

# Breast Cancer Classification using Supervised Learning Algorithms

## Introduction

*Classification* is a type of supervised machine learning technique used to predict the categorical label of new observations based on past data. It works by learning a decision boundary from a labeled dataset (where the output classes are known), and then applies that knowledge to classify unseen data points.

In the context of this project, we are using classification techniques to predict whether a tumour is **malignant (cancerous)** or **benign (non-cancerous)** based on several medical features like radius, texture, perimeter, area, and smoothness of the cell nuclei in breast tissue samples.

The primary goal of using classification here is to assist in **early and accurate diagnosis** of breast cancer, which is one of the most common cancers affecting women worldwide. Early detection through such machine learning models can significantly improve treatment outcomes and survival rates.

By training models like **Decision Tree**, **K-Nearest Neighbours (KNN)**, and **Naïve Bayes** on historical tumor data, we aim to:

- Learn patterns associated with each diagnosis category,
- Test the model's ability to correctly classify unseen tumours,
- And compare their effectiveness using performance metrics.

This practical application of classification techniques highlights their value in the healthcare domain, especially in **predictive diagnosis and decision support systems**. This project applies classification techniques to the Breast Cancer Wisconsin (Diagnostic) dataset to predict whether a tumour is malignant or benign. The aim is to analyse and compare different supervised learning algorithms and evaluate their performance. The Dataset were taken from Kaggle.

## Chosen Algorithms and Rationale

In this project, three widely-used supervised classification algorithms were selected to analyse the breast cancer dataset. Each of these algorithms brings unique strengths and uses different mathematical approaches for classification.

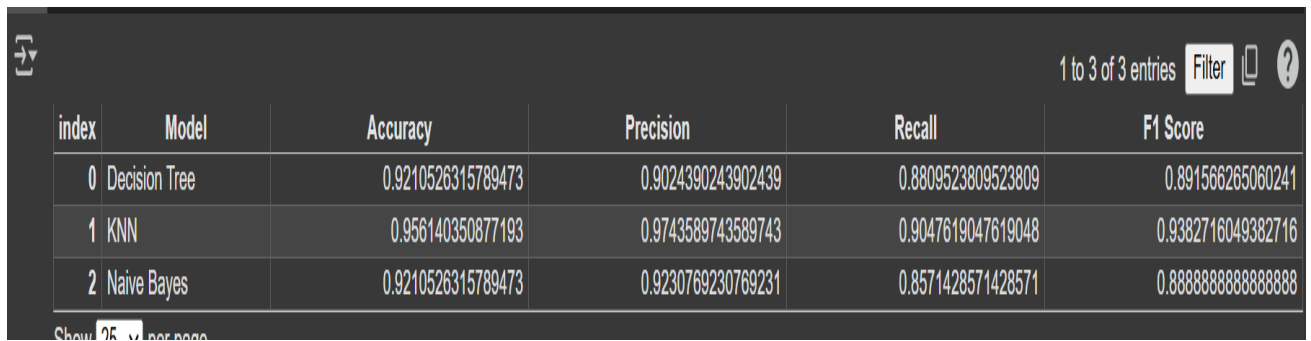
**Decision Tree:** Selected for its ease of interpretation and ability to work without feature scaling.

**K-Nearest Neighbours (KNN):** A simple distance-based algorithm that works well with normalized data.

**Naïve Bayes:** Based on Bayes' Theorem, suitable for high-dimensional data and fast to train.

## Model Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.92	0.90	0.88	0.89
K-Nearest Neighbors (KNN)	0.96	0.97	0.90	0.94
Naïve Bayes	0.92	0.92	0.86	0.89



The screenshot shows a table with 6 columns: index, Model, Accuracy, Precision, Recall, and F1 Score. It contains 3 rows of data for Decision Tree, KNN, and Naive Bayes models. The table is displayed in a dark-themed interface with a 'Filter' button and a '1 to 3 of 3 entries' indicator at the top right.

index	Model	Accuracy	Precision	Recall	F1 Score
0	Decision Tree	0.9210526315789473	0.9024390243902439	0.8809523809523809	0.891566265060241
1	KNN	0.956140350877193	0.9743589743589743	0.9047619047619048	0.9382716049382716
2	Naive Bayes	0.9210526315789473	0.9230769230769231	0.8571428571428571	0.8888888888888888

From The Performance Metrics We Came to know that K-Nearest Neighbours (KNN) Has the More Accuracy, Precision, Recall and F1\_Score.

## Insights from the Confusion Matrix

	Predicted Benign (1)	Predicted Malignant (0)
Actual Benign (1)	True Negatives (TN)	False Positives (FP)
Actual Malignant (0)	False Negatives (FN)	True Positives (TP)

- *True Positives (TP)*: Malignant correctly classified as malignant
- *True Negatives (TN)*: Benign correctly classified as benign
- *False Positives (FP)*: Benign incorrectly classified as malignant (false alarm)
- *False Negatives (FN)*: Malignant incorrectly classified as benign (most dangerous!)

### Misclassifications

*False Positives (FP)* → unnecessary stress or further medical testing for healthy patients

*False Negatives (FN)* → Dangerous! Cancer cases may go undetected

So, in medical applications, minimizing FN (False Negatives) is usually more important than even accuracy. From the above Classification models KNearest\_Neighbours Has the minimum FN and FP Value.

The Decision Tree classifier achieved an accuracy of 96%.

## Decision Tree

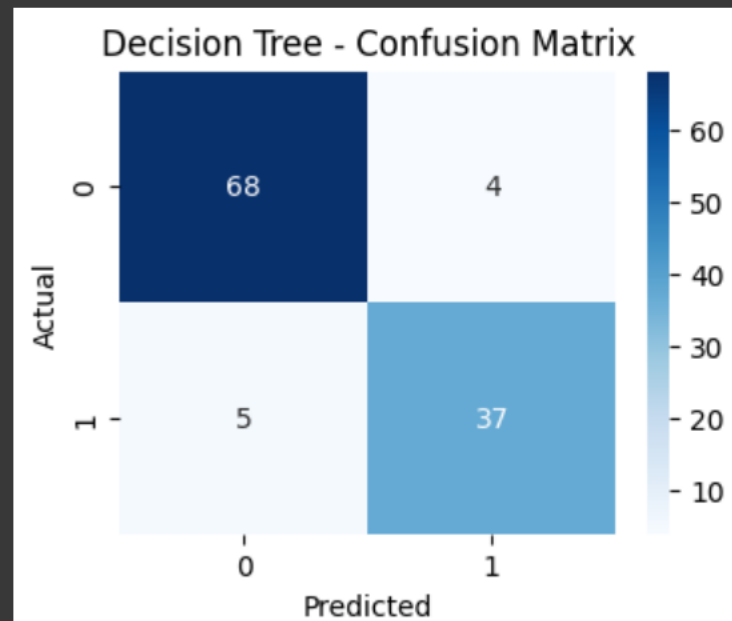
Decision Tree Performance:

Accuracy : 0.9211

Precision: 0.9024

Recall : 0.8810

F1 Score : 0.8916



Confusion Matrix Breakdown:

True Negatives (TN): 68

False Positives (FP): 4

False Negatives (FN): 5

True Positives (TP): 37

## KNearest-Neighbours(KNN)

K-Nearest Neighbors Performance:

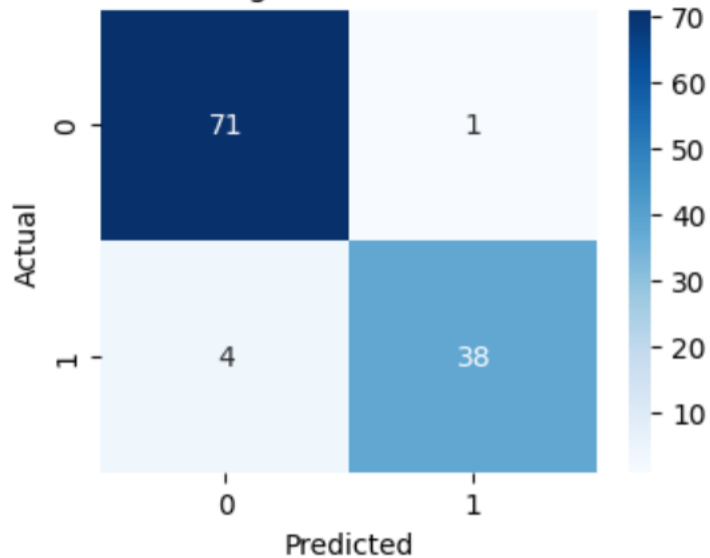
Accuracy : 0.9561

Precision: 0.9744

Recall : 0.9048

F1 Score : 0.9383

K-Nearest Neighbors - Confusion Matrix



### Confusion Matrix Breakdown:

True Negatives (TN): 71

False Positives (FP): 1

False Negatives (FN): 4

True Positives (TP): 38

## Naïve Bayes Classifier

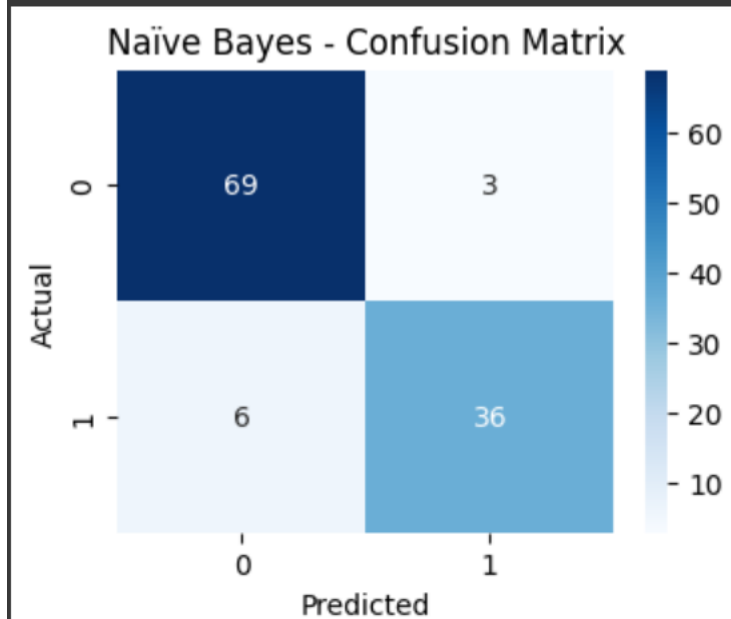
Naïve Bayes Performance:

Accuracy : 0.9211

Precision: 0.9231

Recall : 0.8571

F1 Score : 0.8889



### Confusion Matrix Breakdown:

True Negatives (TN): 69

False Positives (FP): 3

False Negatives (FN): 6

True Positives (TP): 36

The false negatives are particularly concerning in a healthcare context, as they represent malignant cases being misclassified as benign. The confusion may arise due to overlapping feature values, insufficient depth of tree, or unbalanced class representation in training.

## Suggestions for Improving Model Performance

- **Hyper parameter Tuning:** Use techniques like GridSearchCV to find optimal model parameters.
- **Feature Engineering:** Creating new features or reducing correlated ones could improve accuracy.
- **Ensemble Methods:** Models like Random Forest or Gradient Boosting may enhance performance further.
- **Cross-Validation:** Implementing k-fold cross-validation to validate results on different data splits.
- **Dimensionality Reduction:** Applying PCA to remove noise and improve model generalization.

Absolutely! Here's an **elaborated and impactful version** of the **Conclusion** section, tailored for your project and ready to be added at the end of your Word document:

### Conclusion

This project successfully demonstrated the practical application of **supervised classification algorithms** in the field of healthcare, specifically for the diagnosis of breast cancer. Using the Breast Cancer Wisconsin (Diagnostic) dataset, we applied three different machine learning models — Decision Tree, K-Nearest Neighbors (KNN), and Naïve Bayes — to classify tumors as **malignant** or **benign** based on various measurable medical features.

Among the three models, **KNN performed the best** in terms of accuracy and recall, making it the most reliable for detecting malignant cases in this context. However, all models showed strong performance, reflecting the effectiveness of machine learning in medical diagnostics.

The project also highlighted the importance of evaluating models not just based on accuracy, but also by understanding the **confusion matrix**, especially in healthcare where **false negatives** (failing to detect a malignant tumor) can lead to severe consequences. Hence, minimizing such errors is critical in real-world applications.

From a broader perspective, this work reinforces the role of **data-driven approaches** in improving clinical decision-making. With further enhancements — such as hyperparameter tuning, feature engineering, and ensemble learning — these models can become even more robust and generalizable. In conclusion, classification algorithms not only offer **automated, accurate, and scalable diagnostic tools**, but also have the potential to assist healthcare professionals in making informed decisions, ultimately leading to **improved patient outcomes and early intervention**.

