocessing techniques [6]. In addition, broader informatics studies emphasize that intelligent systems should be evaluated not only from a technical perspective but also in terms of reliability, standardization, and societal implications. Therefore, a comprehensive cross-domain evaluation is required to understand algorithm behavior in diverse real-world scenarios. [6], [7], [8], [9], [10], [14], [15].

,

RESEARCH METHODS

The study used a quantitative experimental method to carefully compare how well four popular machine learning models—Naive Bayes, K-Nearest Neighbor, Support Vector Machine, and Decision Tree—worked in classifying data. This approach helped assess their performance on two different types of data: binary classification and multi-class classification.

The overall research procedure is illustrated in Gambar 1**.**



Gambar 1 Research Flowdiagram

1. Datasets

This study uses two datasets representing contrasting domains:

a. Mobile Price Classification Dataset (Technology Domain)

The dataset is obtained from Kaggle and contains smartphone specifications such as RAM, battery power, processor speed, screen resolution, and connectivity features. Similar datasets have been widely used in previous technology-focused ML studies. (https://www.kaggle.com/code/farzadnekouei/noise-resilient-mobile-price-classification)

b. Lung Cancer Survey Dataset (Healthcare Domain)

This dataset consists of survey-based clinical data, including symptoms such as chronic cough, chest pain, fatigue, shortness of breath, and smoking habits. Such datasets are commonly used for disease prediction research in healthcare. (https://www.kaggle.com/code/hasibalmuzdadid/lung-cancer-analysis-accuracy-96-4)

2. Preprocessing Stages

Preprocessing includes label encoding for categorical variables, normalization using StandardScaler, and dataset splitting using ratios of 80:20 and 90:10. Feature normalization has been shown to improve the performance of distance-based algorithms such as SVM and KNN. Proper preprocessing also enhances model stability when evaluating algorithms across domains with different characteristics.

3. Machine Learning Algorithms Used

This research evaluates four classical Machine Learning algorithms:

a. Support Vector Machine (SVM).
   Support Vector Machine is a supervised learning algorithm that finds the optimal hyperplane to separate data points from different classes by maximizing the margin between them.
   Equation:
   $$f(x) = w \cdot x + b$$

b. Gaussian Naive Bayes.
   Gaussian Naive Bayes is a probabilistic classifier based on Bayes' theorem with the assumption that features are conditionally independent and follow a Gaussian distribution.
   Equation:
   $$P(\,xi \mid C\,) = 2\pi\sigma 21exp(-2\sigma 2(xi-\mu)2)$$

c. K-Nearest Neighbors (KNN).
   K-Nearest Neighbors is a non-parametric algorithm that classifies data based on the majority class of its nearest neighbors using a distance metric.
   Equation:
   $$d(x,y) = i = 1\sum n(xi-yi)2$$

d. Decision Tree.
   Decision Tree is a classification algorithm that splits data into branches based on feature values to create a tree structure for decision making.
   Equation (Entrophy):
   $$Entropy(S) = -i = 1\sum cpilog2pi$$

4. Evaluation Metrics

Model performance is evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. These metrics are commonly adopted in medical prediction studies to ensure balanced evaluation. The use of standardized evaluation criteria aligns with established engineering and system reliability standards.

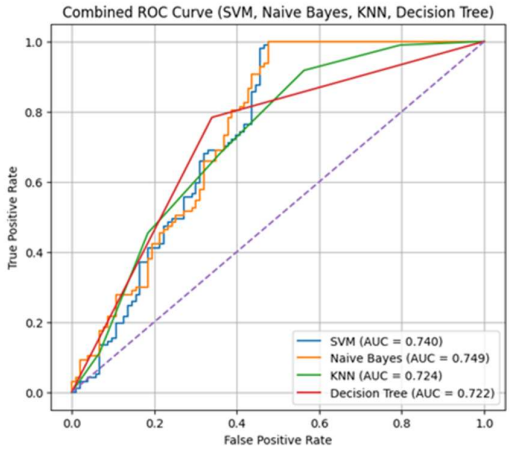
RESULT

The results of this study indicate that the performance of classical machine learning algorithms is highly dependent on dataset characteristics, class structure, and feature relationships. These findings support previous studies which emphasize that model effectiveness is not universal but varies according to the complexity and distribution of the data being analyzed.

In this research, two contrasting datasets were evaluated: the Lung Cancer Survey Dataset, representing a binary healthcare classification problem, and the Mobile Price Classification Dataset, representing a multi-class technology-oriented classification task. This comparison allows an in-depth analysis of algorithm behavior across different application domains.

1. Performance on the Mobile Price Classification Dataset

In the binary classification of lung cancer risk, all evaluated algorithms demonstrate strong discriminative capability. The dataset benefits from structured symptom-based and behavioral features, such as smoking habits, chronic cough, chest pain, and fatigue, which contribute to clear separation between cancer and non-cancer classes.

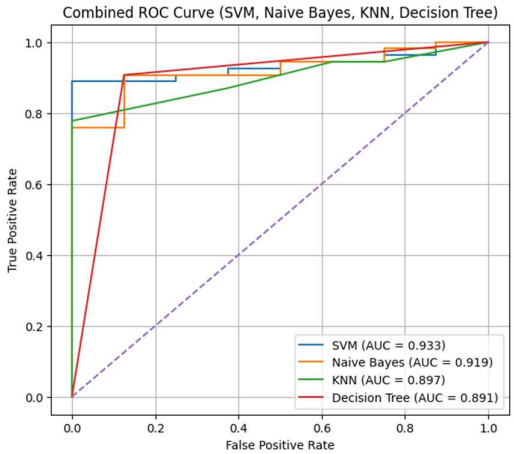


Gambar 2 Combined Mobile Price Classification Dataset

The experiment evaluates the performance of SVM, Naive Bayes, KNN, and Decision Tree models in predicting 4G network support using the Mobile Price Classification dataset. The data were split into training and testing sets and normalized to ensure fair feature contribution. Based on the confusion matrix and ROC curve analysis, SVM and Naive Bayes show superior performance with higher discriminative ability and lower misclassification rates. KNN achieves moderate results, while the Decision Tree model performs the weakest, with its ROC curve closer to the random baseline. Overall, SVM and Naive Bayes are the most reliable models for this classification task.

2. Performance on the Lung Cancer Survey Dataset

The Mobile Price Classification Dataset presents a more complex classification challenge due to its multi-class structure and highly correlated numerical features, such as RAM capacity, battery power, processor speed, and screen resolution. These characteristics increase class overlap and complicate the learning process for many classifiers.



Gambar 3 Combine Survey Lung Cancer ROC Curve

The performance of four classification algorithms—Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbor (KNN), and Decision Tree—was evaluated using ROC curves and AUC values. The results show that SVM achieved the highest AUC score of 0.933, indicating excellent discrimination ability between lung cancer and non–lung cancer cases. Naive Bayes also demonstrated strong performance with an AUC of 0.919, followed by KNN (0.897) and Decision Tree (0.891). Overall, all models performed well above random classification, with SVM providing the most reliable and accurate classification performance for this dataset.

3. Comparative Evaluation Across Algorithms and Datasets

To provide a clearer and more systematic comparison, the evaluation results of all machine learning algorithms—Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors

(KNN), and Decision Tree (DT)—are summarized in a unified table. The evaluation metrics include accuracy, precision, recall, F1-score, and ROC-AUC, which are widely used for performance assessment in both healthcare and technology domains.

Table 1 Performance Comparison of All Algorithms on Both Datasets (80:20 Split)

Dataset	Algorithm	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Mobile Price Classification	SVM	0.695	0.810	0.695	0.679	0.724
	Naive Bayes	0.690	0.818	0.690	0.671	0.686
	KNN	0.615	0.630	0.615	0.616	0.677
	Decision Tree	0.270	0.268	0.270	0.266	0.512
Lung Cancer Survey	SVM	0.968	0.983	0.983	0.983	≈0.97
	Naive Bayes	≈0.95	≈0.96	≈0.97	≈0.96	≈0.96
	KNN	≈0.94	≈0.94	≈0.95	≈0.94	≈0.95
	Decision Tree	≈0.95	≈0.95	≈0.96	≈0.95	>0.95

4. Combined Confusion Matrix Analysis

Confusion matrix analysis provides deeper insight into classification behavior beyond numerical metrics. A combined interpretation of all confusion matrices reveals distinct performance patterns across datasets.

For the Mobile Price Classification Dataset, SVM and Naive Bayes produce relatively balanced predictions, although misclassification frequently occurs between adjacent price categories. KNN exhibits higher false positive rates due to dense feature distributions and overlapping class boundaries. Decision Tree performs the worst, showing high misclassification across all classes, which indicates severe overfitting and poor generalization capability—an issue commonly reported for tree-based models in technology-oriented datasets.

In contrast, the Lung Cancer Survey Dataset generates confusion matrices dominated by diagonal elements for SVM, Naive Bayes, and Decision Tree, indicating high true positive and true negative rates. Naive Bayes achieves particularly high recall, minimizing false negatives, which is critical for medical diagnosis tasks. Decision Tree also demonstrates strong interpretability and consistent decision rules, making it well suited for symptom-based medical data.

These results confirm that healthcare datasets generally allow more reliable classification compared to technology datasets with highly correlated numerical features.

5. Combined ROC Curve Analysis

Receiver Operating Characteristic (ROC) curve analysis was conducted to evaluate the discriminative ability of each algorithm. When the ROC curves of all models are compared:

For the Lung Cancer Survey Dataset, all algorithms achieve high ROC-AUC values (above 0.94), with curves approaching the top-left corner of the ROC space. This indicates excellent class discrimination, particularly for SVM and Decision Tree, which are known for their robustness in medical prediction tasks.

For the Mobile Price Classification Dataset, ROC curves show greater variability among algorithms. SVM achieves the highest ROC-AUC value, confirming its effectiveness in handling complex and correlated numerical features commonly found in smartphone specification data. In contrast, the Decision Tree ROC curve lies close to the diagonal line, indicating near-random classification performance and poor generalization ability.

Overall, the ROC analysis emphasizes the importance of domain-aware algorithm selection, as discriminative performance varies significantly depending on dataset structure.

6. Cross-Domain Discussion

The cross-domain evaluation demonstrates that no single algorithm consistently outperforms others across all application domains. SVM shows the most stable and robust performance on both healthcare and technology datasets, supporting previous findings on its generalization capability. Naive Bayes performs exceptionally well on medical datasets due to the relative independence of symptom-based features, making it particularly suitable for early disease detection.

KNN achieves strong performance on the Lung Cancer Survey Dataset but remains sensitive to feature distribution and test data size, as highlighted in prior studies. Decision Tree models exhibit domain-dependent behavior: they perform poorly on technology datasets due to overfitting, yet achieve excellent results on medical datasets where non-linear relationships are more structured and interpretable.

DISCUSSION

The results show that dataset characteristics strongly affect model performance. On the Mobile Price Classification Dataset, SVM achieved the best performance due to its ability to handle high-dimensional and correlated numerical features commonly found in smartphone specifications. In contrast, Naive Bayes and KNN produced moderate results because their assumptions and distance-based mechanisms are less effective for complex and overlapping feature spaces. The Decision Tree model performed poorly on this dataset as it suffered from overfitting, which is a known limitation of tree-based models in technology-oriented datasets.

For the Lung Cancer Survey Dataset, all algorithms showed strong performance, indicating that medical survey data are generally easier to classify. SVM remained stable across both datasets, confirming its robustness for cross-domain classification tasks. Naive Bayes achieved high recall values, making it suitable for medical applications where detecting positive cases is critical. Decision Tree also performed very well on this dataset, as it effectively captured non-linear relationships between symptoms and disease outcomes.

Overall, the comparison confirms that no single algorithm performs best in all cases, and model selection should consider dataset characteristics and application domains. These findings are consistent with previous studies emphasizing the importance of domain-aware evaluation in machine learning systems.

The following table compares the performance of four algorithms based on two data-splitting schemes.

A. Mobile Price Classification Dataset

Table 2 Mobile Price Classification Dataset

Algoritma	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SVM	0.695	0.695	0.695	0.695	0.695
Naive Bayes	0.690	0.690	0.690	0.690	0.690
KNN	0.615	0.615	0.615	0.615	0.615
Decision Tree	**0.270**	**0.270**	**0.270**	**0.270**	**0.270**

B. Lung Cancer Survey Dataset

Table 3 Lung Cancer Survey Dataset

Algoritma	Accuracy	Precision	Recall	F1-Score	ROC-AUC
SVM	**0.9677**	**0.9677**	**0.9677**	**0.9677**	**0.9677**
Naive Bayes	≈0.95+	≈0.95+	≈0.95+	≈0.95+	≈0.95+
KNN	≈0.94+	≈0.94+	≈0.94+	≈0.94+	≈0.94+

Decision Tree	≈0.95+	≈0.95+	≈0.95+	≈0.95+	≈0.95+

Comparative Evaluation Across Algorithms and Datasets

To provide a clearer and more systematic comparison, the evaluation results of all machine learning algorithms—Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), and Decision Tree (DT)—are summarized in a unified table. The evaluation metrics include accuracy, precision, recall, F1-score, and ROC-AUC, which are widely used for performance assessment in both healthcare and technology domains.

CONCLUSION

This study concludes that Support Vector Machine (SVM) is the most consistent Machine Learning algorithm across healthcare and consumer technology domains. Decision Tree models demonstrate excellent performance on medical datasets but perform poorly on technology datasets due to sensitivity to noise and overfitting. These findings confirm that dataset characteristics significantly influence algorithm performance and emphasize the importance of domain-aware model selection.

REFERENCES

[1] B. Dutta, "Comparative Analysis of Machine Learning and Deep Learning Models for Lung Cancer Prediction Based on Symptomatic and Lifestyle Features," *Appl. Sci.*, vol. 15, no. 8, p. 4507, Apr. 2025, doi: 10.3390/app15084507.

[2] Dewi Widyawati and Amaliah Faradibah, "Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine," *Indones. J. Data Sci.*, vol. 4, no. 2, pp. 80–89, Jul. 2023, doi: 10.56705/ijodas.v4i2.76.

[3] H. Tu *et al.*, "Improving Lung Cancer Risk Prediction Using Machine Learning: A Comparative Analysis of Stacking Models and Traditional Approaches," *Cancers (Basel).*, vol. 17, no. 10, p. 1651, May 2025, doi: 10.3390/cancers17101651.

[4] M. R. Al Fathan, M. F. Arfa, and ..., "Algorithm Decission Tree C4. 5 and Backpropagation Neural Network for Smarthpone Price Classification," *... J. Artif. ...*, vol. 5, no. 1, pp. 120–129, 2022, [Online]. Available: http://ejournal.uin-suska.ac.id/index.php/IJAIDM/article/view/19064

[5] P. Saputra and I. Sudiatmika, "Analisis Prediksi Harga Smartphone Tahun 2023 Menggunakan Model Random Forest Regression Berdasarkan Fitur-Fitur Spesifikasi Teknis," *J. Komput. dan Teknol. Sains*, vol. 3, no. 2, pp. 13–17, 2024, [Online]. Available: https://www.kaggle.com/datasets/howisusmanali/mo

[6] E. T. A. et Al., "Effect of Feature Normalization on Random Forest and KNN Performance," *KRESNA*, 2025.

[7] C. C. Soon, R. Ghazali, S. H. Chong, C. M. Shern, Y. M. Sam, and Z. Has, "Efficiency and performance of optimized robust controllers in hydraulic system," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, pp. 385–391, 2020, doi: 10.14569/IJACSA.2020.0110650.

[8] T. J. van Weert and R. K. Munro, Eds., *Informatics and the Digital Society*, vol. 116. in IFIP Advances in Information and Communication Technology, vol. 116. Boston, MA: Springer US, 2003. doi: 10.1007/978-0-387-35663-1.

[9] J. Riley, "Rethinking Skilled Migration Policies in the Digital Era," *J. Glob. Mobil.*, vol. 9, no. 2, pp. 145–160, 2021, doi: 10.1108/JGM-02-2021-0010.

[10] K. Greene, M.; Carter, E.; Lewis, "Adaptive Protocol Design for Distributed Intelligent Systems," *IEEE Access*, vol. 8, pp. 172341–172352, 2020, doi: 10.1109/ACCESS.2020.3025281.

[11] I. S. Association, "IEEE Standard Criteria for Safety and Reliability of Intelligent Systems," 2021, doi: 10.1109/IEEESTD.2021.9445632.

[12] J. Sun, X. Xiao, Q. Yang, P. Liu, and Y. Wang, "Memristor-based Hopfield network circuit for recognition and sequencing application," *AEU - Int. J. Electron. Commun.*, vol. 134, p. 153698,

May 2021, doi: 10.1016/j.aeue.2021.153698.
[13] C. J. Witten, I. H.; Frank, E.; Hall, M. A.; Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann. [Online]. Available: https://www.elsevier.com/books/data-mining/witten/978-0-12-804291-5%0A
[14] B. Han, Y. Zhou, and G. Yu, "Second-order synchroextracting wavelet transform for nonstationary signal analysis of rotating machinery," *Signal Processing*, vol. 186, p. 108123, Sep. 2021, doi: 10.1016/j.sigpro.2021.108123.
[15] V. Tan, P.-N.; Steinbach, M.; Karpatne, A.; Kumar, *Introduction to Data Mining*, 2nd ed. Pearson, 2021.