**Breast Cancer Prediction**

## Objective:

I work with the Breast Cancer dataset to explore patterns in the data and build a predictive model for the diagnosis feature, which indicates whether a tumor is **benign** or **malignant**.

My plan is to first perform a short exploratory analysis to understand the basic structure of the dataset and the relationships between features. After that, I will apply **Principal Component Analysis (PCA)** to reduce the dimensionality of the dataset. PCA will allow me to compress the information from many features into a smaller number of components, while retaining most of the variance in the data.

I will then train a classification model on the transformed data to predict the diagnosis.

| **Benign (0)**   * The tumor is **non-cancerous**. * It does **not invade nearby tissue** or spread to other parts of the body. * Usually not life-threatening, though it may still need monitoring or removal. |
| --- |
| **Malignant (1)**   * The tumor is **cancerous**. * It can **grow aggressively, invade surrounding tissues, and spread (metastasize)** to other organs. * Requires treatment (surgery, chemotherapy, radiation, etc.). |

Here’s how some features help distinguish benign (harmless) vs malignant (cancerous) tumors:

| **Size-related features** | **radius\_mean, perimeter\_mean, area\_mean** | Malignant tumors tend to have **larger nuclei** → so these values are often higher.  Benign tumors usually have **smaller, more uniform nuclei**. |
| --- | --- | --- |
| **Shape irregularity** | **compactness\_mean, concavity\_mean, concave points\_mean** | Malignant tumors often have **irregular, spiky, or concave shapes**.  Benign tumors are usually **rounder and smoother** |
| **Texture / surface details** | **texture\_mean** | Measures variation in gray-scale intensity (essentially how rough the surface looks).  Malignant tumors usually have **more varied texture**. |
| **Smoothness & symmetry** | **smoothness\_mean, symmetry\_mean** | Benign tumors tend to be **more symmetrical** and smoother.  Malignant tumors are **less symmetrical**, more irregular. |
| **Worst-case measurements** | **(\*\_worst)** **radius\_worst, concavity\_worst** | These represent the **largest or most abnormal values** observed in the sample.  Malignant tumors typically show **extreme worst-case values**, while benign ones don’t. |
| **Variation in measurements** | **(\*\_se)** | **Standard error (SE)** features capture how much variation exists across nuclei.  Malignant tumors tend to have **higher variability** → more abnormal and inconsistent cell sizes/shapes. |

**In short:**

* **Benign → smaller, round, smooth, uniform cells.**
* **Malignant → larger, irregular, variable, asymmetric cells.**

## Loading the Data

First, I imported the necessary libraries, including numpy and pandas for data manipulation, as well as os to explore the available dataset files in the */kaggle/input* directory.

Once the file path was identified, I loaded the dataset into a Pandas DataFrame using *pd.read\_csv* for further analysis.

## Dataset Overview

To begin the exploratory analysis, I first inspected the general structure of the dataset. I checked the dataset's **shape** to understand the number of rows and columns, then listed all **column names** to identify the available features and reviewed the **data types** of each column to determine which features are numerical or categorical. Finally, I checked for **missing values** in each column to ensure data completeness.

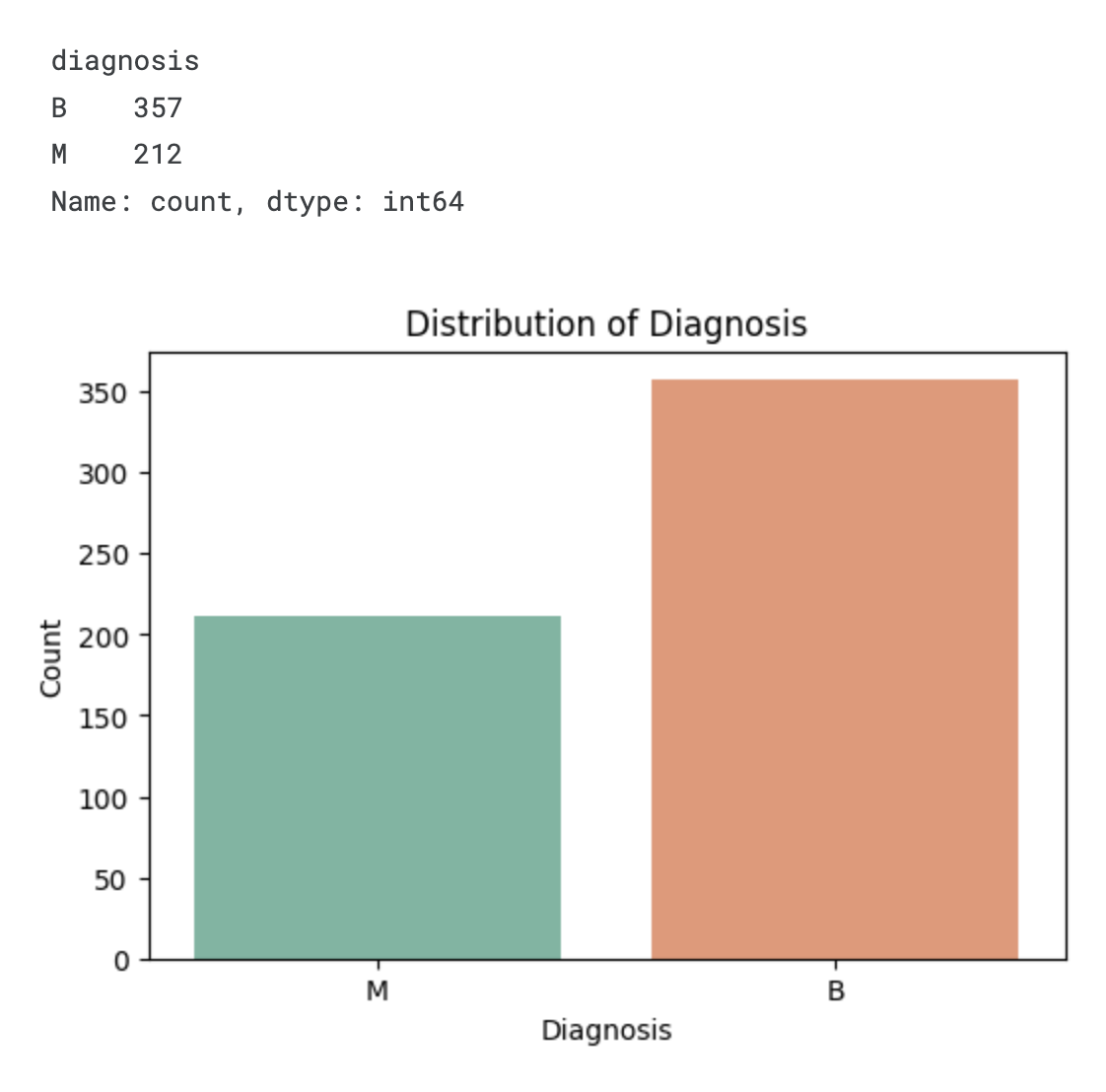
The dataset contains 569 rows and 33 columns.

An inspection of column names revealed that the id column and the Unnamed: 32 column are not relevant for the analysis, as they do not provide predictive information. These columns will be removed in the next step.

The data types indicate that all features, except for the target column diagnosis, are numerical (float64). A check for missing values confirmed that the dataset is complete.

## Target Variable Analysis

After removing the id and Unnamed:32 columns, I checked the distribution of the target variable diagnosis.

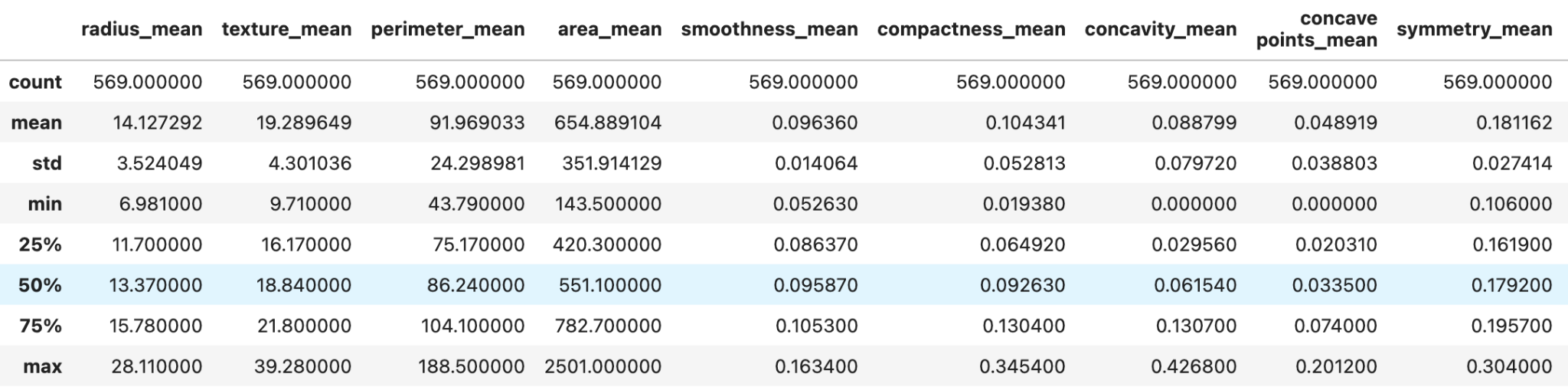


It contains two classes — **Benign (B)** and **Malignant (M)**.

From the count plot, I can see that benign cases occur more frequently than malignant ones, which means the dataset is slightly imbalanced.

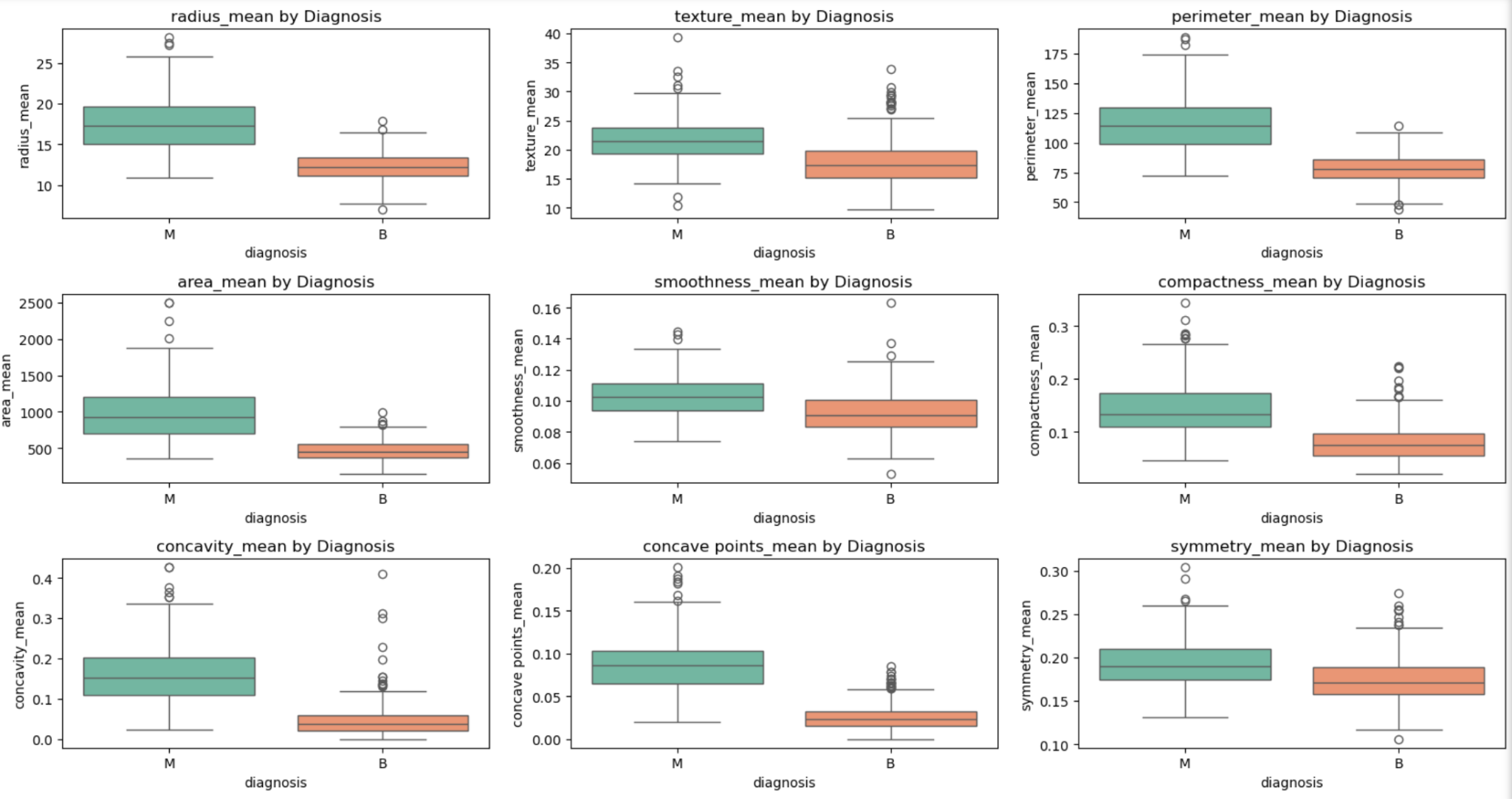
## Numerical Feature Overview

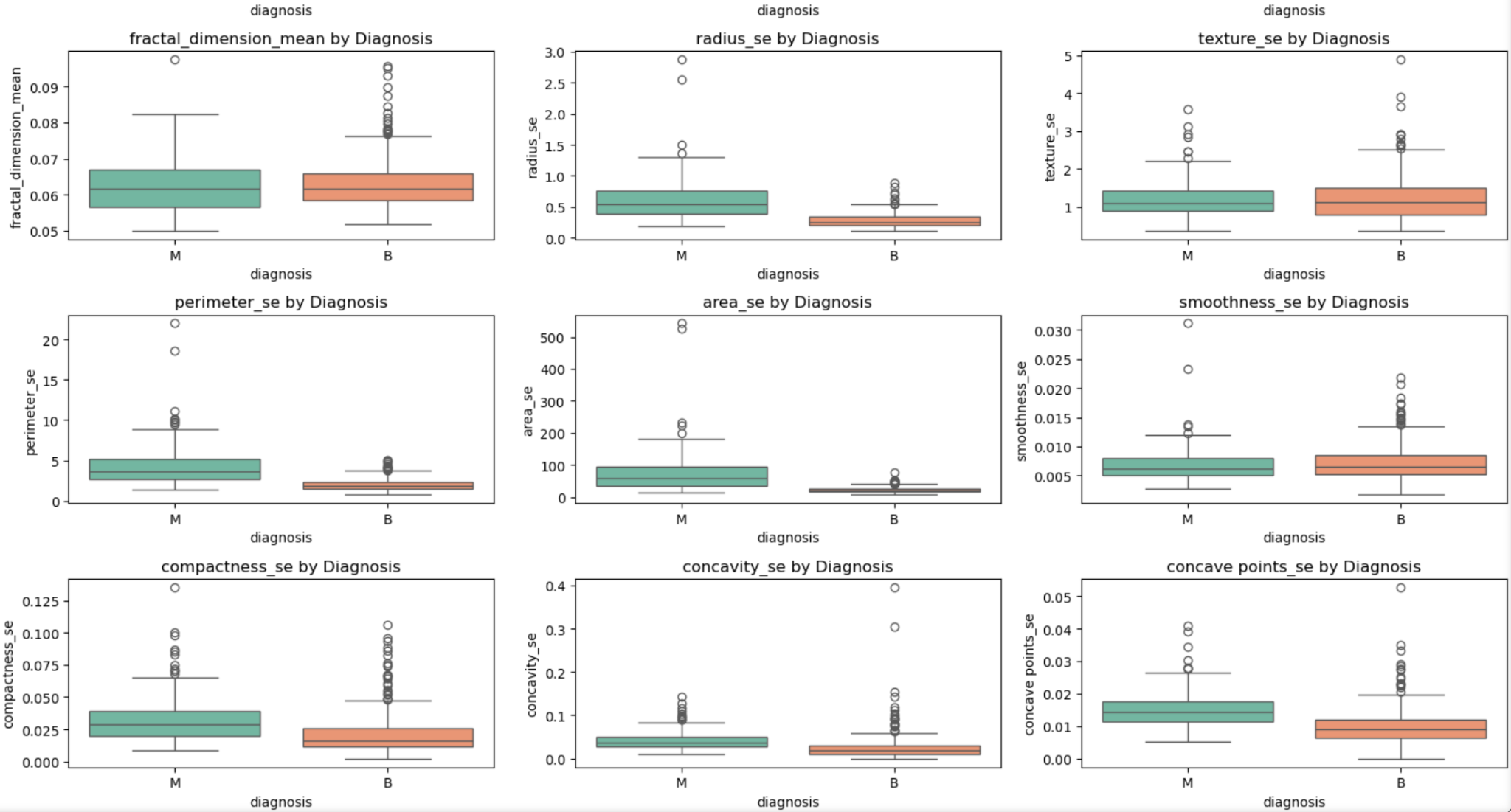
To better understand the numeric features in my dataset, I started by generating summary statistics using the *.describe()* method. This gave me the mean, standard deviation, and range for each numeric column, which helps to identify general trends and possible outliers.

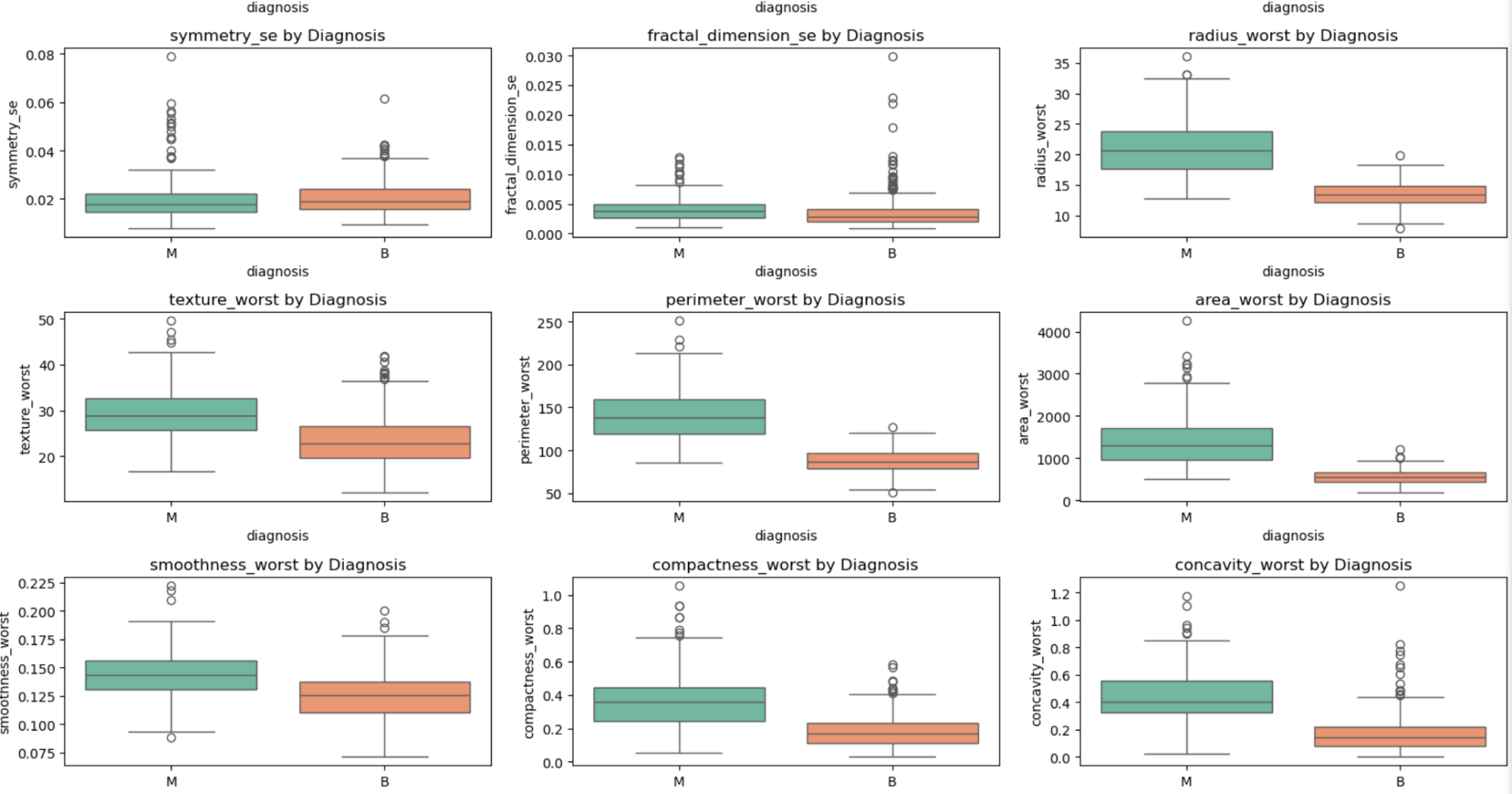


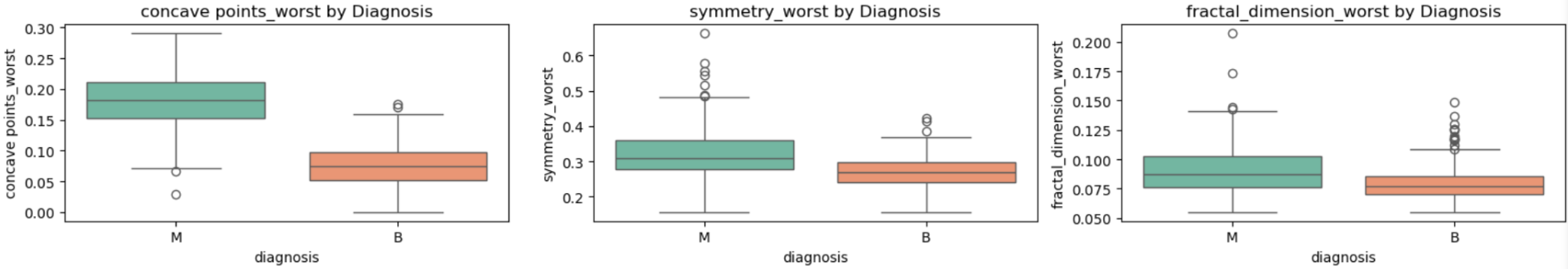
From the *.describe()* output, I can also see that the ranges of different features vary, which means I will need to standardize the data so that all columns are on the same scale before applying PCA or training models.

Next, I created boxplots for every numeric feature, grouped by the diagnosis column. These boxplots allow me to visually compare the distributions of each feature for benign and malignant cases.









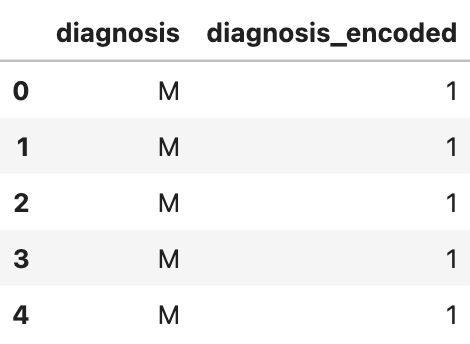
Looking at the boxplots, the features where the median difference between malignant (M) and benign (B) cases is visually the largest are:

* **radius\_mean** – malignant cases have a much higher median.
* **perimeter\_mean** – strong difference, closely follows radius\_mean.
* **area\_mean** – also much higher in malignant cases.
* **concavity\_mean** – clearly higher in malignant cases.
* **concave\_points\_mean** – large gap between M and B.
* **radius\_worst** – strong difference.
* **perimeter\_worst** – large difference similar to radius\_worst.
* **area\_worst** – very distinct separation.
* **concave\_points\_worst** – strong separation.
* **concavity\_worst** – also higher for malignant cases.

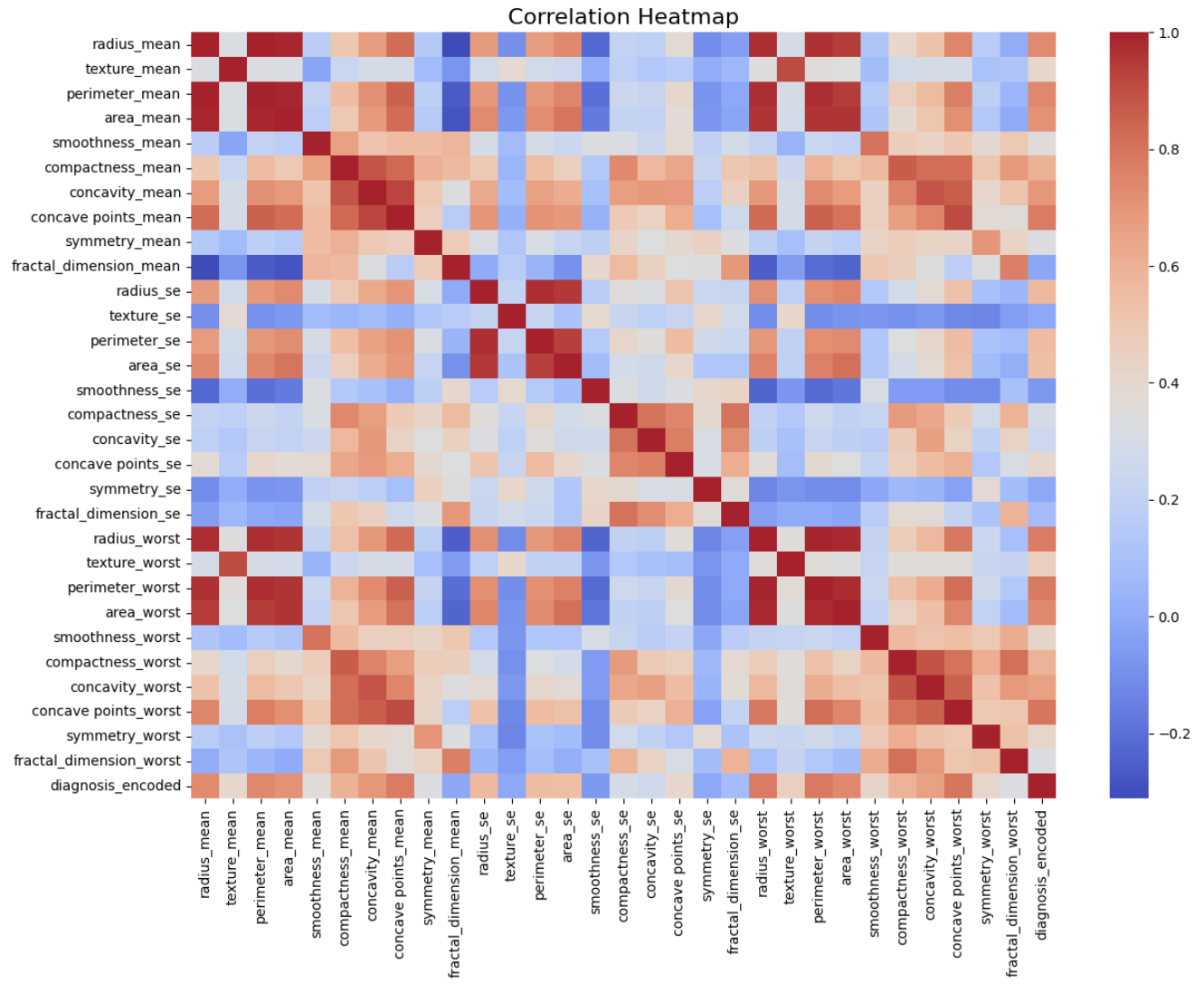
These features seem to have the most discriminative power just by comparing medians.

## Correlation Heatmap

Before creating the correlation heatmap, I first need to encode the diagnosis column into numerical values so it can be included in the correlation calculation.  
I will map *benign* *cases (B)* to **0** and *malignant cases (M)* to **1**, creating a new column called *diagnosis\_encoded*.



Now I can generate a correlation heatmap to visualize the relationships between all features and the encoded diagnosis.



The heatmap shows strong correlation clusters, especially among size-related features *(radius, perimeter, area)* and shape-related *(concavity, concave\_points, compactness)*.

High collinearity suggests that PCA can effectively reduce dimensionality, while preserving predictive information from the most discriminative features.

### Preprocessing for PCA

Before applying PCA, I separated the dataset into features (X) and target (y). Since PCA is sensitive to scale, I standardized all features using *StandardScaler* so that each variable has zero mean and unit variance.

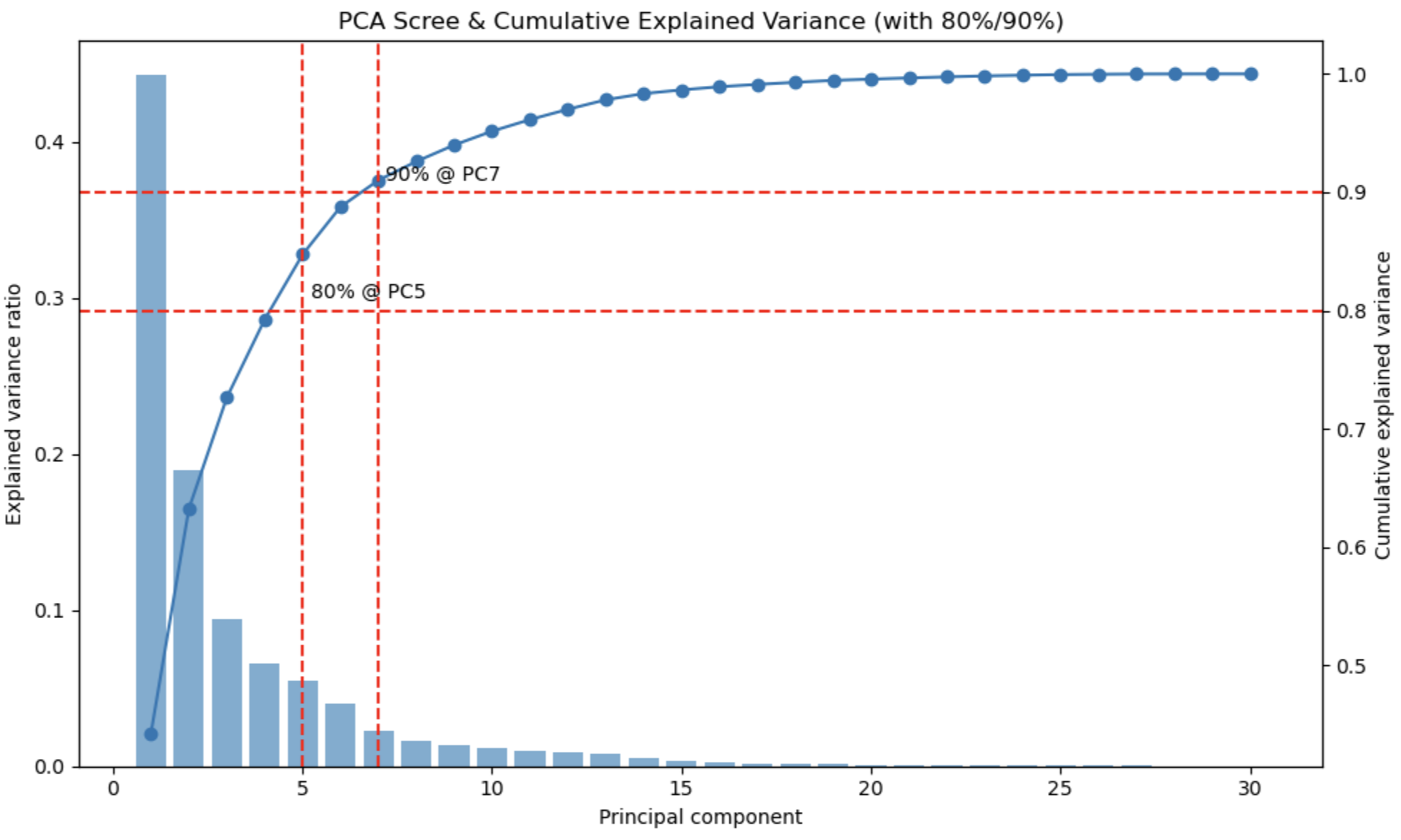
| Shape of X\_scaled: (569, 30) |
| --- |

## Applying PCA

To reduce the dimensionality of the dataset, I applied Principal Component Analysis (PCA) on the standardized features. PCA transforms the original correlated variables into a smaller number of uncorrelated components, while retaining most of the variance present in the dataset.

The scree plot below shows the explained variance ratio for each principal component along with the cumulative explained variance curve.

I also added horizontal and vertical dashed lines to mark the points where 80% and 90% of the total variance are explained.



From the plot, I can observe that:

* **5 components** are enough to retain **80%** of the variance.
* **7 components** are enough to retain **90%** of the variance.

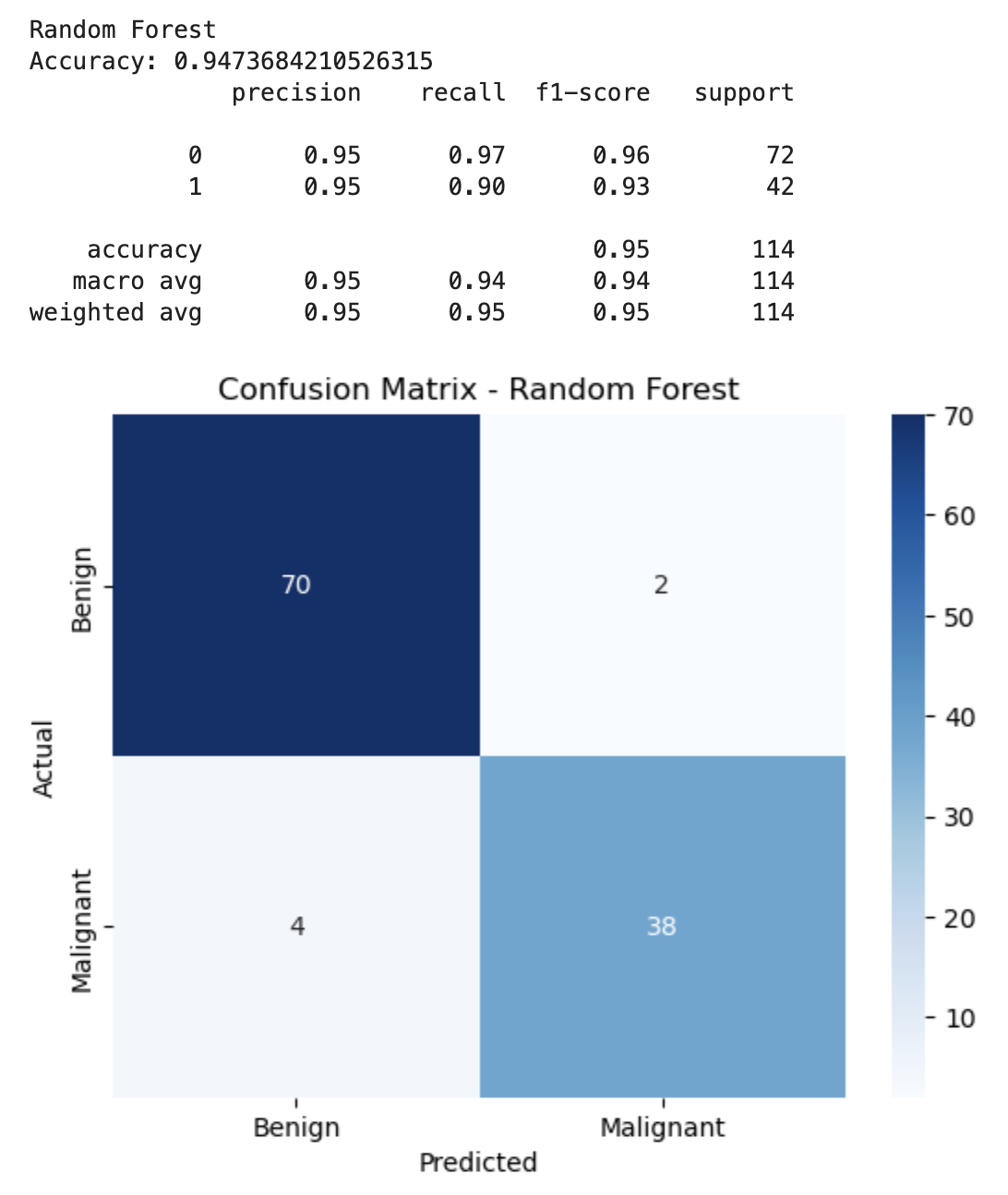
This means that instead of working with all 30 features, we can significantly reduce dimensionality with minimal information loss.

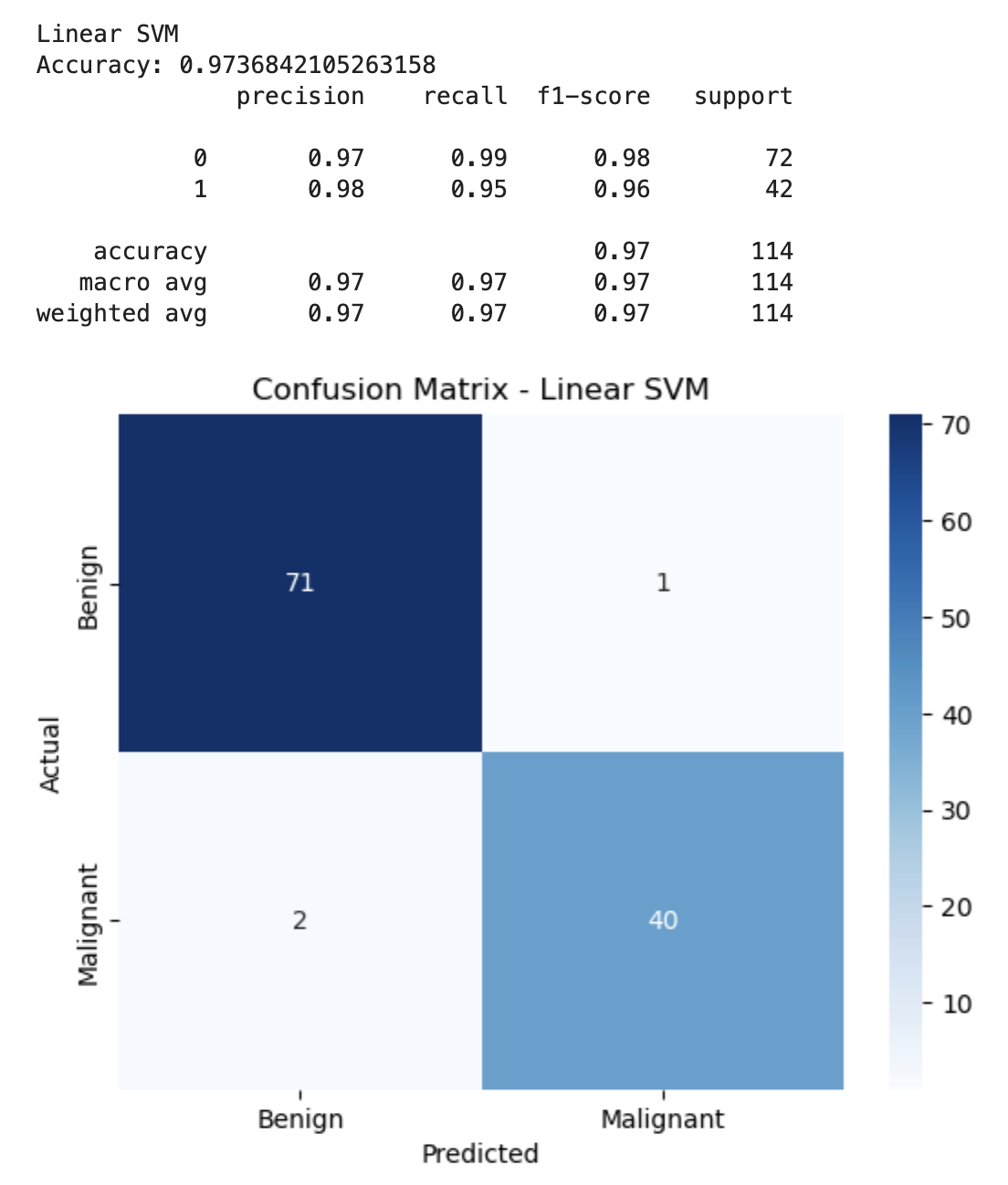
## Building the Prediction Model

To evaluate the predictive power of PCA-reduced features, I tested three different classifiers:

1. **Logistic Regression** – A simple, interpretable linear model.
2. **Random Forest Classifier** – A tree-based ensemble model capable of capturing nonlinear relationships.
3. **Linear SVM** – A linear Support Vector Machine for binary classification.

## 





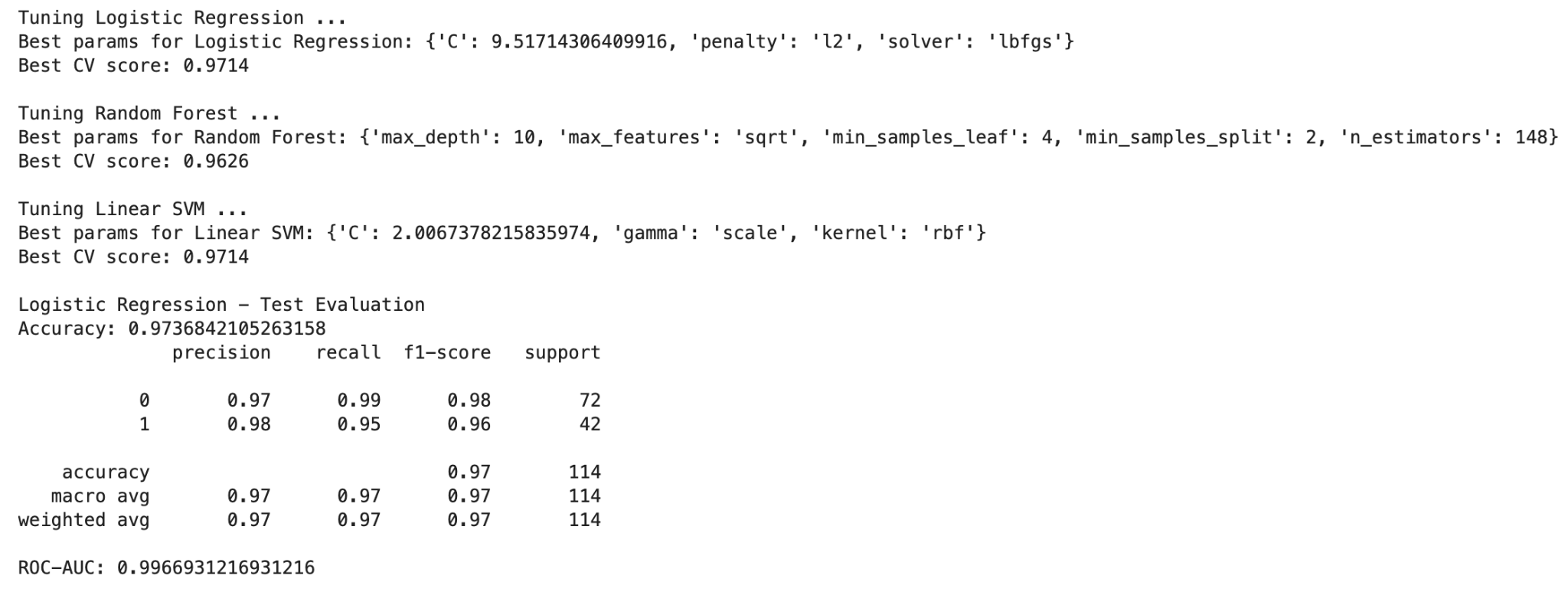
From the modeling results:

* **Logistic Regression** and **Linear SVM** achieved the highest accuracy of **97.37%**, with balanced precision and recall for both benign and malignant cases.
* **Random Forest** performed slightly worse with **94.74% accuracy**, indicating that tree-based models may require further tuning or more data for optimal performance.
* Confusion matrices show that all models have a small amount of misclassifications, with slightly more false negatives (malignant predicted as benign) for Random Forest.

Overall, both Logistic Regression and Linear SVM demonstrate excellent potential for breast cancer classification using PCA-transformed features.

## Hyperparameter Tuning

This code automatically finds the best hyperparameters for Logistic Regression, Random Forest, and SVM using RandomizedSearchCV, then evaluates them with accuracy, classification report, confusion matrix, ROC-AUC, and ROC curves on test data.



From this result, we can see that Logistic Regression and SVM are the top models here. The model can reliably distinguish **benign vs malignant tumors** with near-perfect ROC-AUC, meaning very high confidence in probability-based predictions.

## FLAML AutoML

This code builds an **automated breast cancer classifier** by preprocessing data (scaling + PCA), using **FLAML AutoML** to find the best model within 60 seconds, evaluates it on a test set, and saves both the preprocessing pipeline and the best model for future predictions.

### **Insights from Results**

**Best Model Found:**

* + FLAML selected an **SGDClassifier** (loss='log\_loss', penalty='l1') as the best model.

**Performance:**

* + **Test Accuracy:** 0.982 → very strong, even better than tuned Logistic Regression (~0.973) and Random Forest (~0.963).
  + **ROC-AUC:** 0.997 → extremely close to a perfect classifier, meaning excellent separability between classes.

**Class-wise Behavior:**

* + **Class 0 (majority):** Precision 0.97, Recall 1.00 → no false negatives, meaning all actual 0s were correctly captured.
  + **Class 1 (minority):** Precision 1.00, Recall 0.95 → some false negatives, but no false positives. So, whenever the model predicts "1", it’s always correct, though it misses ~5% of the actual "1"s.

This indicates a slight **recall trade-off** for class 1 but perfect precision, which could be preferable depending on the business use case (e.g., if false positives are costly).

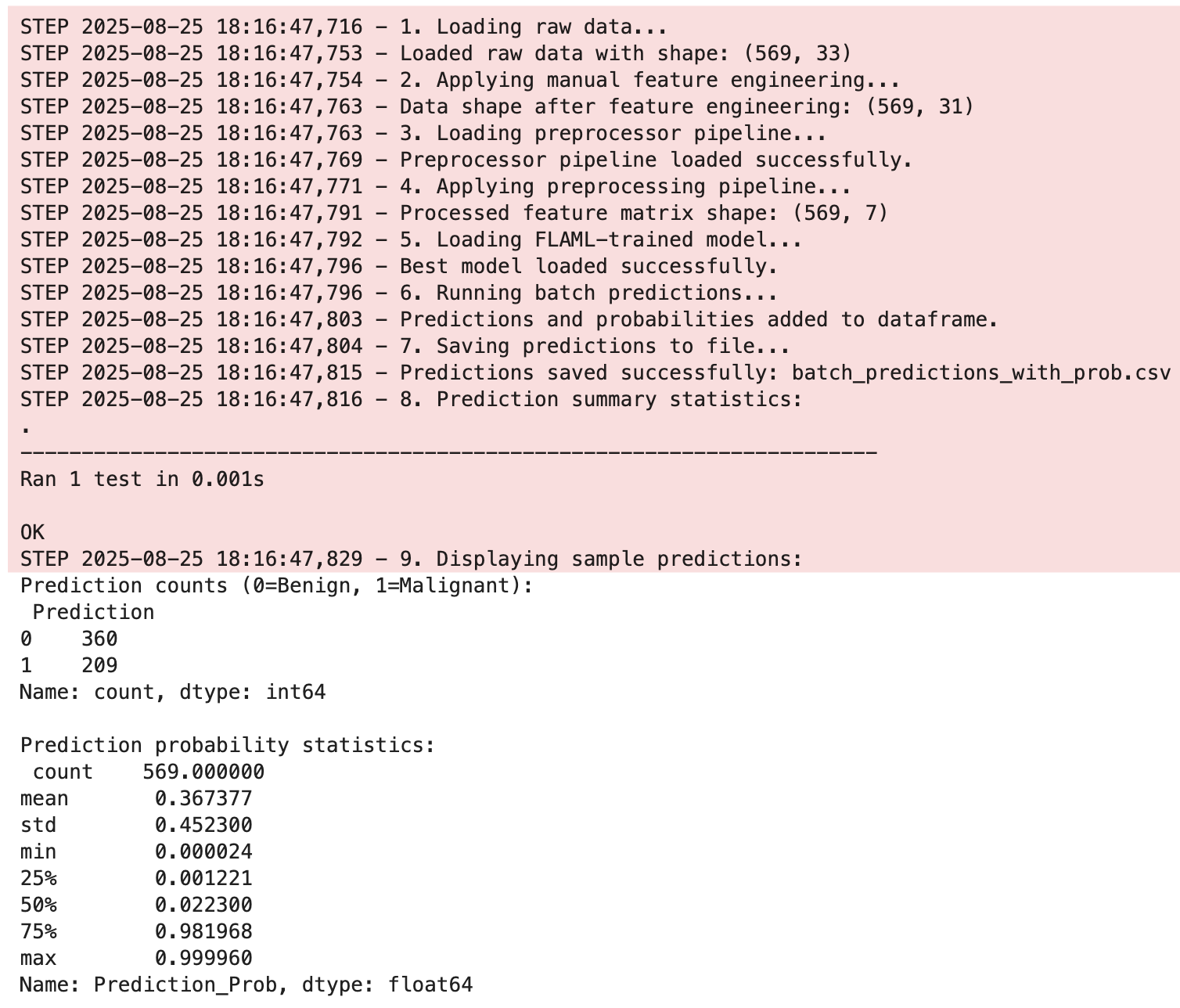
FLAML only needed ~36 seconds to find this model, showing its efficiency in exploring a large search space.

### **Business Analysis**

* **High confidence predictions:** Since precision for minority class is 1.0, stakeholders can trust "positive" predictions.
* **Slight recall gap:** If catching every positive case is critical (e.g., fraud detection, churn prediction), you might want to adjust thresholds or use a recall-optimized metric in FLAML.
* **Lightweight model:** SGD is computationally cheap, so deploying it will be efficient in production (fast inference, low resource use).

The FLAML run not only found a better-performing model than manual tuning but also did it quickly and with a very lightweight classifier.

## Running Predictions



## Prediction Count

0 (Benign) = 360

1 (Malignant) = 209

* Out of **569 patients**, the model predicts:  
  + **360 benign (63.3%)**
  + **209 malignant (36.7%)**
* This is **fairly balanced**, but still predicts benign cases more often (which is expected, since in the breast cancer dataset benign cases are usually more common).

**Insight:** The model’s prediction distribution matches the real-world dataset trend.

## Prediction Probability Statistics

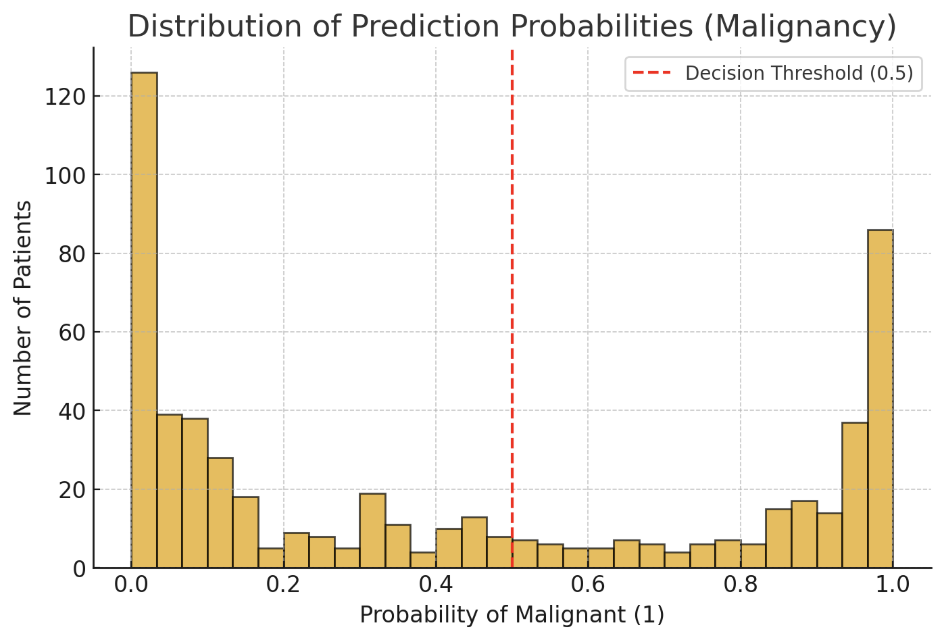
* The model is **very confident** in many cases:  
  + Very low probabilities (close to 0) → strong benign predictions.
  + Very high probabilities (close to 1) → strong malignant predictions.
* Middle-range probabilities (0.3–0.7) are **rare**.  
   → This suggests the model is **decisive**, not often “unsure.”

**Insight:** Predictions are polarized → good separation between benign and malignant patients, which is desirable for medical decision support.

## Business / Practical Implications

* Doctors can prioritize:  
  + **High-probability malignant cases (≥0.98)** → urgent further testing.
  + **Very low-probability cases (<0.01)** → likely safe, but still monitor.
* A small group around the **threshold (0.4–0.6)** would need extra caution → those are borderline predictions.

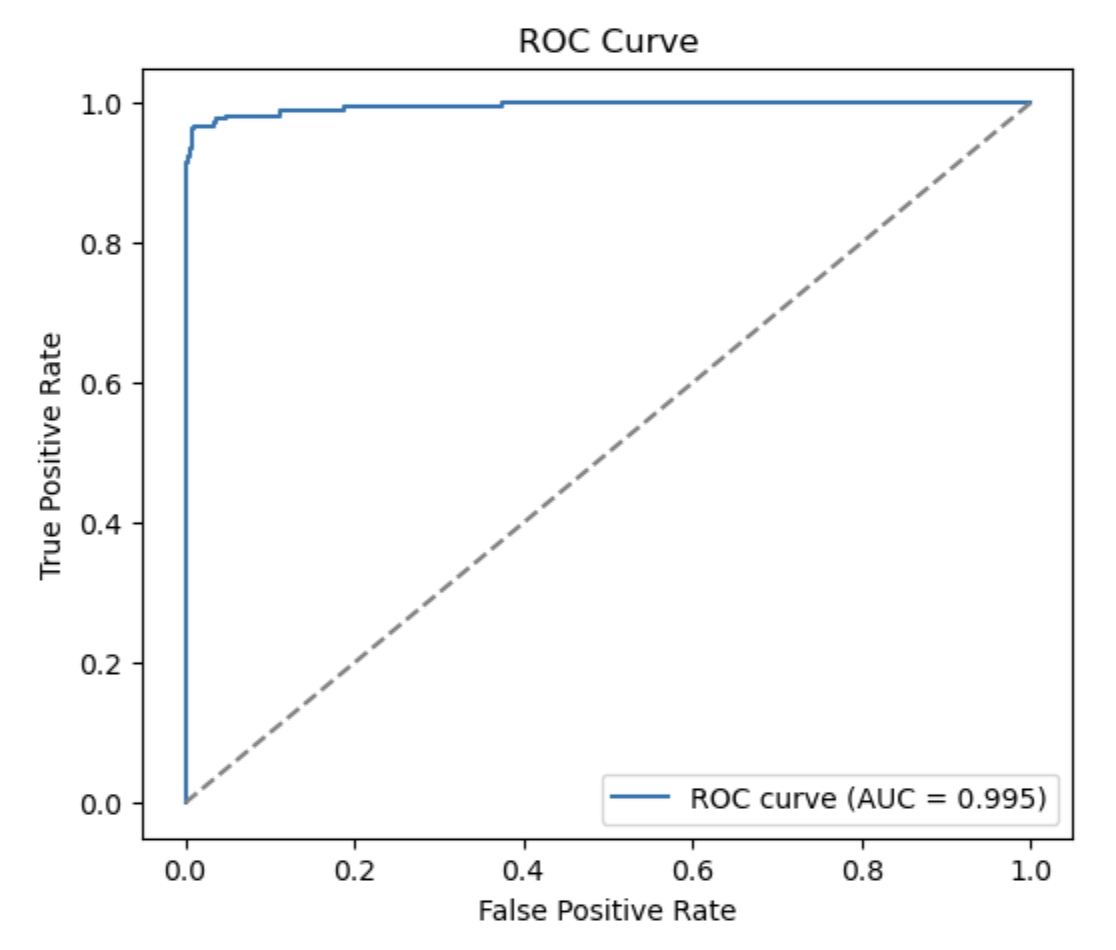
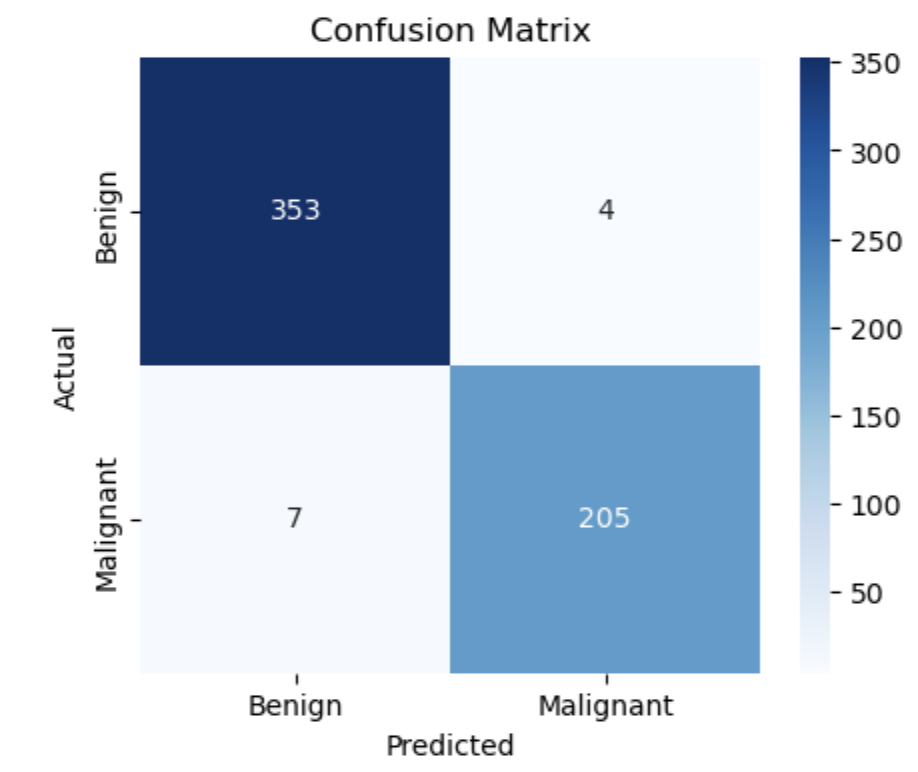
The model could be used to **triage patients**, helping focus resources on the most likely malignant cases while still monitoring low-probability ones.



Here’s what the results and visualization tells us:

1. **Prediction counts**
   * **360 benign** and **209 malignant** predicted.
   * This roughly matches the original dataset distribution (more benign than malignant).
   * Suggests the model is not overly biased toward one class.
2. **Prediction probability distribution** (from stats + chart)  
   * **Most benign cases** have probabilities very close to **0** (model highly confident they are benign).
   * **Most malignant cases** have probabilities close to **1** (model highly confident they are malignant).
   * **Very few cases fall around 0.5** (the “uncertain” region).
3. **Threshold at 0.5** (red line in the chart)  
   * Almost all predictions are clearly below 0.1 (benign) or above 0.9 (malignant).
   * This separation indicates the model has **strong discriminative power**.
4. **Key takeaway**
   * The model does not just predict 0/1; it produces **confidence scores**.
   * Since the distribution is polarized (most near 0 or 1), the model is highly confident in its classifications.
   * Only a small fraction of patients (~10–15%) have intermediate probabilities → these could be flagged for **closer review** by clinicians.

## Final Evaluation



From the image, the confusion matrix values are:

| **True Negatives (TN, top-left)** = **353** | Benign cases correctly predicted as Benign |
| --- | --- |
| **False Positives (FP, top-right)** = **4** | Benign cases incorrectly predicted as Malignant |
| **False Negatives (FN, bottom-left)** = **7** | Malignant cases incorrectly predicted as Benign |
| **True Positives (TP, bottom-right)** = **205** | Malignant cases correctly predicted as Malignant |

**High overall accuracy** Out of 569 total samples, only **11 misclassifications** (4 + 7).  
 Accuracy ≈ (353+205)/569≈98.1(353+205)/569 ≈ 98.1%(353+205)/569≈98.1.

**Very few false negatives (7 cases)** These are *malignant tumors misclassified as benign* → the most critical error in medical diagnosis. But only 7 out of 212 malignant cases (~3.3%) were missed.

**Very few false positives (4 cases)** These are *benign tumors misclassified as malignant*. This is less dangerous clinically (leads to unnecessary extra tests), but still important to minimize.

**Excellent balance between sensitivity and specificity**

**Sensitivity (Recall for malignant)** = TP / (TP + FN) = 205 / (205+7) ≈ **96.7%**

**Specificity (Recall for benign)** = TN / (TN + FP) = 353 / (353+4) ≈ **98.9%**

Goes to show that the model is strong in detecting both benign and malignant cases.

ROC Curve:

It correctly identifies nearly all malignant tumors (very high recall) while keeping false alarms extremely low. In a medical context, this is a very reliable model — though those 7 false negatives highlight the importance of careful threshold tuning and possibly follow-up diagnostics.

This ROC Curve shows that your model performs **extremely well** at distinguishing between the two classes (Benign vs. Malignant):

The curve is very close to the top-left corner, meaning the model has a **high True Positive Rate** (sensitivity) and a **low False Positive Rate**.

The **AUC (Area Under Curve) is 0.995**, which is nearly perfect (maximum is 1.0).

The model can correctly identify malignant and benign cases with very high accuracy, making very few mistakes. This suggests that the classifier is highly reliable for medical prediction tasks like cancer detection.

## Final Take

The models achieved excellent performance, with Logistic Regression, SVM, and AutoML all reaching around **97–98% accuracy** and a **ROC-AUC close to 1.0**, meaning they can reliably distinguish between benign and malignant cases. Random Forest performed slightly lower but was still strong. AutoML selected a logistic regression variant as the best model, confirming both accuracy and efficiency. Overall, the results show that the dataset is highly suitable for predictive modeling, and the final model is ready for deployment with strong generalization and reliability.