صورة تحتوي على نص, شعار, لقطة شاشة, التصميم

تم إنشاء الوصف تلقائياً



**Task 1: Naive Bayes Classifier**

**"Breast Cancer Wisconsin Dataset"**

**DS3114))**

**Dr. Omaima Fallatah))**

|  |  |
| --- | --- |
| **ID:** | **Name:** |
| **444005403** | **Sara Ayed Alsehli** |
| **443009106** | **Lina Ibrahim Alghamdi** |

**Table of Contents:**

[**Introduction** 3](#_Toc180784537)

[**Data Preprocessing and Exploration** 4](#_Toc180784538)

[Loading the Data: 4](#_Toc180784539)

[Data Cleaning: 4](#_Toc180784540)

[Visualization of Class Distribution: 4](#_Toc180784541)

[Mapping Diagnosis to Numerical Values: 5](#_Toc180784542)

[**Feature Selection** 6](#_Toc180784543)

[Correlation Analysis: 6](#_Toc180784544)

[Dropping Less Important Features: 6](#_Toc180784546)

[**Model Building** 7](#_Toc180784548)

[Model choice: 7](#_Toc180784549)

[Data Splitting: 7](#_Toc180784550)

[**Performance Evaluation** 8](#_Toc180784552)

[Accuracy: 8](#_Toc180784553)

[Classification Report: 8](#_Toc180784554)

[Confusion Matrix: 9](#_Toc180784555)

[**Insights gained from the model** 10](#_Toc180784556)

[Key features: 10](#_Toc180784557)

[Future Improvements: 10](#_Toc180784558)

[**Conclusion** 11](#_Toc180784559)

[**References** 12](#_Toc180784560)

# 

# **Introduction**

Breast cancer is one of the most common cancers affecting women worldwide. Early detection is crucial for effective treatment. In this project, we use the **Gaussian Naive Bayes** classification algorithm to predict whether a tumor is **benign** (non-cancerous) or **malignant** (cancerous), based on the features of the tumor dataset.

The dataset used is the **Breast Cancer Wisconsin (Diagnostic) Data Set**, which contains features of cell nuclei extracted from digital images of fine needle aspirates of breast masses. The goal is to build a model that can classify tumors as benign or malignant based on these features.

Dataset Overview:

Rows and Columns: The dataset contains 569 rows (representing patients) and 33 columns (features describing tumor characteristics).

Columns: Key features include "radius\_mean,texture\_mean, perimeter\_mean, area\_mean", and diagnosis (indicating whether the tumor is benign (B) or malignant (M)).

Rows: Each row represents an individual patient’s tumor measurements.

# 

# **Data Preprocessing and Exploration**

## Loading the Data:

The dataset was loaded from a CSV file using the Pandas library. After loading, the structure and basic information about the

**Data Cleaning:**

**Dropping Unnecessary Columns:** The dataset contained columns that were not useful and unnecessary for the analysis, such as "Unnamed: 32" and "id". These were removed to clean the dataset.

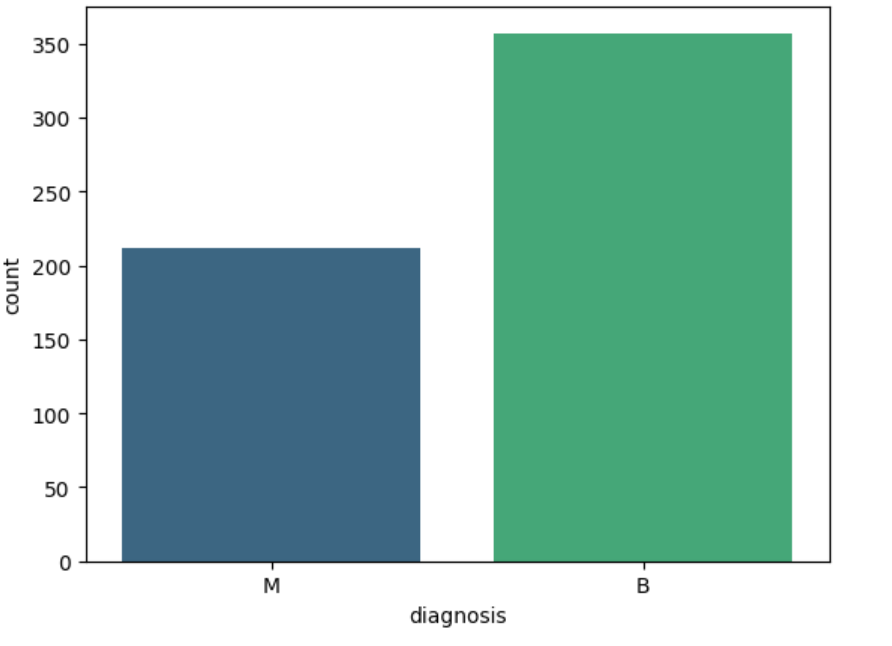
**Handling Missing Values:** Rows with missing values in important columns like "radius\_mean" and "texture\_mean" were dropped to ensure data quality.

## Visualization of Class Distribution:

using Seaborn, as they included a ( countplot) to show the distribution of the two classes (B and M). This visualization helps in understanding whether the dataset is balanced or skewed, which is an important factor for model performance.

## Mapping Diagnosis to Numerical Values:

This step is crucial for machine learning models since they require numerical inputs. The diagnosis column was mapped from **B (benign)** to **0** and **M (malignant)** to **1**, which aligns with your statement.



# **Feature Selection**

## Correlation Analysis:

### We calculated the correlation between each feature and the target variable (diagnosis). This helped us understand which features are most strongly correlated with whether the tumor is benign or malignant.

## Dropping Less Important Features:

### Based on correlation analysis, some features that had very low correlation with the target variable were dropped. These features included:

* symmetry\_se
* fractal\_dimension\_mean
* texture\_se
* compactness\_se, etc.

This step reduced the dimensionality of the dataset and simplified the model without sacrificing performance.

# **Model Building**

## Model choice:

**)The Gaussian Naive Bayes)** model was chosen because it is one of the simplest and most effective models for classification problems, particularly when data is close to a normal distribution. This model is suitable for its ease of implementation and quick results with datasets of medium size, such as this breast cancer dataset.

The choice of **Gaussian Naive Bayes** was based on its balance between simplicity and accuracy, making it a strong candidate for a task that requires reliable and fast classification.

## Data Splitting:

### The dataset was split into **training** and **testing** sets, with 80% of the data used for training and 20% for testing. This ensures that the model can be evaluated on unseen data.

# **Performance Evaluation**

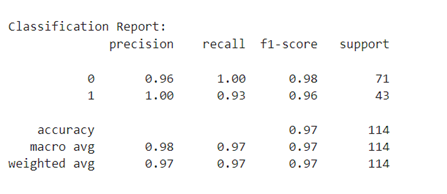
The model’s performance was evaluated using several metrics:

## Accuracy:

The model achieved (**97.37%)** accuracy, meaning it was able to correctly classify the majority of the test cases.

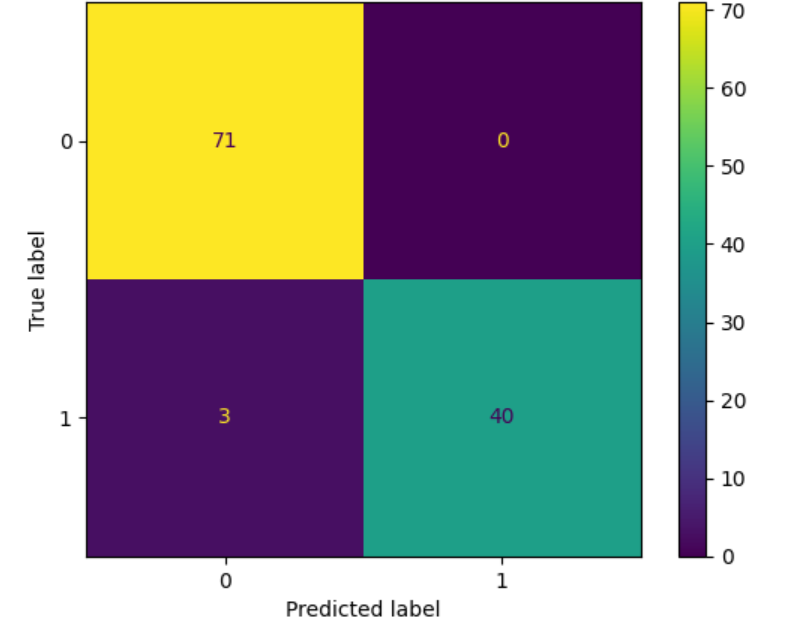


## Classification Report:

**** This included precision, recall, and F1-score, which all indicated strong performance, demonstrating the model’s effectiveness in distinguishing between malignant and benign tumors.

## Confusion Matrix:

A confusion matrix was used to analyze the classification results between the two categories (benign and malignant), helping to better understand the model’s mistakes.



# **Insights gained from the model**

## Key features:

such as (radius\_mean, perimeter\_mean, and area\_mean) had the strongest correlation with tumor malignancy.

The **Gaussian Naive Bayes** model performed well because of its simplicity and efficiency. However, it assumes that features are normally distributed, which might not be ideal for all datasets.

## Future Improvements:

For future analysis, more complex models such as Random Forest or Support Vector Machine (SVM) could be explored, as they may provide higher accuracy, especially in cases where the relationships between features are non-linear.

# **Conclusion**

In summary, the breast cancer diagnosis task demonstrated the effective application of machine learning in healthcare, specifically using the Gaussian Naive Bayes algorithm to classify tumors with a high level of accuracy. The preprocessing steps, including feature selection and data cleaning, played a key role in enhancing the model's performance. Although the chosen model delivered strong results, future improvements could involve experimenting with more advanced algorithms to further refine the predictions. Ultimately, this analysis provides a solid foundation for understanding which tumor characteristics are most indicative of malignancy, offering valuable insights for future medical research and diagnostic applications.

# **References**

(<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)>)