# K-Nearest Neighbor (K-NN) Analysis Report

Zihan Yan

## Objective

The objective of this assignment is to gain familiarity with applying the K-Nearest Neighbor (K-NN) algorithm on a dataset and explaining the results.

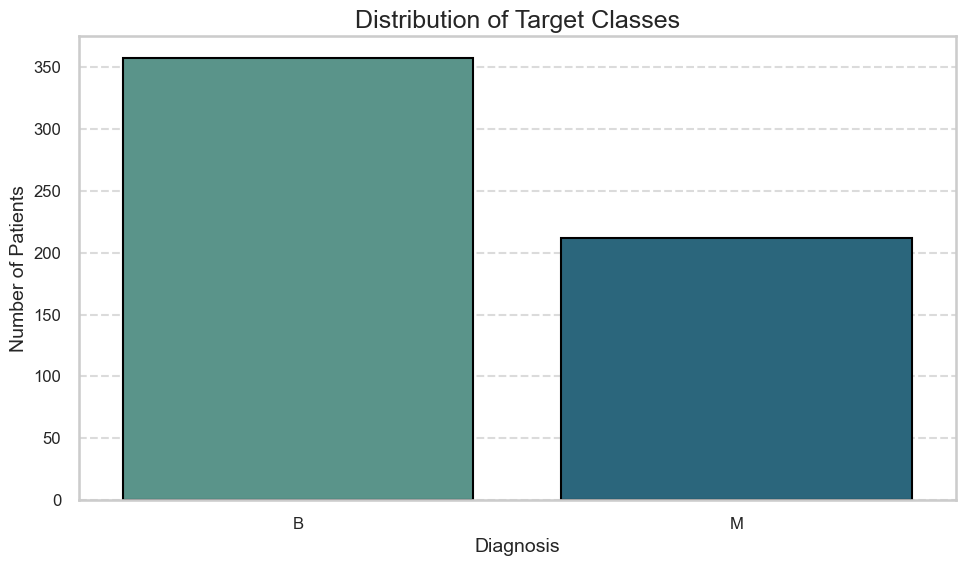
## Dataset Overview

The dataset used for this assignment is the Breast Cancer dataset (KNNAlgorithmDataset.csv). The target variable is the diagnosis column, which has two categories:

* M = Malignant (cancerous)
* B = Benign (non-cancerous)

## Data Dimensionality

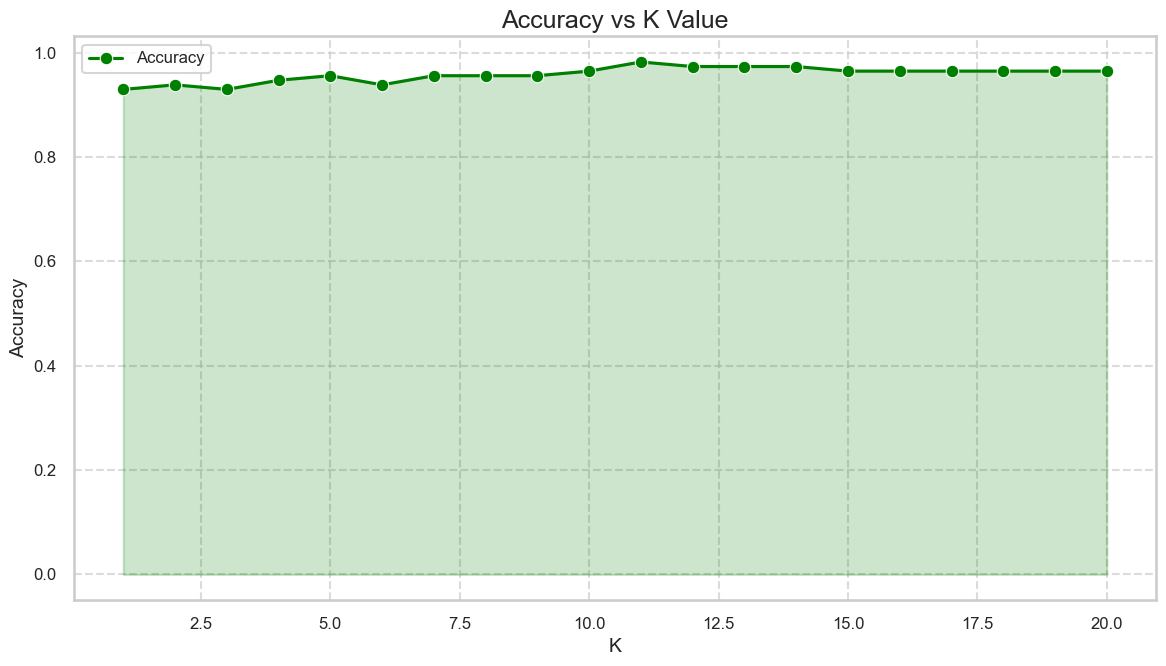
The first step in the analysis is to determine the dimensionality of the dataset. This includes calculating the total number of patients, the number of attributes (categories), data types, any missing values, and the number of patients in each target class.

1. Total number of patients: 569
2. Number of attributes: 33 (including the target variable 'diagnosis')
3. Data types: Mostly numerical (float64) and one categorical (object) for 'diagnosis'.
   1. id int64
   2. diagnosis object
   3. radius\_mean float64
   4. texture\_mean float64
   5. perimeter\_mean float64
   6. area\_mean float64
   7. smoothness\_mean float64
   8. compactness\_mean float64
   9. concavity\_mean float64
   10. concave points\_mean float64
   11. symmetry\_mean float64
   12. fractal\_dimension\_mean float64
   13. radius\_se float64
   14. texture\_se float64
   15. perimeter\_se float64
   16. area\_se float64
   17. smoothness\_se float64
   18. compactness\_se float64
   19. concavity\_se float64
   20. concave points\_se float64
   21. symmetry\_se float64
   22. fractal\_dimension\_se float64
4. Missing values: Only the 'Unnamed: 32' column contains all missing values.
5. Number of patients in each target class:
   1. 357 patients are diagnosed as Benign (B)
   2. 212 patients are diagnosed as Malignant (M)

## Determine the optimal K value

To determine the optimal K value for the K-Nearest Neighbor (K-NN) algorithm, two methods are used:

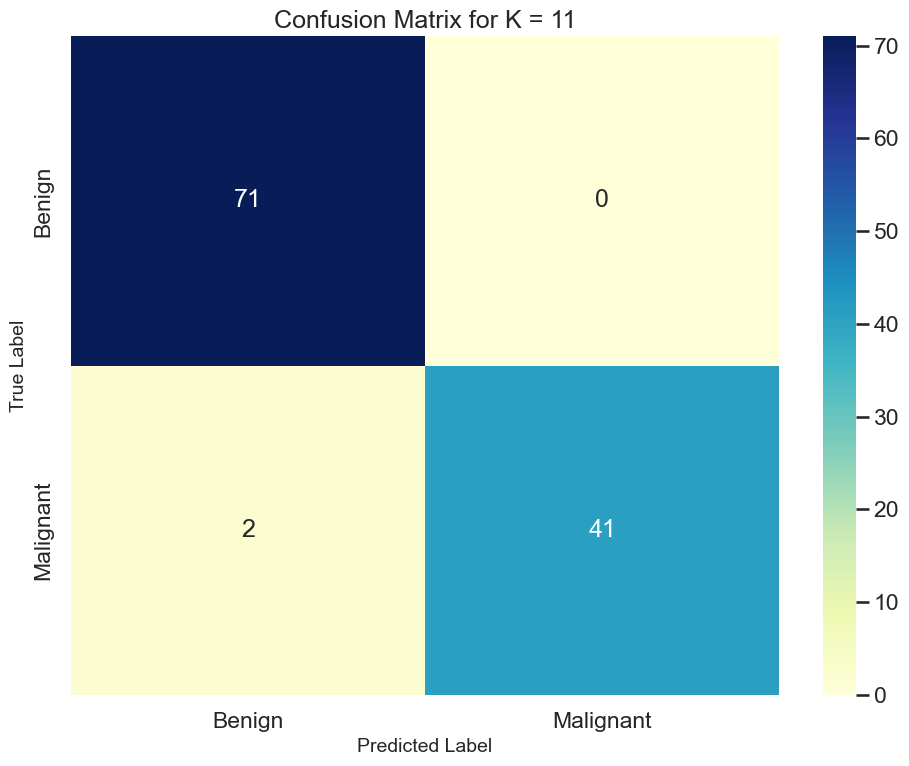
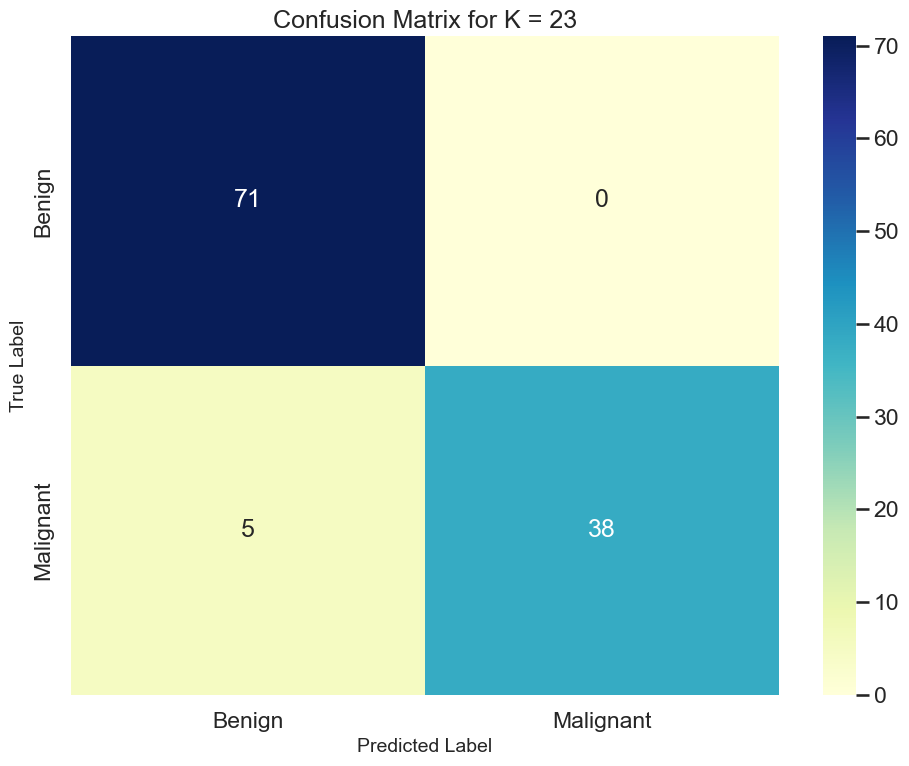
1. Square Root of the Sample Size: This method suggests using K= \sqrt{569} ≈23.
2. Error Rate Plot: The optimal K value from the error plot was found to be K = 11.
   1. **Error Rate Plot**: The plot indicated that the error rate reached its minimum at K=11, confirming this as the optimal K value.
   2. **Accuracy Comparison Plot**: A bar plot comparing accuracies showed a advantage for K=11.

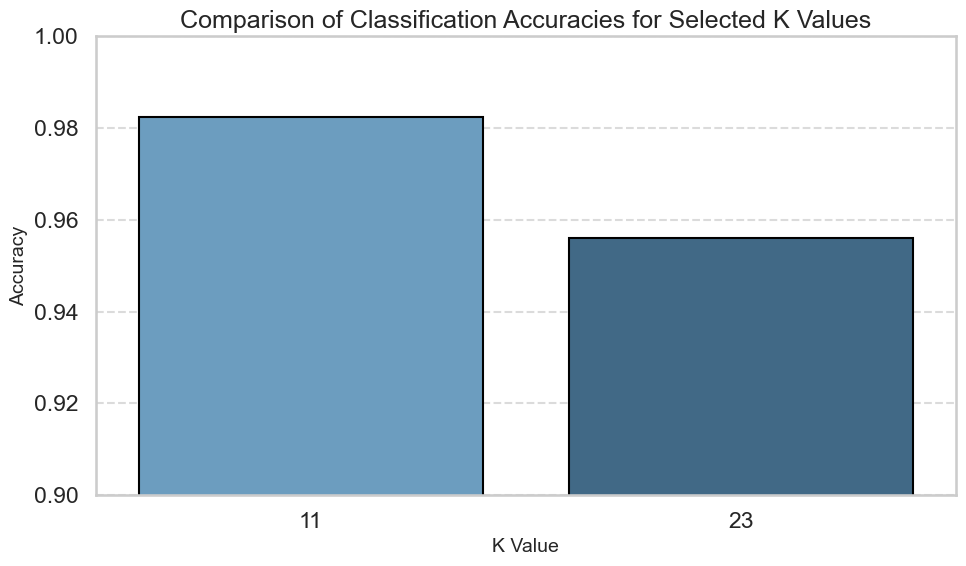


**Apply K-NN using the K values**

1. Comparison of classification accuracies among the K values.

* For K=23: The classification accuracy was **95.61%**.
* For K=11: The classification accuracy was **98.25%**.
* This indicates that the smaller K value (K = 11) performed better in this scenario, achieving a higher accuracy in classifying patients correctly.





1. Comparison of the number of patients misclassified for each target class (M, B) for each K value.
   1. For K=23: 5 benign patients were misclassified as malignant, and no malignant patients were misclassified.
   2. For K=11: Only 2 benign patients were misclassified as malignant, with no malignant misclassifications.
   3. This comparison demonstrates that the K value derived from the error rate plot (K = 11) not only achieved higher accuracy but also reduced the number of misclassified benign patients.

