MACHINE LEARNING AND NEURAL NETWORKS (CM3015)

MIDTERM REPORT ON

THE COMPARATIVE ANALYSIS OF DECISION TREE AND K-NEAREST NEIGHBOR ALGORITHMS PERFORMANCE ON TWO DATASETS.

JUNE, 2025.

ABSTRACT

Machine learning algorithm selection is crucial for effective model deployment, therefore comparing different algorithms performance across varying datasets is important. This report presents a comparative analysis of two fundamental machine learning algorithms; K-Nearest Neighbors (KNN) and Decision Tree, on the Iris and Breast Cancer datasets from the SciKit-learn library. The study evaluates their classification performance using key metrics (accuracy, precision, recall, and F1-score) under different hyperparameter settings (KNN: k=3 vs. k=5; Decision Trees: depth=3 vs. depth=5). Results indicate that KNN performs optimally with fewer neighbors (k=3), achieving 95% accuracy on Breast Cancer and 100% on Iris, while Decision Trees benefit from increased depth (depth=5), improving Breast Cancer accuracy from 89% to 91%.

INTRODUCTION

Machine learning algorithm selection is crucial for effective model deployment [1]. Different machine learning algorithms are available and can be used to gain valuable insights, intelligently analyze, and develop the corresponding smart and automated applications from datasets [2]. Different algorithms perform variably across different datasets and different classification tasks. For instance, studies [3],[4] comparing model performance across business intelligence and cybersecurity datasets show that Random Forest outperforms achieving high accuracy rates (up to 96.3% in classification) and precision metrics. Gradient Boosting algorithm provides comparable accuracy levels as random forest [3]. It is however important to note that performance of algorithms vary with varying datasets, thus an algorithm that performs very well on a dataset may do poorly on another due to overfitting [5].

This report highlights the comparative analysis of the performance of K-Nearest Neighbors (KNN) and Decision Tree classifiers in classification tasks, using the Iris and Breast Cancer datasets from the SciKit-learn library.

Dataset Justification

The Iris dataset originates from the seminal work of British statistician and biologist Ronald A. Fisher The measurements were collected by American botanist Edgar Anderson from three Iris species (Iris setosa, Iris versicolor, and Iris virginica) in the Gaspé Peninsula (Canada). Fisher used Anderson's data to demonstrate how quantitative traits (sepal/petal dimensions) could classify species, making it one of the first examples of supervised machine learning [6]. The dataset comprises 150 samples, evenly distributed across three classes: Setosa, Versicolor, and Virginica, with 50 samples per class. Each entry in the dataset includes four numerical attributes measured in centimeters: sepal length, sepal width, petal length, and petal width. These attributes

serve as predictive features for determining the class of the iris plant which specifies the species classification (Setosa, Versicolor, or Virginica) for each sample [7].

The Breast cancer dataset is derived from theUniversity of California, Irvin's Machine Learning Repository and contains features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass [8]. It includes 30 attributes such as radius, texture, perimeter, area, and smoothness, which are crucial for classification tasks.The dataset categorizes tumors into two classes: malignant and benign [8].

The Iris dataset is relatively simple, with 4 numerical features and 3 balanced classes, making it ideal for assessing basic classification performance. The Breast Cancer dataset is more complex, with 30 numerical features and an imbalanced binary classification task (malignant vs. benign), allowing for an evaluation of how the algorithms handle higher dimensionality and class distribution.

Algorithm Justification

K-Nearest Neighbors (KNN) is a fundamental classification approach known for its simplicity and computational efficiency [9]. The method requires only two primary parameters; the number of neighbors (k) and a distance metric, thereby making it straightforward to implement. However, its performance depends significantly on proper data scaling and local data patterns [9].

Decision Trees represent a versatile machine learning technique applicable to both classification and regression tasks [10]. These models operate by recursively partitioning data based on feature values, creating an interpretable tree structure [10]. Decision Trees offer several advantages, including minimal data preprocessing requirements and compatibility with diverse data types

[10]. The algorithm's effectiveness is influenced by factors such as tree depth, feature selection criteria, and dataset characteristics [11].

KNN is an instance-based algorithm that classifies samples based on proximity to neighboring data points, making it highly dependent on feature scaling and local data structure. Decision Trees are rule-based algorithms that recursively partition the feature space using hierarchical decision rules, inherently handling non-linear relationships. By comparing these fundamentally different methodologies, this report provides insights into how local similarity measures (KNN) versus global feature splits (Decision Trees) perform under varying conditions like hyperparameter tuning and different datasets.

METHODOLOGY

This report is a comparative analysis of the performance metrics between K-Nearest Neighbors (KNN) and Decision Tree algorithms in classification tasks using the Iris and Breast Cancer datasets from the SciKit-learn library. The table below provides a summary of the datasets used in this report, detailing their characteristics.

Table 1:

Characteristic	Iris Dataset	Breast Cancer Dataset
Domain	Botany/Taxonomy	Medical Diagnostics
Total Samples	150 (50 per class)	569 (357 benign, 212 malignant)
Number of Classes	3 (Setosa, Versicolor, Virginica)	2 (Benign, Malignant)

Number of Features	4 (all numerical)	30 (all numerical)
Class Balance	Perfectly balanced	Imbalanced (63% benign, 37% malignant)

Model Training

The KNN model was implemented from scratch using python Numpy libraries. The Euclidean distance was calculated using the formula given below:

$\sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$

where:

x_1 = First data point's feature 1 value

x_2 = Second data point's feature 1 value

y_1 = First data point's feature 2 value

y_2 = Second data point's feature 2 value [12]

For the KNN implementation, the Euclidean distance metric was employed to measure similarity between data points. The algorithm was evaluated using two configurations: one with k=3 neighbors and another with k=5 neighbors. These values were selected to assess how the choice of neighborhood size influences classification performance. Decision tree algorithm used in this analysis was imported from the sklearn.tree library. To examine the impact of the decision tree model's complexity, two depth settings were tested: a shallow tree with max_depth=3 and a deeper tree with max_depth=5. Both KNN and Decision tree algorithms were trained on the same stratified training data to maintain consistency in evaluation. Performance metrics were

then computed on the held-out test set to facilitate a direct comparison between the two classification approaches.

Evaluation Metrics

Datasets were divided into training and testing subsets using an 80-20 split ratio. The following performance metrics were assessed on both algorithms:

- Accuracy: (TP+TN)/(TP+TN+FP+FN)
- Precision: TP/(TP+FP)
- Recall: TP/(TP+FN)
- F1-score: Harmonic mean of precision and recall

RESULTS AND ANALYSIS

KNN ALGORITHM

Across both datasets, the KNN algorithm showed better performance metrics when k value =3 than when k value = 5. Figure 1 shows the graphic distribution of the performance with the two tested hyperparameters. The KNN algorithm also demonstrated excellent performance on the Iris dataset compared to the breast cancer, achieving consistent high accuracy with both k values. Figure 2 shows the graphic distribution of KNN performance across both datasets. KNN maintained strong performance on the Breast Cancer dataset, with better performance at k = 3 compared to k=5. Table 2 Summarizes the performance of the KNN algorithm.

Table 2: Table showing KNN algorithm performance metrics

METRIC	K = 3	K = 5
(Accuracy - Breast Cancer)	95%	93%
(Accuracy - Iris)	100%	97%
(Precision - Breast Cancer)	95%	93%
(Precision - Iris)	100%	96%
(Recall - Breast Cancer)	94%	93%
(Recall - Iris)	100%	97%
(F1 - Breast Cancer)	94%	93%
(F1 - Iris)	100%	97%

Figure 1: Graph showing distribution of KNN performance metrics across parameters

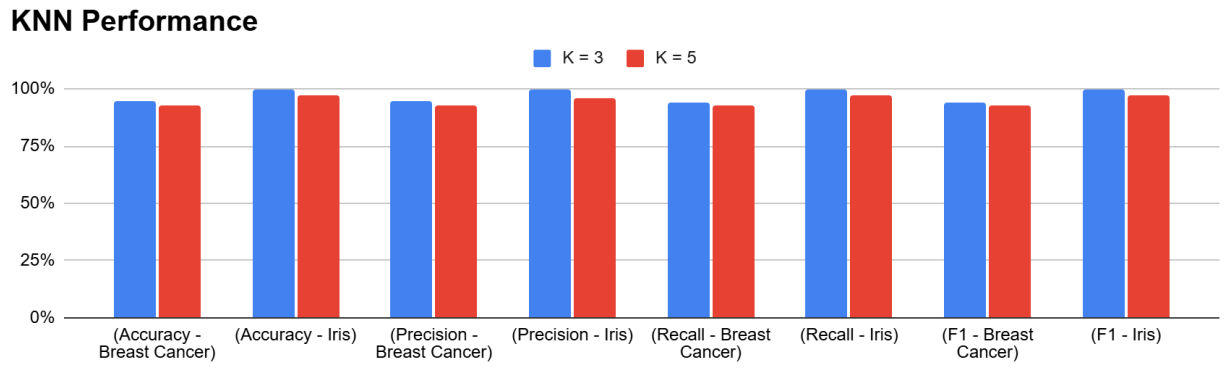
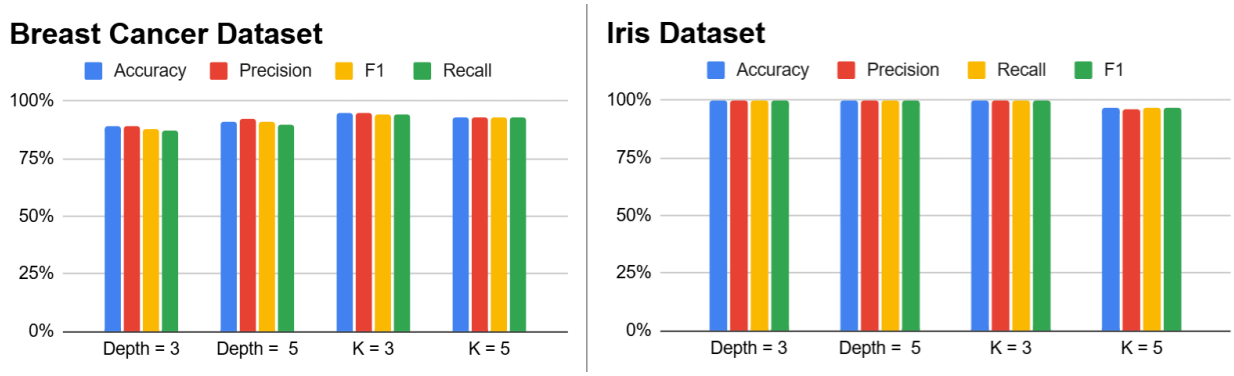


Figure 2: Graph comparing performance metrics of both algorithms across parameters



DECISION TREE ALGORITHM

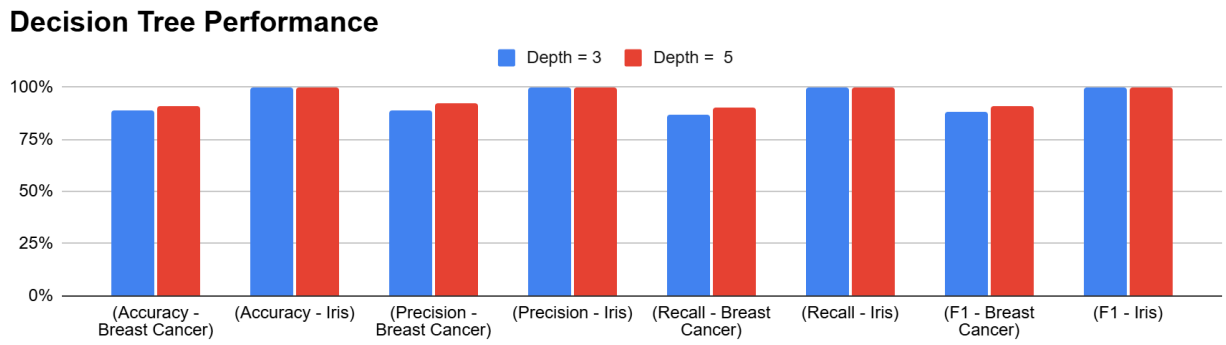
Across both datasets, the decision tree algorithm showed better performance metrics when depth =5 than when depth = 3. Figure 3 shows the graphic distribution of the algorithm's performance with the two tested hyperparameters (Depth = 5 and depth =3). For the Breast Cancer dataset, the deeper tree, depth = 5 showed superior performance compared to the shallower depth, depth = 3. Table 3 Summarizes the performance of the decision tree algorithm. In contrast, the Iris dataset showed perfect classification (100% across all metrics) at both tree depths, indicating the dataset's inherent simplicity and clear decision boundaries.

Table 3: Table showing Decision Tree algorithm performance metrics

METRIC	Depth = 3	Depth = 5
(Accuracy - Breast Cancer)	89%	91%
(Accuracy - Iris)	100%	100%
(Precision - Breast Cancer)	89%	92%
(Precision - Iris)	100%	100%

(Recall - Breast Cancer)	87%	90%
(Recall - Iris)	100%	100%
(F1 - Breast Cancer)	88%	91%
(F1 - Iris)	100%	100%

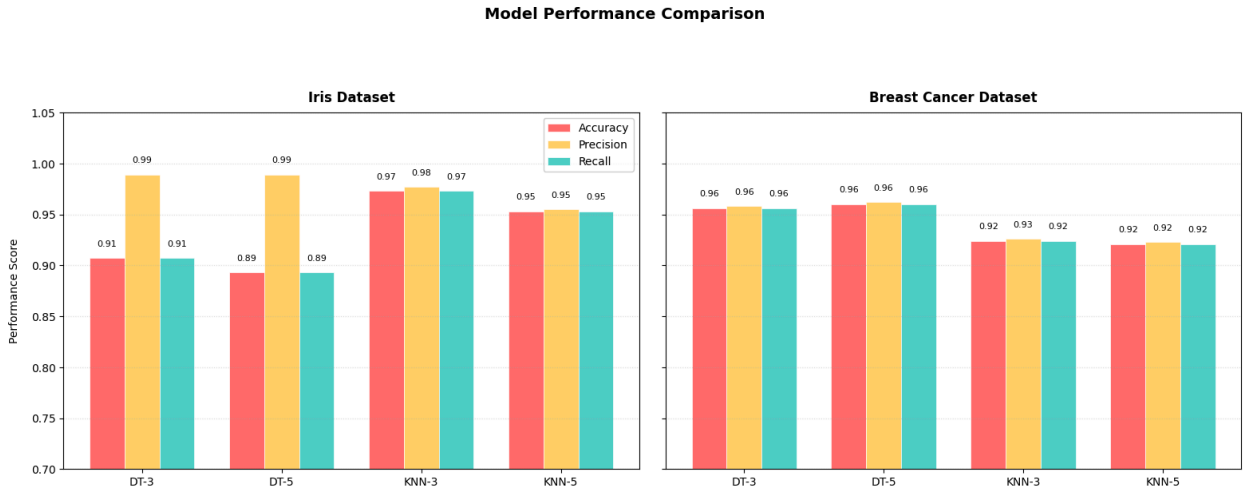
Figure 3: Graph showing distribution of Decision tree performance metrics across parameters



CROSS VALIDATION

The cross-validation results reveal consistent accuracy of decision tree on both datasets (92.6–96.0%) with low standard deviation (±0.02–0.03), indicating stable performance across folds. Depth variation had minimal impact on Iris but slightly better precision at depth=3 for Breast Cancer. KNN achieved higher accuracy on Breast Cancer (95.6–96.0%) compared to Iris (89.3–90.7%), with notably lower variance (±0.02 vs. ±0.09–0.11). Find more details in figure 4 below.

Figure 4: Graph showing Models' cross validation performance



DISCUSSION

The results reveal interesting patterns about how the two algorithms perform under different conditions. For the KNN algorithm, setting the number of neighbors (k) to 3 generally led to better results than using k=5. This was true for both the Breast Cancer and Iris datasets, though the difference was more noticeable with the Iris dataset. With k=3, KNN achieved 95% accuracy on the Breast Cancer dataset and a perfect 100% on the Iris dataset, suggesting it handles simpler datasets like Iris exceptionally well.

On the other hand, the Decision Tree algorithm performed better when allowed to grow deeper (depth=5) rather than being restricted to depth=3. This was particularly clear with the Breast Cancer dataset, where accuracy improved from 89% to 91% with the deeper tree. However, the Iris dataset was so straightforward that the Decision Tree achieved perfect scores regardless of depth, reinforcing that simpler datasets don't always need complex models..

The cross-validation results add another layer of insight. Decision Trees showed stable performance across different data splits, with only small variations in accuracy. KNN, while excellent on the Breast Cancer dataset, was less consistent on the Iris dataset, especially with higher k values. This suggests that KNN may be more sensitive to parameter choices depending on the dataset.

CONCLUSION

This comparative analysis demonstrates that algorithm performance depends heavily on both dataset characteristics and hyperparameter selection. KNN achieves optimal results with fewer neighbors (k=3), excelling particularly on the simpler Iris dataset with perfect classification, while maintaining strong 95% accuracy for Breast Cancer diagnosis. Decision Trees benefit from increased depth (depth=5), showing measurable improvement on the more complex Breast Cancer dataset. Cross-validation confirms Decision Trees offer greater stability across data splits, whereas KNN exhibits higher variance, especially on Iris. These findings highlight that KNN is preferable for straightforward classification tasks, while deeper Decision Trees better handle nuanced patterns in complex data. The results emphasize that effective model deployment requires matching algorithm strengths to dataset properties through careful parameter tuning. Future work could explore hybrid approaches to combine KNN's precision with Decision Trees' robustness.

REFERENCES

[1] Paleyes, A., Urma, R.G. and Lawrence, N.D., 2022. Challenges in deploying machine learning: a survey of case studies. *ACM computing surveys*, *55*(6), pp.1-29.

[2] Sarker, I.H., 2021. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), p.160.

[3] Al Bony, M.N.V., Das, P., Pervin, T., Shak, M.S., Akter, S., Anjum, N., Alam, M., Akter, S. and Rahman, M.K., 2024. Comparative performance analysis of machine learning algorithms for business intelligence: a study on classification and regression models. *Frontline Marketing, Management and Economics Journal*, *4*(11), pp.72-92.

[4] Sayed, M.A., Sarker, M.S.U., Al Mamun, A., Nabi, N., Mahmud, F., Alam, M.K., Hasan, M.T., Buiya, M.R. and Choudhury, M.Z.M.E., 2024. Comparative analysis of machine learning algorithms for predicting cybersecurity attack success: a performance evaluation. *The American Journal of Engineering and Technology*, *6*(09), pp.81-91.

[5] Kwon, O. and Sim, J.M., 2013. Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications*, *40*(5), pp.1847-1857.

[6] Mithy, S.A., Hossain, S., Akter, S., Honey, U. and Sogir, S.B., 2022. Classification of iris flower dataset using different algorithms. *International Journal of Scientific Research in Mathematical and Statistical Sciences.* 9(6), pp. 1-10.

[7] Thirunavukkarasu, K., Singh, A.S., Rai, P. and Gupta, S., 2018, December. Classification of IRIS dataset using classification based KNN algorithm in supervised learning. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)* (pp. 1-4). IEEE.

[8] Hemapriya, B., Devi, K. and Harini, K., 2023, January. Automatic Scikit-learn based detection and classification of Breast Cancer using Machine./Learning techniques. In *2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)* (pp. 1-8). IEEE.

[9] Moldagulova, A. and Sulaiman, R.B., 2017, May. Using KNN algorithm for classification of textual documents. In *2017 8th international conference on information technology (ICIT)* (pp. 665-671). IEEE.

[10] Dehghani, A.A., Movahedi, N., Ghorbani, K. and Eslamian, S., 2023. Decision tree algorithms. In *Handbook of hydroinformatics* (pp. 171-187). Elsevier.

[11] Ordyniak, S. and Szeider, S., 2021, May. Parameterized complexity of small decision tree learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 7, pp. 6454-6462).

[12] Abba, I. (2023) KNN algorithm – K-nearest neighbors classifiers and model example, Available at: *https://www.freecodecamp.org/news/k-nearest-neighbors-algorithm-classifiers-and-model-example/* (Accessed: 17 June 2025).