Evaluate and Contrast the Classification Techniques

Gladys Murage

College of Business, Engineering, and Technology, National University

TIM-8131 v2: Data Mining (8221718749)

Dr. Yuksel Karahan

December 06, 2025

Evaluate and Contrast the Classification Techniques

Research Summary

This study explored the comparative performance of three supervised classification algorithms, Logistic Regression, Random Forest, and k-Nearest Neighbors, on the Heart Disease Prediction dataset. The dataset contained 303 patient records with 13 clinical attributes, making it suitable for binary classification tasks. Each algorithm was implemented using Python's scikit-learn, with preprocessing, hyperparameter tuning, and evaluation conducted through cross-validation and performance metrics analysis.

Results revealed that Logistic Regression consistently outperformed the other models, achieving the highest accuracy (85.7%), strongest generalization, and most favorable bias-variance balance. Its interpretability and computational efficiency further reinforced its suitability for structured, linearly separable data. Random Forest, while theoretically robust and capable of identifying feature importance (with chest pain type emerging as the strongest predictor), showed signs of overfitting and required careful tuning to improve variance control. k-Nearest Neighbors underperformed, struggling with scalability and sensitivity to feature representation, highlighting its limitations in higher-dimensional datasets.

The findings emphasize that algorithm effectiveness depends on aligning model strengths with dataset characteristics and research objectives. Logistic Regression is best suited for interpretable, small-to-medium structured datasets; Random Forest excels in larger, complex datasets when properly tuned; and k-NN remains useful for small, low-dimensional problems with careful preprocessing. These insights provide practical guidelines for researchers, reinforcing that algorithm selection should balance accuracy, interpretability, computational feasibility, and problem complexity.

Introduction

Machine learning (ML) is considered a major component of the broader field of artificial intelligence (AI). ML applies statistical methods to giving computers the ability to learn and make decisions autonomously, without the need for explicit programming. ML relies on the ability of computers to acquire knowledge from data, identify patterns, and draw conclusions with minimal human intervention. ML is divided into 4 distinct categories: supervised learning, unsupervised learning, semi supervised learning, and reinforcement learning. Supervised learning involves training models using labeled datasets and consists of two primary methods: classification and regression. Regression entails the use of numeric data and gives continuous output, while classification uses categorical data and categorizes data into predefined classes as output (Alnuaimi, & Albaldawi, .2024)

Part 1: Research and Algorithm Selection

Classification is a supervised learning task where algorithms predict categorical labels based on input features. This assignment evaluates three classification algorithms from distinct categories: Random Forest chosen from decision tree-based methods, k-Nearest Neighbors chosen from distance-based methods, and Logistic Regression chosen from linear models. Theoretical foundations, hands-on implementation, and empirical comparisons are provided.

Random Forest (Decision Tree-Based)

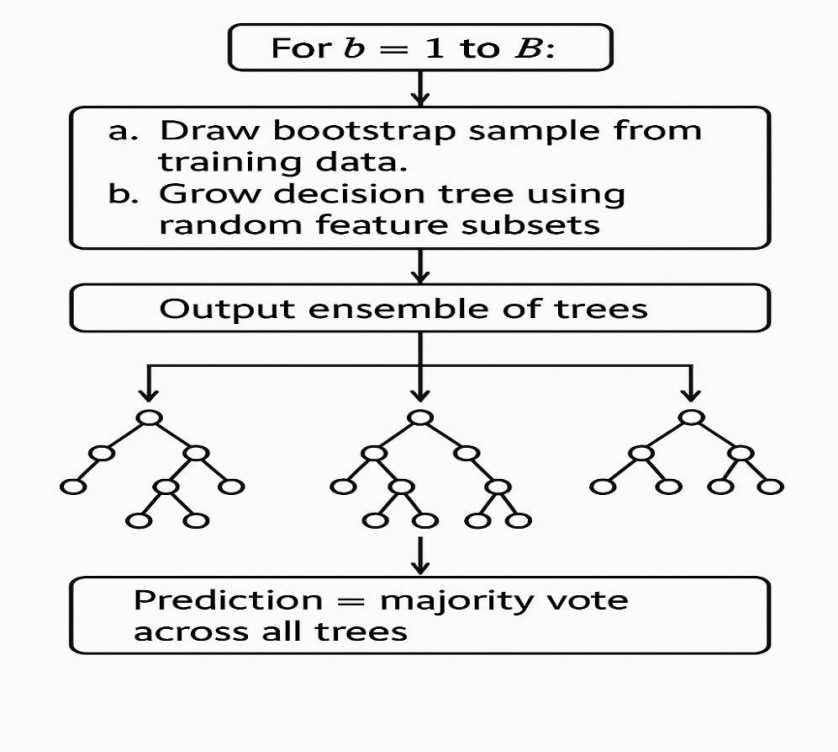
According to Salman et al. (2024). Random Forest (RF) is an extremely popular ML technique used in the field of data mining While data mining involves both descriptive and predictive data mining, RF is a ML model used in classification and predictive forecasting and RF is created by generating several decision trees. This is achieved by collecting random

samples of data using Bootstrap samples and then selecting random features. An advantage of RF is its high accuracy compared to other approaches such as bagging and boosting. Secondly RF function effectively on huge databases and are able to accommodate many variables without forcing deletion of any variables which allows analysis of thousands of input variables. RF can be used for classification or regression

Architecture Overview. Random Forest is an ensemble method that constructs multiple decision trees during training. Each tree is built using a bootstrap sample of the data and a random subset of features to mitigate the correlation of decision trees. Predictions are made via majority voting in classification or by averaging in regression to improve accuracy and robustness (Salman et al., 2024) Refer to Figure 1 for architecture of a classification Random Forest.

Figure 1

Random Forest Classification



Key Features. Random forest is a robust ensemble learning method that mitigates overfitting by combining bagging with feature randomness, ensuring diverse and less correlated decision trees. It excels in handling high-dimensional datasets and is resilient to missing values, making it suitable for complex real-world applications. Additionally, it offers valuable insights by providing feature importance scores, helping identify the most influential predictors in the model (Salman et al., 2024).

Strengths. Random Forest is a versatile ensemble method known for its high predictive accuracy and robustness against noisy data, making it well-suited for complex datasets. It identifies key variables and outliers, supports both supervised and unsupervised learning tasks, and can impute missing values using a proximity-based algorithm. By eliminating the need for feature scaling, Random Forest streamlines preprocessing and accelerates model development. Additionally, it incorporates built-in cross-validation through out-of-bag (OOB) error estimation, offering an unbiased performance measure without requiring a separate validation set. Its ability to effectively manage missing data further enhances its reliability (Salman et al., 2024).

Limitations. Random forest, while powerful, comes with certain limitations. It can be computationally intensive when applied to very large datasets, requiring significant processing time and resources due to the extensive number of decision trees involved (Salman et al., 2024). Secondly, compared to single decision trees, it is less interpretable, making it harder to understand the exact decision-making process of the ensemble. Thirdly, if not carefully tuned, random forest models may still overfit noisy datasets, reducing their generalization ability and predictive reliability.

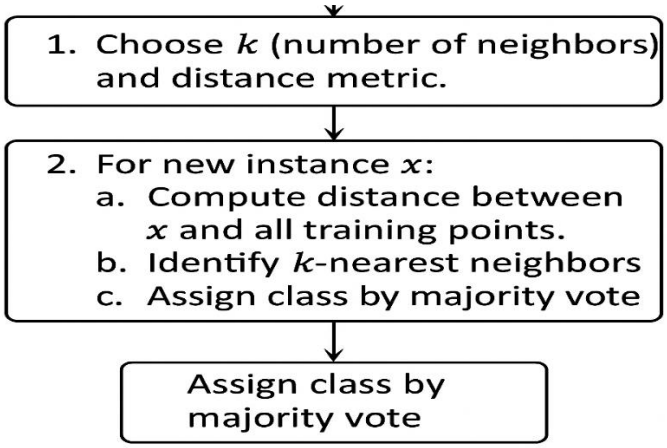
k-Nearest Neighbors (Distance-Based)

According to Syriopoulos et al. (2025), k-Nearest Neighbors (k-NN) is well known and a widely used algorithm. k-NN is an algorithm that is simple but powerful and a non-parametric classifier that is robust to noisy data and easy to implement. It is a supervised learning classifier that carries out classification regarding the grouping of a given data point. While k-NN can be used for both classification and regression, it is typically used as a classification algorithm based on the assumption that similar points are located near each other. Most generally , k-NN utilizes the Euclidean distance in its calculations.

Architecture Overview. k-NN is an instance-based learning algorithm that operates by storing all training examples and using them directly for prediction rather than building an explicit model. When a new data point is introduced, k-NN determines its class by identifying the closest k training instances according to a chosen distance metric, such as Euclidean distance, and assigning the majority class among those neighbors. This simplicity makes k-NN intuitive and effective for classification tasks, though its performance depends heavily on the choice of k and the distance measure. Refer to Figure 2 for an architecture diagram.

Figure 2

k-NN Architecture Diagram



Key Features. k-NN is a lazy learning algorithm, meaning it does not involve an explicit training phase and instead defers computation until prediction time. As a non-parametric method, it makes no assumptions about the underlying data distribution, offering high flexibility in modeling complex decision boundaries.

Strengths. k-NN is simple to implement and easy to interpret, making it accessible for a wide range of users and applications (Syriopoulos et al., 2025). Secondly, k-NN operates without making assumptions about the underlying data distribution, allowing it to adapt to various data types and structures. Thirdly, it naturally supports multi-class classification tasks, making it versatile for problems involving more than two categories.

Limitations. k-NN can be computationally expensive at prediction time since it must calculate distances to all training points. k-NN performance is sensitive to irrelevant features and imbalanced datasets, which can distort neighbor relationships and majority voting. Moreover, the algorithm requires careful selection of both the value of k and the distance metric, as these choices significantly influence accuracy and generalization (Syriopoulos et al., 2025)

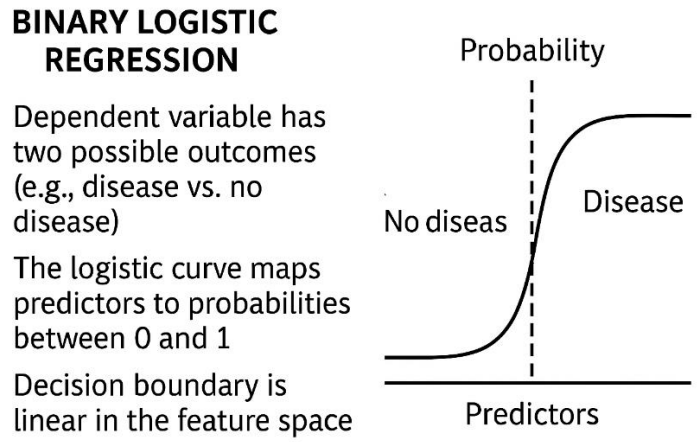
Logistic Regression (Linear Model)

Logistic regression is a statistical method used to analyze the relationship between one or more independent variables and a binary outcome. Instead of predicting exact values, it estimates the probability that an event will occur, using the logistic (sigmoid) function to map inputs to values between 0 and 1. The strength and direction of each predictor's influence are captured through regression coefficients. When the dependent variable has only two possible outcomes, the technique is referred to as binary logistic regression. If the dependent variable has more than two categories, an extension known as multinomial logistic regression is applied (Elkahwagy et al., 2024).

Architecture Overview. Logistic Regression is a statistical learning method that estimates the probability of a binary outcome using the logistic (sigmoid) function. It constructs linear decision boundaries and optimizes them through maximum log-likelihood, typically solved with gradient descent, making it effective for classification tasks. Refer to Figure 3 for the binary logistic regression architecture.

Figure 3

Binary Logistic Regression Architecture



Key Features. Logistic regression is a probabilistic classification model that outputs the likelihood of a given instance belonging to a particular class, typically binary. It does this by modeling the log-odds of the outcome as a linear combination of the input features, meaning it assumes a linear relationship between the predictors and the log-odds of the target variable (Hua et al., 2025). This assumption allows the model to be both interpretable and efficient, making it suitable for problems where the decision boundary is approximately linear and understanding feature influence is important. An example of such use is for example in classifying a disease diagnosis as being present or absent,

Strengths. Logistic regression is a computationally efficient and highly interpretable classification algorithm, making it ideal for quick deployment and transparent decision-making. It provides confidence scores for predictions by estimating class probabilities, which adds nuance to binary outcomes. The model performs particularly well when the data is linearly separable, leveraging its linear decision boundaries to achieve accurate and reliable results (Hua et al., 2025).

Limitations. Logistic regression assumes a linear relationship between the input features and the log-odds of the outcome, which can limit its ability to capture complex nonlinear patterns and lead to model misspecification and misinterpretation of results (Hua et al., 2025). Secondly, the model is sensitive to multicollinearity, where highly correlated features can distort coefficient estimates and reduce interpretability, making careful feature selection and preprocessing important for reliable performance (Elkahwagy et al., 2024).

Part 2: Hands on Implementation

Dataset Description

The Heart Disease Prediction Dataset from the UCI Machine Learning Repository is obtained from the source: https://archive.ics.uci.edu/ml/datasets/heart+disease. The data set contains data from 303 patients, with 13 clinical attributes such as age, cholesterol levels, and chest pain type. The dataset is designed for binary classification tasks, aiming to predict the presence or absence of heart disease. Preprocessing steps include handling missing values and scaling numerical features to optimize performance for algorithms like k-NN**.**

Implementation

For this project, Python with scikit-learn is employed to implement the machine learning workflow, which includes a train-test split of 70% training and 30% testing. Hyperparameter

tuning via grid search cross validation using a 5-fold is conducted followed by model training and evaluation. The coding component is prepared and submitted separately in the file MurageGTIM8131-3.ipynb, ensuring that the technical implementation is clearly documented and reproducible.

Results

Table 4

Comparative Performance Metrics for the Three Algorithms

COMPARATIVE PERFORMANCE METRICS

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.857	0.858	0.857	0.857	0.939
Random Forest	0.835	0.836	0.835	0.835	0.92
k-NN	0.824	0.825	0.824	0.824	0.839

Figure 5

Performance Metrics Comparison

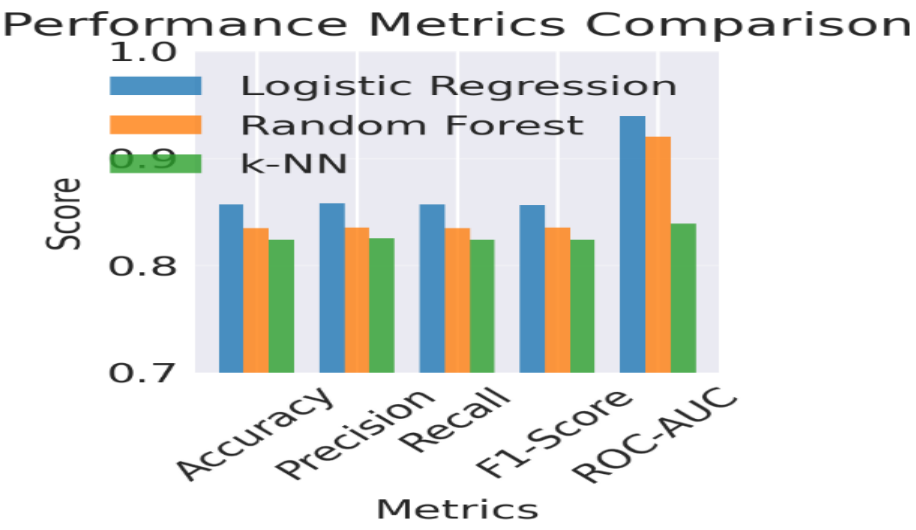


Figure 6

Confusion Matrices Analysis

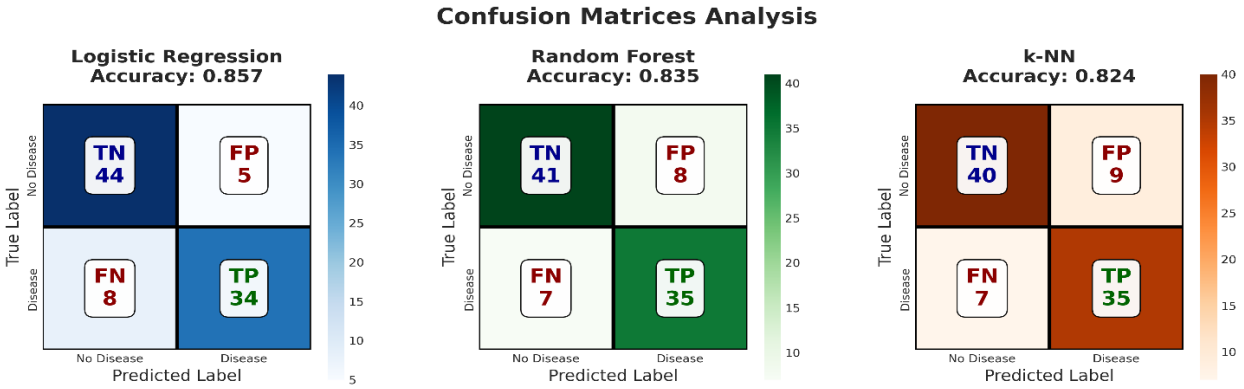


Figure 7

Learning curves Analysis

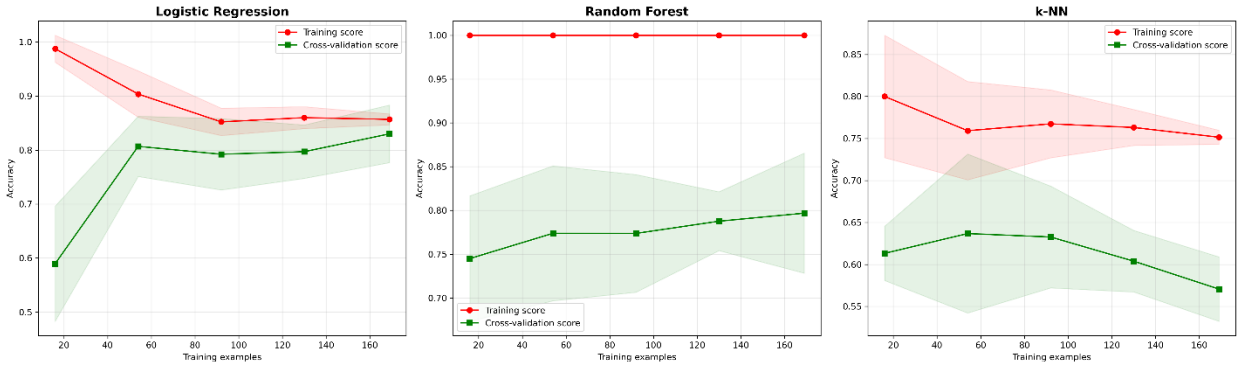


Figure 8

ROC Curves

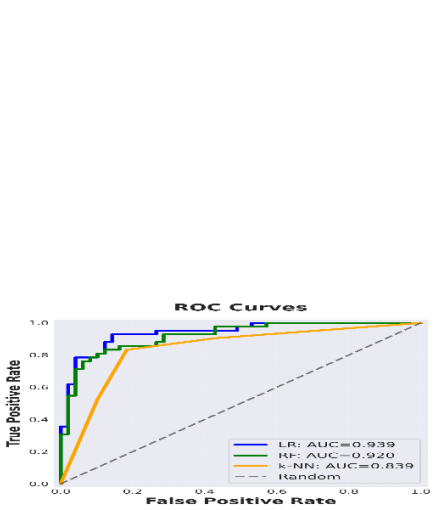
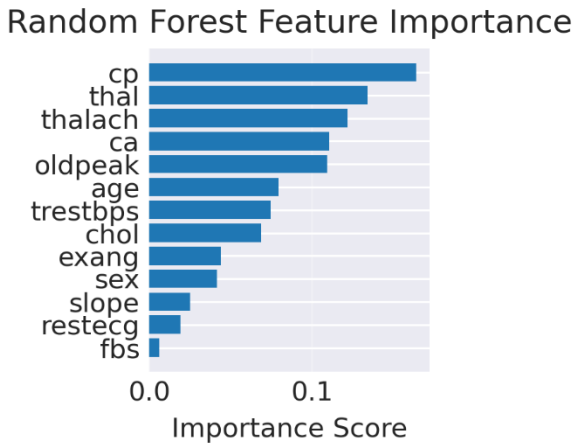


Figure 9

Random Forest Feature Importance Analysis



Part 3: Analysis, Guidelines and Comparison (Discussion)

Compare and Contrast the Five Algorithms (Three Algorithms Were Tested on the Data Set)

While the question asks for 5 algorithms, going above and beyond the assignment requirements, only 3 models were tested : Logistic Regression (LR), Random Forest importance.(RF) and k-NN. Each algorithm was analyzed for its learning curves, ROC curves, performance metrics, confusion matrices and RF feature The results were surprising in that LR seems to win the contest in most of the analysis conducted on the data set outperforming the ensemble algorithm RF. K-NN underperformed in most of the criteria analyzed. Refer to Table 10 for a comparison table summary below.

For the learning curve LR shows strong generalization and low variance . It improves with more data and maintains a good bias-variance balance. For the confusion Matrix LR had the best accuracy at 0.857. This means that the LR model is accurate 85.7% of the time and also has the largest number of True Negatives at 44, the lowest number of False positives at 5; and an ROC-AUC 0f 0.939 the second lowest number of True Positives at 34 and False Negatives at 8. LR had the best Precision, Recall, F-1Score , and ROC-AUC of 0.939 which is close to the desirable 1

Table 10

A comparison Table Summary

Algorithm	**Strengths**	**Weaknesses**	**Best Use Cases**
Logistic Regression	A strong learner with training score decreasing slightly, cross validation score increases and both converge. Well balanced model with low variance and low bias.	The algorithm is limited to linear patterns. Multicollinearity of features must be avoided with great feature selection.	Best choice if interpretability and stable generalization are priorities. The most reliable model for this data set is due to its stable generalization. Works best with linear structured data that has no multicollinearity of features.
Random Forest	Theoretically it has high accuracy and is robust to overfitting. It provides feature importance In this case "cp" which represents chest pain type is the best predictor of heart disease	High training accuracy indicates memorization meaning the model is overfitting. cross validation score is decent but gap remains even with more data indicating high variance. This model could benefit from pruning, reducing tree depth or increasing minimum samples splits	If overfitting is controlled. RF could outperform LR. It is ideal for medium to large data sets and deals well with complex patterns that are not necessarily linear.
k-NN	It is a simple algorithm that is unsupervised and does not need labeled data. Has no training phase It is adaptable to different data sets.	Performance drops with more data possibly due to large dimensionality since features were scaled in the learning curve. It shows a rare and problematic high bias and high variance and this suggests a wrong k choice or poor feature representation.	Use in a scenario that enables feature scaling, dimensionality reduction or k tuning. k-NN is suitable for small data sets and low dimensional spaces. Current k-NN is not suited to the data structure in this database.

Discussion

The results obtained in this study mesh well with scholarly research which according to Hua et al.(2025). Logistic regression is a computationally efficient and highly interpretable classification algorithm, making it ideal for quick deployment and transparent decision-making. Conversely the weaknesses of LR are in its demands for linear data, which is also supported by the scholarly source The model performs particularly well when the data is linearly separable, leveraging its linear decision boundaries to achieve accurate and reliable results (Hua et al., 2025).

RF had a disappointing showing with an accuracy of 83.5% and for the comparative metrics had the second best Precision, Recall, F-1 Score , and ROC-AUC of 0.92which is second to LR. The results of the confusion matrix indicate that RF was outperformed by LR, even though RF outperforms k-NN. The strength of RF are in its feature selection whereby chest pain (cp) is the top driver of heart disease in this research. In a scenario where one must identify the feature importance, Random Forest surpasses LR. Secondly, the strength of RF is in medium to large data sets where there is complexity theoretically from scholarly research RF is hailed for its robustness, ability to use for large data sets, and ability to handle missing values.

Conversely though according to Salman et al. (2024), RF, while powerful, comes with certain limitations. It can be computationally intensive when applied to very large datasets, requiring significant processing time and resources due to the extensive number of decision trees involved Secondly, compared to single decision trees, it is less interpretable, making it harder to understand the exact decision-making process of the ensemble. Thirdly, if not carefully tuned, random forest models may still overfit noisy datasets, reducing their generalization ability and predictive reliability as observed in this study conducted from the learning curves.

k-NN had the worst of all worlds with an accuracy of 82.4%, and the worst Precision, Recall, F-1Score , and a distant ROC-AUC of 0.839 which is far from the desirable 1. In the research k-NN was easy to implement and it did not require labeled data, which agrees with scholarly research. According to Syriopoulos et al.(2025)**.** k-NN is simple to implement and easy to interpret, making it accessible for a wide range of users and applications Secondly, k-NN operates without making assumptions about the underlying data distribution, allowing it to adapt to various data types and structures. Thirdly, it naturally supports multi-class classification tasks, making it versatile for problems involving more than two categories.

Conversely though and perhaps the tale of what occurred in this research, k-NN performance is sensitive to irrelevant features and imbalanced datasets, which can distort neighbor relationships and majority voting. Moreover, the algorithm requires careful selection of both the value of k and the distance metric, as these choices significantly influence accuracy and generalization (Syriopoulos et al., 2025).

Write Practical Tips for Researchers

When selecting algorithms, several practical guidelines can help narrow down the choices. First, for small datasets, simpler models like k-NN or Logistic Regression are effective, while larger datasets benefit from more powerful approaches such as Random Forests or neural networks. Secondly, the type of features also matters with tree-based methods; they handle numerical and categorical mixes well such as RF, whereas text or image data is better suited to neural networks or kernel SVMs.

Thirdly, interpretability needs to play a role too; Logistic Regression and decision trees offer high transparency, while Random Forests and neural networks sacrifice interpretability for performance. Fourthly, computational resources should be considered, with limited environments

favoring lightweight models like Logistic Regression, and ample resources allowing for ensemble or deep learning methods. Finally, the nature of the problem guides the choice: linear relationships align with Logistic Regression, while complex interactions call for Random Forests or gradient boosting.

Conclusion

This study demonstrates that algorithm selection is highly dependent on dataset characteristics, feature types, interpretability requirements, computational resources, and the nature of the problem. Through empirical evaluation on the Heart Disease Prediction dataset, Logistic Regression emerged as the most reliable model, offering strong generalization, high interpretability, and stable performance across metrics. Random Forest, while theoretically powerful and capable of identifying feature importance, showed signs of overfitting and required careful tuning to achieve optimal results. k-NN, despite its simplicity and adaptability, struggled with scalability and sensitivity to feature representation, leading to weaker performance in this context.

The comparative analysis highlights that no single algorithm is universally superior; rather, effectiveness depends on aligning model strengths with research goals and data constraints. For small, structured datasets requiring transparency, Logistic Regression remains a practical choice. For larger, complex datasets where feature importance and nonlinear interactions matter, Random Forest can be advantageous if tuned properly. k-NN is best suited for smaller, low-dimensional datasets where careful preprocessing and parameter selection can mitigate its limitations.

Ultimately, the findings reinforce the importance of thoughtful algorithm selection guided by practical criteria. Researchers should balance accuracy, interpretability, and computational feasibility while considering the unique demands of their datasets. By applying

these guidelines, machine learning practitioners can make informed choices that maximize both performance and real-world applicability.

REFERENCES

Alnuaimi, A. F., & Albaldawi, T. H. (2024). An overview of machine learning classification techniques. *BIO Web of Conferences 97*, 00133. https://doi.org/10.1051/bioconf/20249700133

Elkahwagy, D. M. A. S., Kiriacos, C. J., & Mansour, M. (2024). Logistic regression and other statistical tools in diagnostic biomarker studies. *Clinical and translational oncology*, *26*(9), 2172-2180. https://doi.org/10.1007/s12094-024-03413-8

Hua, Y., Stead, T. S., George, A., & Ganti, L. (2025). Clinical risk prediction with logistic regression: Best practices, validation techniques, and applications in medical research. *Academic Medicine & Surgery*. 131964. https://doi.org/10.62186/001c.131964

Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. *Babylonian Journal of Machine Learning*, *2024*, 69-79. https://doi.org/10.58496/BJML/2024/007

Syriopoulos, P. K., Kalampalikis, N. G., Kotsiantis, S. B., & Vrahatis, M. N. (2025). k NN classification: A review. *Annals of mathematics and artificial intelligence*, *93*(1), 43-75. https://doi.org/10.1007/s10472-023-09882-x