**TIRAMISU RESEARCH**
Tiramisu Research Journal vol. 1, Jan 2025

# IRIS FLOWER SPECIES CLASSIFICATION USING K-NEAREST NEIGHBOURS

Moses Sinanta P.W.J.[1]
[1]Spring Garden Research Division

**Abstract**

This study explores the use of the K-Nearest Neighbours (KNN) algorithm to classify Iris species from the Iris dataset. The dataset includes features such as sepal length, sepal width, petal length and petal width, with the goal of predicting one of the three species: *Setosa*, *Versicolor* or *Virginica*. The study investigates the effect of different values for the number of neighbours ($k$), ranging from 1 to 5, on classification accuracy. The results show high performance across all tested values of $k$, with the best accuracy achieved at $k = 3$ and $k = 5$. The research demonstrates the effectiveness of KNN in species classification and highlights the importance of selecting an appropriate $k$ value. This study is conducted for educational purposes only, aiming to provide insight into machine learning classification algorithms.

**Key words**: *K-Nearest Neighbours*, *Iris dataset*, *classification*, *machine learning*.

## INTRODUCTION

The classification of Iris species has long been a standard benchmark problem in machine learning. The Iris dataset consists of 150 samples from three different species of Iris flowers. By analysing the flowers' features, such as its petal and sepal dimensions, the goal is to accurately classify the species.

Machine learning has proven to be a powerful tool for classification tasks, with various algorithms designed to categorize data into predefined classes. One popular method for classification is the $k$-Nearest Neighbours (KNN), which works by finding the most similar data points to make predictions. KNN is particularly appealing due to its simplicity and effectiveness, especially for problems where the decision boundaries are not easily defined.

The goal of this study is to classify Iris species using the KNN algorithm, focusing on determining the best hyperparameter: the number of neighbours ($k$). This study aims to explore how different values of $k$ impact model performance and accuracy. This study is conducted for educational purposes only, aiming to demonstrate the practical application of KNN in a real-world dataset.

## THEORETICAL BASIS

**Figure 1** The three Iris species: *Setosa* (top left), *Versicolor* (top right) and *Virginica* (bottom).

The Iris dataset, introduced by Ronald A. Fisher [10] is a widely used dataset in machine learning. It contains 150 samples from three species of Iris flowers–*Setosa*, *Versicolor* and *Virginica*–each with four features: sepal length, sepal width, petal length and petal width. This dataset is often used to demonstrate and benchmark classification algorithms due to its simplicity and well-defined class labels [8] [5].

Machine learning is a field of artificial intelligence focused on developing algorithms that allow computers to learn from and make predictions on data [3]. It involves training a model on labelled data and then evaluating its performance on unseen data. There are several types of machine learning algorithms including supervised learning, where the model is trained with input-output pairs [6], and unsupervised learning, where the model finds patterns without labelled outcomes [4].

$k$-Nearest Neighbours (KNN) [9] is a simple, non-parametric algorithm used for classification and regression tasks. It works by identifying the $k$ closest training samples in the feature space and predicting the class label based on the majority vote of these neighbours. KNN's performance is sensitive to the choice of $k$ and the distribution of data, making it important to carefully tune its parameters for optimal results [1].

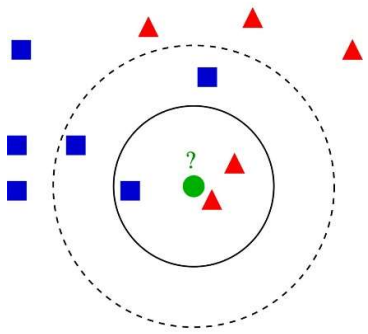


**Figure 2** Visualization of KNN algorithm [2].

Previous works using the Iris dataset have explored a variety of machine learning algorithms [5], with many studies focusing on comparing different classification algorithms. KNN has been widely tested on the Iris dataset due to its simplicity and effectiveness, often serving as a baseline model. Studies have also investigated the impact of different hyperparameters, such as the number of neighbours [7], and explored other methods for improving KNN's performance [1].

## METHODOLOGY

The Iris dataset used in this study is saved in CSV format, with each row representing a single flower sample and each column corresponding to one of the four features: sepal length, sepal width, petal length and petal width. Additionally, the dataset includes a label column specifying the species of each sample (*Setosa*, *Versicolor* or *Virginica*). The dataset consists of 150 instances, with 50 samples from each species, providing a balanced class distributions suitable for classification.

To prepare the data for modelling, we apply feature scaling using `StandardScaler` from the scikit-learn library in Python. This scaling standardizes the features by removing the mean and scaling them to unit variance, ensuring that each feature contributes equally to the distance calculation in the KNN algorithm. The dataset is then split into training and testing sets using the 75-25 ratio, where 75% of the data is used for training the model and 25% is reserved for evaluating its performance.

For the KNN algorithm, we will explore the effect of different values of the hyperparameter $k$, ranging $[1,10]$, to determine the optimal number of neighbours for classification. Model performance will be evaluated using several metrics: accuracy, precision, recall and F1 score. These metrics will be computed for each value of $k$ to assess the model's robustness and identify the best-performing $k$ value configuration.

## RESULTS EVALUATION

The overall results from the KNN classification on the Iris dataset shows good performance across all values of $k$ tested. For $k = 1$ and $k = 2$, the accuracy is consistently high at around 97%, with other classification metrics close to 0.97 , indicating strong classification abilities. Notably, for $k = 3$ and $k = 5$, the performance improves to a perfect score of 100%, indicating that KNN perfectly identified all the instances in the testing set.

However, the model performed slightly worse for $k = 4$ compared to its neighbours. Although the accuracy remained at around 97%, the precision and recall were slightly lower, compared to the perfect classification achieved with $k = 3$ and $k = 5$. This drop in performance suggests that the value $k = 3$ and $k = 5$ might better capture the optimal decision

**Table 1** KNN classification result on the Iris dataset across different values of $k$.

| $k$ value | Accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| 1 | 97.37% | 0.9757 | 0.9737 | 0.9736 |
| 2 | 97.37% | 0.9759 | 0.9737 | 0.9737 |
| 3 | 100.00% | 1.0000 | 1.0000 | 1.0000 |
| 4 | 97.37% | 0.9757 | 0.9737 | 0.9737 |
| 5 | 100.00% | 1.0000 | 1.0000 | 1.0000 |

boundaries for this dataset, whereas $k = 4$ might have introduced some noise due to the balancing of neighbours in the decision process.

Aside from the performance dip at $k = 4$, the remaining results align closely with the optimal outocmes. Both $k = 3$ and $k = 5$ resulted in flawless classification, achieving 100% accuracy. This indicates that, within the observed range of $k$, KNN's performance was highly sensitive to the choice of $k$. Overall, the results suggest that KNN is a strong classifier for the Iris dataset, with the best performance observed at $k = 3$ and $k = 5$.

## CONCLUSION

In this study, we applied the K-Nearest Neighbour (KNN) algorithm to classify the Iris dataset, exploring different values of the hyper-parameter $k$ ranging [1,5]. The results revealed that KNN performed exceptionally well across all values of $k$, with perfect accuracy achieved for $k = 3$ and $k = 5$. However, a slight decrease in performance was observed for $k = 4$, where the classification metrics were lower compared to the other values of $k$. Overall, the best performance was obtained with $k = 3$ and $k = 5$, indicating that these values of $k$ are optimal for this classification task.

For future work, further exploration can be made by testing a wider range of $k$ values beyond $k = 5$, including higher values, and experimenting with different distance metrics, such as Manhattan or Mikowski distance, to determine if they yield better performance. Additionally, cross-validation techniques could be employed to ensure that the results are robust and not subject to the particular train-test split used in this study. Further investigation could also involve using advance feature selection methods or combining KNN with other algorithms in an ensemble model to improve classification accuracy.


## ACKNOWLEDGEMENTS

I would like to express my gratitude to my daughter, Deniz Özorman for her help in translating some reading materials in this study.


## CITATION FORMAT

Sinanta, Moses. (2025). "Iris Flower Species Classification Using K-Nearest Neighbours." *Tiramisu Research Journal*, 1, 1-4.

## GITHUB REPOSITORY

https://github.com/MosesSinanta/classification-iris-knn.

## REFERENCES


[1] Halder, Rajib Kumar, et al. (2024). "Enhancing K-Nearest Neighbor Algorithm: A Comprehensive Review and Performance Analysis of Modifications." *Journal of Big Data*, 11(1), 113.

[2] Wikipedia. (2024). "K-Nearest Neighbors Algorithm." Accessed on 25 January 2025 from en.wikipedia.org.

[3] Sharifani, Koosha and Mahyar Aimini. (2023). "Machine Learning and Deep Learning: A Review of Methods and Applications." *World Information Technology and Engineering Journal*, 10(7), 3897-3904.

[4] Naeem, Samreen, et al. (2023). "An Unsupervised Machine Learning Algorithms: Comprehensive Review."

*International Journal of Computing and Digital Systems*, 13(1), 911-921.

[5] Mithy, S.A., et al. (2022). "Classification of Iris Flower Dataset Using Different Algorithms." *International Journal of Scientific Research in Mathematical and Statistical Sciences*, 9(6), 1-10.

[6] Suyal, Manish and Parul Goyal. (2022). "A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms Based on Supervised Learning." *International Journal of Engineering Trends and Technology*, 70(7), 43-48.

[7] Çelik, Ahmet. (2022). "Improving Iris Dataset Classification Prediction Achievement by Using Optimum *k* Value of KNN Algorithm." *Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi*, 3(2), 23-30.

[8] Rao, T. Srinivas, et al. (2021). "Iris Flower Classification Using Machine Learning". *Network*, 9(6), 2082-2090.

[9] Cover, Thomas and Peter Hart. (1967). "Nearest Neighbour Pattern Classification." *IEEE Transactions on Information Theory*, 13(1), 21-27.

[10] Fisher, Ronald A. (1936). "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics*, 7(2), 179-188.