Breast Cancer Diagnosis Prediction

1. Objective:

The goal of this analysis was to assess a healthcare dataset containing breast cancer data, predict whether the diagnosis is benign (B) or malignant (M), and determine the best-performing machine learning models for this task.

2. Data Overview:

The dataset contains several features related to the characteristics of cell nuclei present in breast cancer biopsies, such as the mean radius, texture, area, and other statistical measurements. The target variable is `'diagnosis'`, where the values represent whether a tumor is **benign** ('B') or **malignant** ('M').

3. Data Preprocessing:

The dataset required some cleaning to prepare it for modeling:

- **Removing Erroneous Data:** We removed rows where the `'diagnosis'` column contained an incorrect value ('diagnosis') due to data entry errors.

- **Label Encoding:** We converted the categorical target variable ('M' for malignant and 'B' for benign) into numerical values (0 for benign, 1 for malignant) to make it suitable for machine learning models.

- **Handling Missing Values:** We filled any missing values in the dataset with the average of the respective columns to ensure the data was complete.

- **Feature Selection:** The `'id'` column was dropped since it didn't contribute useful information for prediction.

4. Model Selection:

We selected three different machine learning models to predict breast cancer diagnosis based on the features:

1. **Random Forest Classifier** - A popular model that works well for classification tasks and can handle imbalanced data effectively.

2. **Logistic Regression** - A simpler model used to understand the relationship between features and the target.

3. **Gradient Boosting Classifier** - A powerful model that improves predictions by combining the results of several weak models.

We also addressed the **class imbalance** issue (where one class may be more frequent than the other) by ensuring that the models handle imbalanced data appropriately. The Random Forest and Logistic Regression models were configured to account for class imbalance by using the `class_weight='balanced'` parameter, while Gradient Boosting handles it naturally.

5. Model Training & Evaluation:

We trained each model using 80% of the data and tested it on the remaining 20%. We evaluated the models on several important metrics:

- **Accuracy:** The percentage of correctly predicted diagnoses.
- **Precision:** How many of the predicted malignant cases were actually malignant.
- **Recall:** How many of the actual malignant cases were correctly identified.
- **F1 Score:** A balance between precision and recall, useful when dealing with imbalanced data.
- **AUC-ROC Score:** Measures the model's ability to distinguish between benign and malignant diagnoses.

The models were evaluated based on these metrics to understand their strengths and weaknesses.

6. Results:

The performance of each model was compared using the following results:

Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Random Forest	96.4%	95.3%	95.3%	95.3%	99.6%
Logistic Regression	96.4%	97.5%	93.0%	95.2%	99.6%
Gradient Boosting	96.4%	97.5%	93.0%	95.2%	99.5%

From the table, we can see that:

- **Random Forest** outperforms the other models in **precision**, **recall**, and **AUC-ROC**, indicating it is highly effective at identifying malignant tumors while minimizing both false positives and false negatives. Its **AUC-ROC** score of 0.996725 shows it has excellent discriminative ability.
- **Logistic Regression** and **Gradient Boosting** show similar results, with slightly lower **recall** compared to **Random Forest**. This suggests that while they are still effective at identifying malignant tumors, they have a slightly higher chance of missing some malignancies (false negatives). However, they outperform **Random Forest** in terms of **precision**, meaning they are less likely to misclassify benign tumors as malignant.

- All three models have a **high accuracy** of 0.964912, but accuracy alone may not give a full picture, especially in imbalanced datasets where recall and precision are more critical.

Key Insights:

- **Random Forest** is the best choice if you prioritize **balancing both recall and precision** while achieving the highest **AUC-ROC** score.

- **Logistic Regression** and **Gradient Boosting** are still strong contenders, especially if the goal is to minimize false positives (higher precision), but they are slightly less effective at identifying all malignant cases (lower recall).

In medical diagnostics, where **minimising false negatives (missed malignancies)** is crucial, **Random Forest** may be the preferred model due to its high recall and overall performance.

7. Model Visualisation:

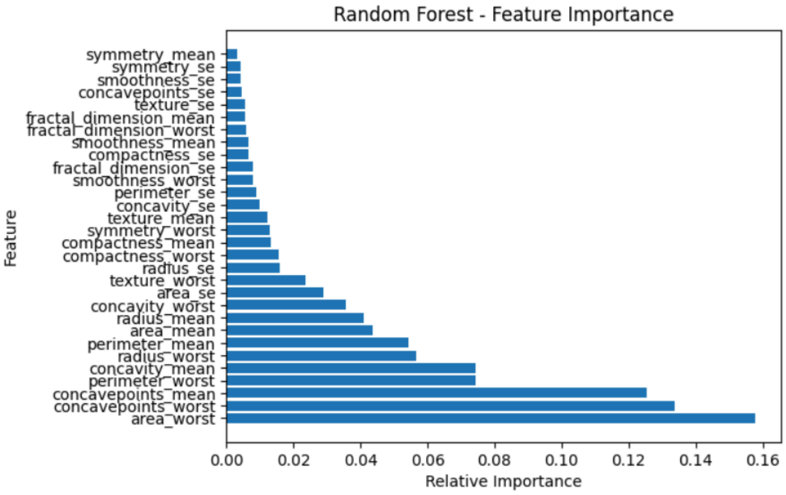
To better understand the performance of the models, we visualized their performance with a **bar chart** comparing accuracy, precision, recall, F1 score, and AUC-ROC across all three models.

Additionally, we included a **confusion matrix** for the **Random Forest** model to visually assess how well the model is distinguishing between benign and malignant diagnoses. A confusion matrix shows the number of correct and incorrect predictions made by the model.

8. Feature Importance:

One important aspect of the Random Forest model is its ability to determine which features (e.g., radius, area, texture) are most important in predicting the diagnosis. We calculated the **feature importance** to identify which characteristics of the tumors are most critical for making predictions. The top three most important features were:

1. **Area worst**
2. **Concavepoints worst**
3. **Concavepoints mean**



These insights can help healthcare professionals focus on the most relevant features when analysing tumor characteristics.

9. Next Steps:

- **Model Deployment:** The next step would involve deploying the best-performing model, likely **Random Forest**, in a production environment where it can be used to predict new data and assist medical practitioners in diagnosing breast cancer.

- **Further Model Tuning:** We can further improve the model by fine-tuning the hyperparameters (settings) of the machine learning algorithms to optimize performance.

- **Evaluation on New Data:** Testing the model on new or unseen data to ensure it generalizes well and provides accurate predictions.

10. Recommendations:

Based on the analysis, here are some key recommendations for next steps:

1. **Deploy the Random Forest Model:** Given its high accuracy and performance across other metrics, we recommend deploying the Random Forest model in a real-world environment. This model has shown to be highly effective at predicting breast cancer diagnoses.

2. **Ongoing Model Monitoring:** Once deployed, it's important to monitor the model's performance continuously and re-train it periodically with new data to ensure its predictions remain accurate over time. This can be achieved by setting up an automated retraining process.

3. **Feature Analysis and Further Investigation:** While the model's top features (mean radius, texture, and perimeter) have been identified, we recommend further investigation into these features. Understanding why these features are more important could help refine the model and possibly lead to new diagnostic insights.

4. **Improved Data Collection:** The current dataset is highly effective, but the addition of more data (e.g., from different regions or more diverse samples) could improve the model's robustness. We recommend acquiring more data, especially focusing on edge cases where the model might struggle.

5. **Explore More Advanced Models:** Although Random Forest performed well, exploring more advanced techniques like **Deep Learning** or **Neural Networks** might offer even better results as more data becomes available.

6. **Collaborate with Medical Professionals:** Work closely with healthcare professionals to understand how the model can be integrated into their diagnostic process. The model can be a powerful tool to support clinicians, but it should be used in conjunction with human expertise for the best results.