k-NN algorithm in automated Brest Cancer detection

Nicholas Stephen

04/04/2025

1. Introduction

The k-nearest neighbours (k-NN) algorithm is a supervised machine learning algorithm used in classification tasks. The principle is based on comparing an uncategorised datapoint to similar datapoints which are categorised and providing classification to the uncategorised datapoint based on similarities to the classified points. There are many advantages of k-NN including its simplicity, it is very quick to train and does not make any assumptions about the dataset [1]. The k-NN algorithm can be used in many applications, with one being the identification of cancerous (malignant) or noncancerous (benign) breast tumours [2]. Automation of this task could yield many benefits including faster processing, increase capacity to test more tumours at a given time and freeing human resources to allow doctors more time to focus on patient care. This mini project will investigate the k-NN algorithm in breast cancer detection, aiming to identify the optimal parameters used for k-NN and evaluate the performance of the k-NN algorithm. Section 2 will detail the method, with the results presented in Section 3 and the conclusions outlined in Section 4.

2. Method

This project was conducted using Python. The dataset, used in research by Mangasarian *et al.* [2], consisted of 569 breast tumour biopsies that were either malignant or benign with over 30 specific cell measurements in each of the biopsies. After cleaning and preprocessing the data set, the full data was split into a 70:30 ratio of for algorithm training/testing, maintaining a similar proportion of malignant and benign cases. The k-NN was applied using the **KNeighborsClassifier** function with the **GridSearchCV** function used to find the optimal parameter combination with the parameters considered listed in Table 1. The choice on the number of neighbours **(k)** was setup to consider only odd numbers and the square root of the test dataset size being the highest possible **k** value [1]. Details on the **algorithm** and **weights** definitions in the **KNeighborsClassifier** function can be found elsewhere [3].

k	Weights	Algorithm
3, 5, …, 17, 19	'uniform', 'distance'	'auto','ball_tree', 'kd_tree', 'brute'

Table 1: List of parameters considered in the KNeighborsClassifier function.

After determination of the optimal parameters, these were applied to the **KNeighborsClassifier** function on the test dataset. The predictions can be examined using True Positive, True Negative, False Negative or False Positive results [4]. In this project, a true positive prediction is when the algorithm correctly identifies a malignant tumour, while a true negative prediction is correct identification of a benign tumour. A false negative prediction incorrectly identifies a malignant tumour as benign, and a false positive prediction incorrectly identifies a benign tumour as malignant. The model performance on both datasets was evaluated using the following metrics [4]

1: Accuracy - The sum of correctly identified cases divided by the total number of cases.

2: Sensitivity - The number of true positives divided by the sum of true positives and false negatives.

3: Specificity - The number of true negatives divided by the sum of true negatives and false positives.

3. Results

It was found that the parameters which gave the highest accuracy (99.0%), sensitivity (97.3%), and specificity (100.0%) were **k = 3**, **weights = uniform**, and **algorithm = auto**. Given that there are more benign tumours than malignant tumours, it is reasonable to have a lower value of **k**, as a higher **k** would bias the predictions towards benign. Also it is well established that class imbalance reduces the accuracy of the k-NN algorithm [1] which highlights the importance of having an equal distribution of categories when using the k-NN algorithm.

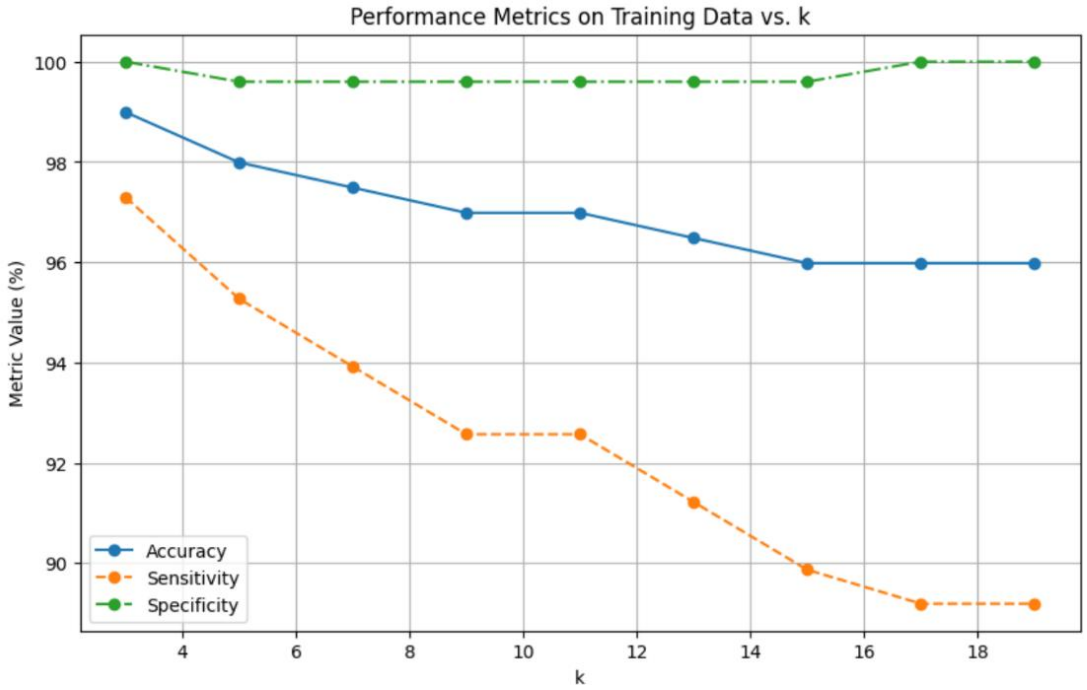

Figure 1: Plot of k-NN algorithm accuracy (blue line), sensitivity (orange line) and specificity (green line) in percentage as a function of **k** on the training dataset. The fixed parameters were **weights = uniform** and **algorithm = auto**.

Figure 1 outlines the accuracy, sensitivity and specificity for different values of **k** on the training data set, showing the accuracy, sensitivity and specificity decrease as **k** increases. Furthermore, the sensitivity is lower compared to accuracy and specificity, irrespective of **k**, implying that the algorithm may be mainly predicting false negatives as errors , *i.e.*, classifying malignant tumour as benign. In this context, this is dangerous for patient health as the tumours could risk spreading across the body.

Figure 2 illustrates the prediction results in a confusion matrix heatmap and shows that 106 benign tumours and 56 malignant tumours have been correctly identified. However, there were in total 9 tumours incorrectly identified with 8 of these being false negative. Using Fig 2, the accuracy (94.7%), sensitivity (87.5%) and specificity (99.1%) was determined.

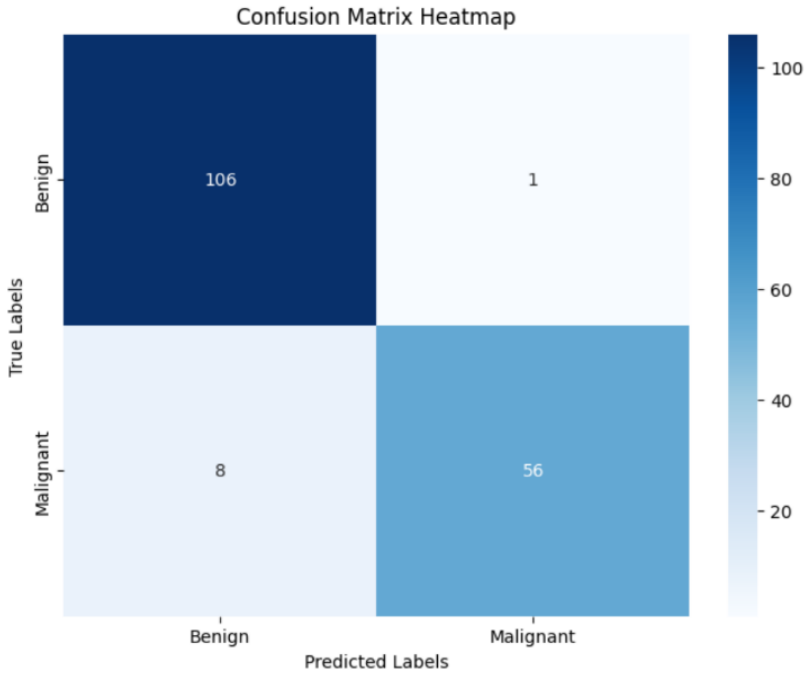

Figure 2: Confusion matrix heatmap detailing true negative (top left square), false positive (top right square), false negative (bottom left square), and true positive (bottom right square) results of the k-NN algorithm applied to the test dataset. Parameters used were **k = 3, weights = uniform** and **algorithm = auto**.

4. Conclusions

In terms of accuracy, selectivity and specificity, the training set's optimal value of **k** was found to be 3. Furthermore, applying these parameters led to correct classification for over 94% of the tumours in the test dataset. However, concerns about the number of false negative cases indicate that additional considerations and refinements would be required. Overall, this project has provided introductory insight into how the k-NN algorithm can be used in classification tasks such as identifying malignant and benign tumours.

5. References

[1] B. Lantz, *Machine Learning with R: Expert techniques for predictive modeling*, 3rd ed. Packt Publishing Ltd, 2019.

[2] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, 'Breast Cancer Diagnosis and Prognosis Via Linear Programming', *Oper. Res.*, vol. 43, no. 4, pp. 570–577, Aug. 1995, doi: 10.1287/opre.43.4.570.

[3] 'KNeighborsClassifier', scikit-learn. Accessed: Apr. 01, 2025. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[4] S. Walters, M. Campbell, and D. Machin, *Medical Statistics: A Textbook for the Health Sciences, 5th Edition*, 5th ed. Wiley-Blackwell, 2021. Accessed: Apr. 01, 2025. [Online]. Available: https://www.wiley.com/en-us/Medical+Statistics%3A+A+Textbook+for+the+Health+Sciences%2C+5th+Edition-p-9781119423645