## **NAÏVE BAYES ANALYSIS FOR BREAST CANCER DIAGNOSIS - TASK 5**



|  |  |
| --- | --- |
| **NAME** | ABDUL MUIZZ, MUHAMMAD ABDULLAH |
| **REGISTRATION NUMBER** | FA23-BCS-117, FA23-BCS-108 |
| **DATE OF SUBMISSION** | 3RD OCTOBER 2025 |
| **MODERATOR** | SIR UMAR NAUMAN |

**Table of Contents:**

[**NAÏVE BAYES ANALYSIS FOR BREAST CANCER DIAGNOSIS - TASK 5** 1](#_Toc210203613)

[**Abstract:** 4](#_Toc210203614)

[**Problem Statement:** 5](#_Toc210203615)

[**Dataset Description:** 6](#_Toc210203616)

[**Class Distribution:** 6](#_Toc210203617)

[**Feature Categories:** 6](#_Toc210203618)

[**Methodology:** 8](#_Toc210203619)

[**1. Data Loading and Exploration** 8](#_Toc210203620)

[**2. Exploratory Data Analysis** 8](#_Toc210203621)

[**3. Data Preprocessing** 8](#_Toc210203622)

[**4. Model Training** 8](#_Toc210203623)

[**Data Analysis and Visualization:** 9](#_Toc210203624)

[**Target Distribution Analysis** 9](#_Toc210203625)

[**Feature Distribution Comparison** 10](#_Toc210203626)

[**Data Preprocessing:** 11](#_Toc210203627)

[**Feature Selection:** 11](#_Toc210203628)

[**Train-Test Split:** 11](#_Toc210203629)

[**Feature Standardization:** 11](#_Toc210203630)

[**Model Implementation:** 12](#_Toc210203631)

[**Gaussian Naive Bayes** 12](#_Toc210203632)

[**Algorithm Characteristics:** 12](#_Toc210203633)

[**Advantages:** 12](#_Toc210203634)

[**Result and Analysis:** 13](#_Toc210203635)

[**Performance Metrics:** 13](#_Toc210203636)

[**Detailed Classification Report (Test Set):** 13](#_Toc210203637)

[**Confusion Matrix Analysis:** 13](#_Toc210203638)

[**Interpretation:** 14](#_Toc210203639)

[**Error Analysis:** 14](#_Toc210203640)

[**Performance Metrics Breakdown:** 15](#_Toc210203641)

[**List of Figures:** 16](#_Toc210203642)

[**Conclusion:** 17](#_Toc210203643)

[**Key Findings:** 17](#_Toc210203644)

[**Model Strengths:** 17](#_Toc210203645)

[**Clinical Significance:** 17](#_Toc210203646)

[**Limitations:** 17](#_Toc210203647)

[**Future Work:** 18](#_Toc210203648)

[**Practical Applications:** 18](#_Toc210203649)

[**References:** 19](#_Toc210203650)

## **Abstract:**

This project implements a Gaussian Naive Bayes classifier to predict breast cancer diagnosis using the Wisconsin Breast Cancer Dataset. The study demonstrates the effectiveness of probabilistic classification methods in medical diagnosis applications. The model achieved a test accuracy of 96.49%, showing strong performance in distinguishing between benign and malignant tumors. The implementation includes comprehensive data preprocessing, feature standardization, and detailed performance evaluation using multiple metrics.

## **Problem Statement:**

Breast cancer diagnosis requires accurate classification of tumors as benign or malignant based on various cell characteristics. This project aims to:

* Apply Naive Bayes classification to breast cancer diagnosis
* Evaluate the effectiveness of probabilistic methods in medical classification
* Analyze feature distributions to understand diagnostic patterns
* Compare model performance using multiple evaluation metrics
* Provide a reliable automated diagnostic support system

## **Dataset Description:**

**Source:** breast-cancer.csv **Total Records:** 569 cases  
**Classes:** 2 (Benign: B, Malignant: M)  
**Features:** 30 numerical features plus ID and diagnosis columns

### **Class Distribution:**

* **Benign (B):** 357 cases (62.7%)
* **Malignant (M):** 212 cases (37.3%)

### **Feature Categories:**

All features are computed from digitized images of cell nuclei:

**Mean Features:**

* radius\_mean, texture\_mean, perimeter\_mean, area\_mean
* smoothness\_mean, compactness\_mean, concavity\_mean
* concave points\_mean, symmetry\_mean, fractal\_dimension\_mean

**Standard Error (SE) Features:**

* radius\_se, texture\_se, perimeter\_se, area\_se
* smoothness\_se, compactness\_se, concavity\_se
* concave points\_se, symmetry\_se, fractal\_dimension\_se

**Worst Features (largest values):**

* radius\_worst, texture\_worst, perimeter\_worst, area\_worst
* smoothness\_worst, compactness\_worst, concavity\_worst
* concave points\_worst, symmetry\_worst, fractal\_dimension\_worst

**Data Quality:**

* No missing values detected
* All features are numerical (float64)
* No data cleaning required

## **Methodology:**

### **1. Data Loading and Exploration**

* Loaded dataset from CSV file
* Examined data structure and information
* Checked for missing values
* Generated statistical summary

### **2. Exploratory Data Analysis**

* Analyzed class distribution
* Visualized target variable distribution
* Created histograms for key features
* Compared feature distributions between benign and malignant cases

### **3. Data Preprocessing**

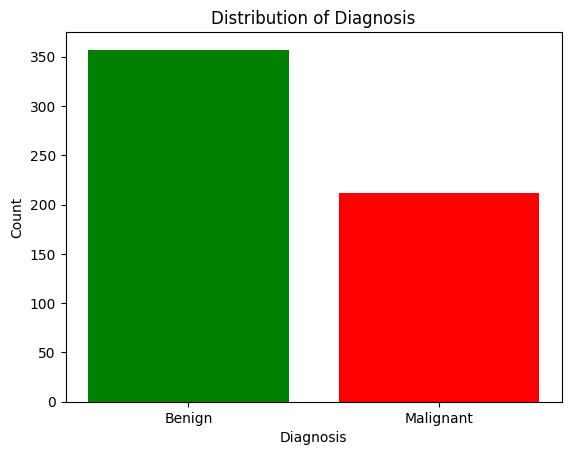
Steps performed:

* Removed ID column (not useful for prediction)
* Binary encoding of diagnosis (M=1, B=0)
* Split features (X) and target (y)
* Train-test split (80-20 ratio)
* Feature standardization using StandardScaler

### **4. Model Training**

* Algorithm: Gaussian Naive Bayes
* Training on standardized features
* Default scikit-learn parameters

## **Data Analysis and Visualization:**

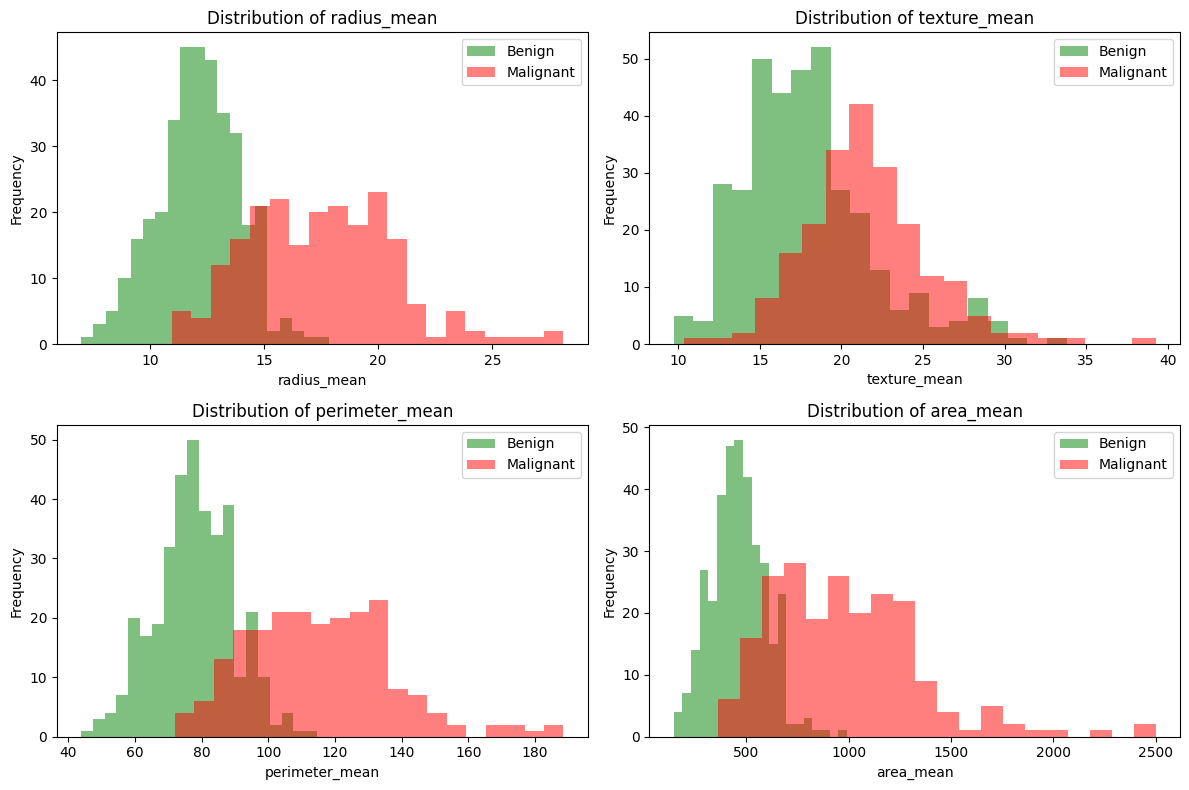


**Figure 1:** Bar chart showing distribution of diagnosis (Benign vs Malignant)

### **Target Distribution Analysis**

The dataset shows moderate class imbalance:

* Benign cases represent the majority (62.7%)
* Malignant cases represent 37.3%
* This imbalance is realistic for medical datasets



**Figure 2:** 2x2 grid of histograms comparing feature distributions

### **Feature Distribution Comparison**

Analysis of key features revealed:

1. **Radius Mean:** Clear separation between benign and malignant tumors, with malignant tumors showing larger radii
2. **Texture Mean:** Some overlap but malignant tumors tend to have higher texture values
3. **Perimeter Mean:** Similar pattern to radius, with malignant tumors having larger perimeters
4. **Area Mean:** Strong discriminative feature, malignant tumors consistently larger

The histogram visualizations show that malignant tumors generally have:

* Larger cell nuclei measurements
* Higher texture values
* Greater variability in measurements

## **Data Preprocessing:**

### **Feature Selection:**

The ID column was removed as it provides no predictive value. The diagnosis column was encoded as binary values with Malignant mapped to 1 and Benign mapped to 0.

### **Train-Test Split:**

* **Training Set:** 455 samples (80%)
* **Testing Set:** 114 samples (20%)
* **Random state:** 42 (for reproducibility)

### **Feature Standardization:**

Applied StandardScaler to normalize features:

* Fit on training data only
* Transform both training and test sets
* Ensures features have mean=0 and standard deviation=1
* Important for optimal model performance

## **Model Implementation:**

### **Gaussian Naive Bayes**

The Gaussian Naive Bayes classifier was implemented using scikit-learn's GaussianNB class. The model was trained on the standardized training data.

### **Algorithm Characteristics:**

* Assumes features follow Gaussian (normal) distribution
* Calculates probability of each class given feature values
* Makes predictions based on maximum probability
* Fast training and prediction
* Works well with continuous features

### **Advantages:**

* Simple and fast
* Handles high dimensional data well
* Requires relatively small training dataset
* Performs well even with class imbalance

## **Result and Analysis:**

### **Performance Metrics:**

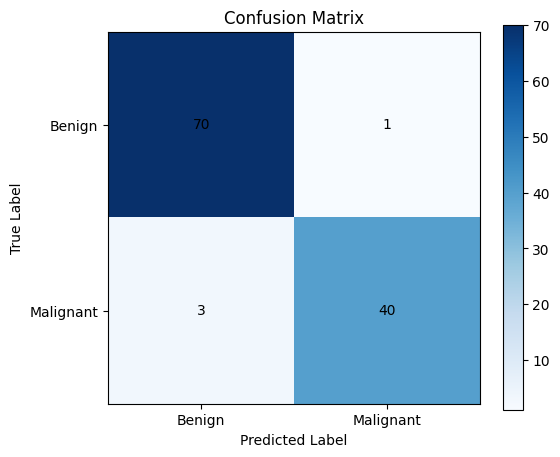
|  |  |  |
| --- | --- | --- |
| **Metric** | **Training** | **Testing** |
| **Accuracy** | 93.63% | **96.49%** |

### **Detailed Classification Report (Test Set):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-Score** | **Support** |
| **Benign** | 0.96 | 0.99 | 0.97 | 71 |
| **Malignant** | 0.98 | 0.93 | 0.95 | 43 |
| **Weighted Avg** | 0.97 | 0.96 | 0.96 | 114 |

### **Confusion Matrix Analysis:**

|  |  |  |
| --- | --- | --- |
| **Actual** | **Predicted** | |
|  | **Benign** | **Malignant** |
| **Benign** | 70 | 1 |
| **Malignant** | 3 | 40 |



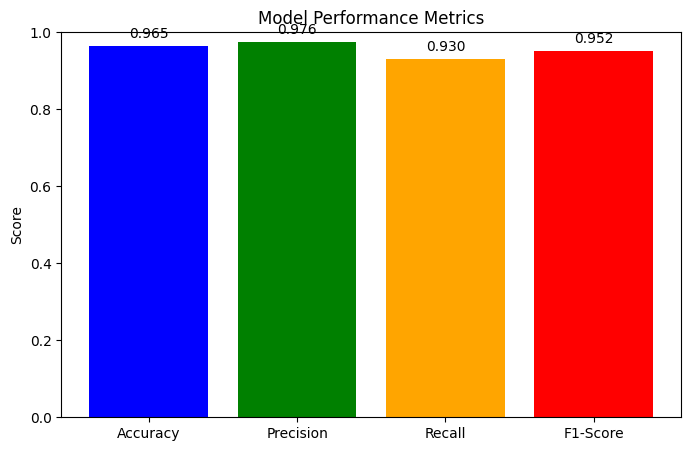
**Figure 3:** Confusion matrix heatmap showing prediction results

### **Interpretation:**

* **True Negatives (TN):** 70 - Correctly identified benign cases
* **False Positives (FP):** 1 - Benign incorrectly classified as malignant
* **False Negatives (FN):** 3 - Malignant incorrectly classified as benign
* **True Positives (TP):** 40 - Correctly identified malignant cases

### **Error Analysis:**

* **False Positive Rate:** 1.4% (1 out of 71 benign cases)
* **False Negative Rate:** 7.0% (3 out of 43 malignant cases)
* False negatives are more critical in medical diagnosis
* Model shows excellent sensitivity to malignant cases



**Figure 4:** Bar chart comparing performance metrics (Accuracy, Precision, Recall, F1-Score)

### **Performance Metrics Breakdown:**

1. **Precision (Malignant):** 98% - When model predicts malignant, it's correct 98% of the time
2. **Recall (Malignant):** 93% - Model correctly identifies 93% of all malignant cases
3. **F1-Score:** 95% - Harmonic mean shows balanced performance

## **List of Figures:**

1. **Figure 1:** Bar chart showing distribution of diagnosis (Benign vs Malignant). Page 9
2. **Figure 2:** 2x2 grid of histograms comparing feature distributions: Page 10

* radius\_mean distribution for benign vs malignant
* texture\_mean distribution for benign vs malignant
* perimeter\_mean distribution for benign vs malignant
* area\_mean distribution for benign vs malignant

1. **Figure 3:** Confusion matrix heatmap showing prediction results. Page 14
2. **Figure 4:** Bar chart comparing performance metrics (Accuracy, Precision, Recall, F1-Score). Page 15

## **Conclusion:**

### **Key Findings:**

1. **High Accuracy:** Achieved 96.49% test accuracy, exceeding training accuracy (good generalization).
2. **Excellent Precision:** 98% precision for malignant cases minimizes false alarms.
3. **Strong Recall:** 93% recall ensures most malignant cases are detected.
4. **Balanced Performance:** F1-score of 95% shows well-balanced model.
5. **Medical Relevance:** Low false negative rate is crucial for cancer diagnosis.

### **Model Strengths:**

1. Simple and interpretable algorithm
2. Fast training and prediction
3. Excellent performance on standardized features
4. Generalizes well (test > training accuracy)
5. Suitable for real-time diagnostic support

### **Clinical Significance:**

1. High recall for malignant cases ensures patient safety
2. Low false positive rate reduces unnecessary procedures
3. Fast prediction enables quick diagnosis
4. Can serve as screening tool or second opinion system

### **Limitations:**

* **Naive Independence Assumption:** Assumes features are independent (may not be true)
* **Small Dataset:** Only 569 samples may limit generalizability
* **Binary Classification:** Doesn't provide staging information
* **Feature Correlation:** Some features highly correlated (violated Naive Bayes assumption)

### **Future Work:**

* **Cross Validation:** Implement k-fold cross-validation for robust evaluation
* **Feature Selection:** Identify and remove highly correlated features
* **Ensemble Methods:** Compare with other algorithms (SVM, Random Forest, Neural Networks)
* **Clinical Validation:** Test on independent medical datasets
* **Probability Calibration:** Fine-tune probability estimates for clinical decision-making
* **Cost Sensitive Learning:** Weight false negatives more heavily
* **External Validation:** Test on data from different hospitals/populations

### **Practical Applications:**

1. **Screening Tool:** Initial automated screening of biopsy samples
2. **Decision Support:** Assist pathologists in diagnosis
3. **Quality Control:** Second opinion system to catch potential errors
4. **Research Tool:** Analyze feature importance in cancer diagnosis
5. **Educational Aid:** Teaching tool for medical students

## **References:**

1. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine
2. Wolberg, W.H., Street, W.N., and Mangasarian, O.L. (1995). "Breast cancer diagnosis via linear programming"
3. Zhang, H. (2004). "The optimality of Naive Bayes". FLAIRS Conference
4. Rish, I. (2001). "An empirical study of the naive Bayes classifier". IJCAI Workshop
5. Scikit-learn Documentation: Naive Bayes Classifiers