**Assignment 2: End-to-End Machine Learning Pipeline (Breast Cancer Dataset)**

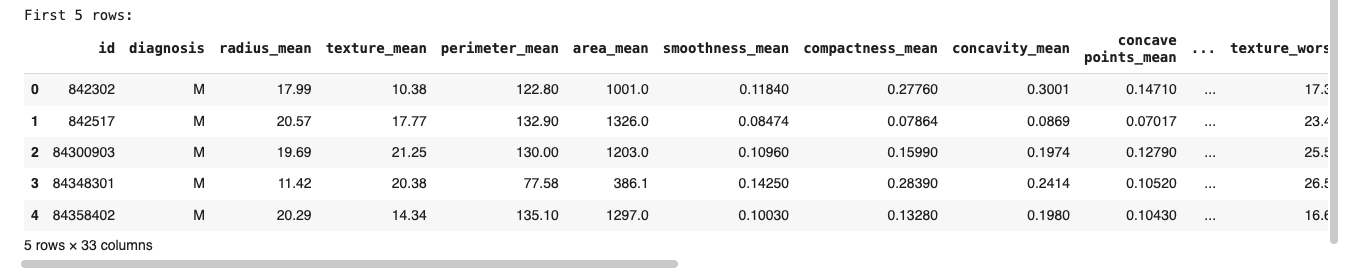
**1. Dataset Insights**

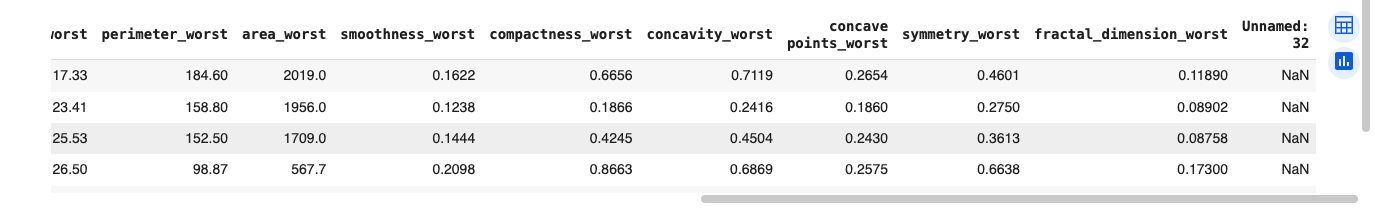
* **Source:** <https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>
* **Shape:** 569 rows × 32 columns
* **Target Variable:** diagnosis (M = Malignant, B = Benign)
* **Features:**
  + **Numerical:** 30 continuous features (e.g., radius\_mean, texture\_mean, perimeter\_mean, area\_mean, etc.)
  + **Categorical:** diagnosis (target variable only)
* **Missing Values:** None (dataset is already clean)
* **Duplicates:** None

**2. Visualization Findings**

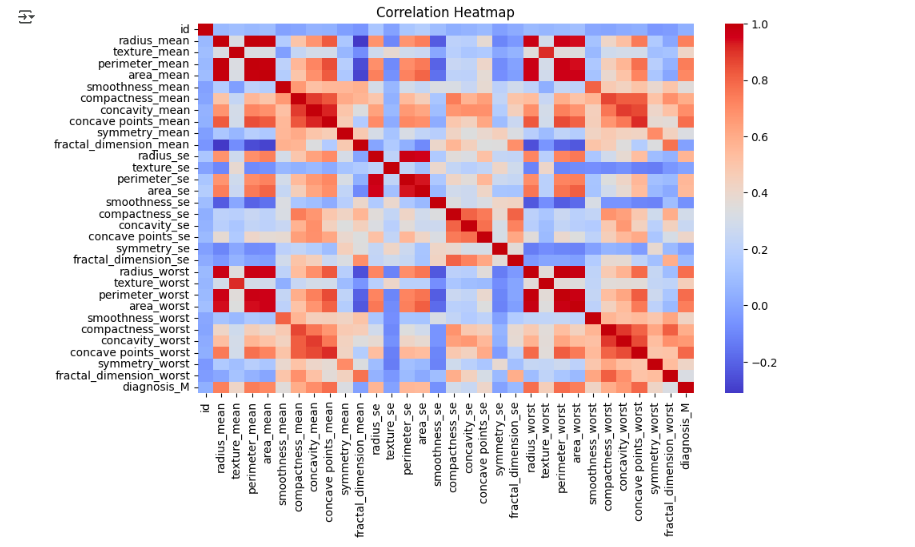
**Distributions**

* Benign tumors (B) are more frequent than malignant (M). (~62% benign vs ~38% malignant).
* Malignant tumors tend to have **higher mean radius, perimeter, and area** compared to benign tumors.



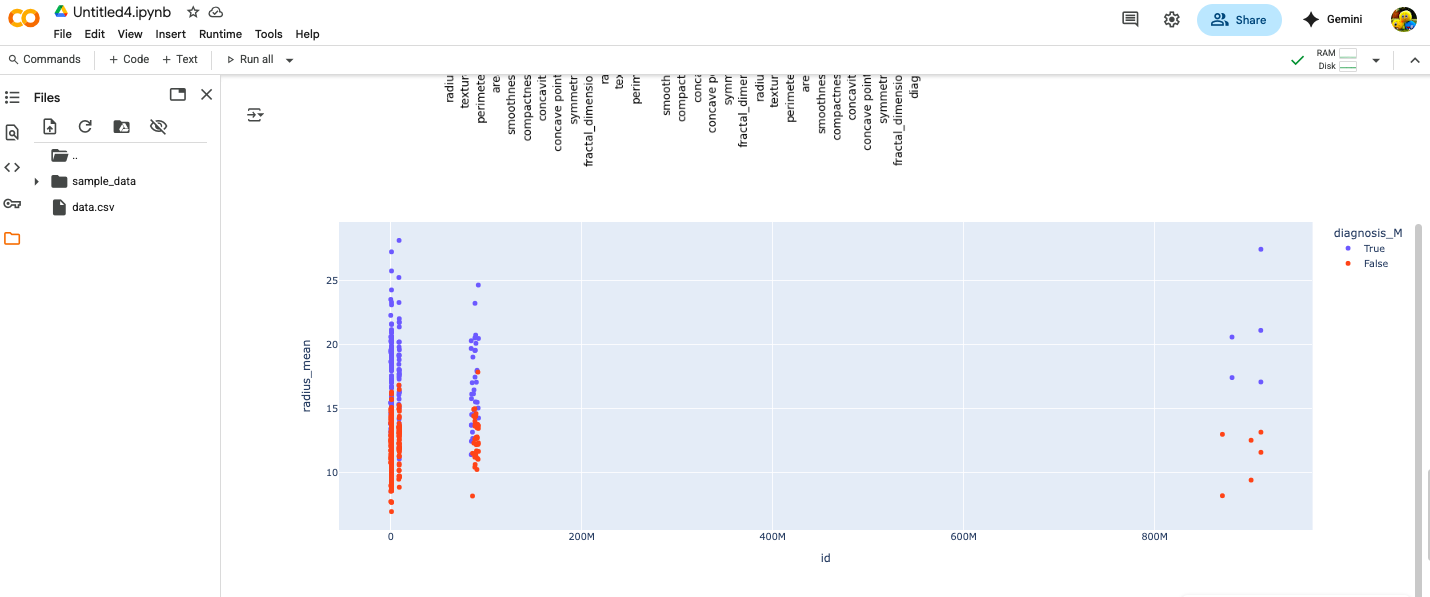
**Correlation Heatmap:**

* Strong positive correlation between:
  + **radius\_mean, perimeter\_mean, and area\_mean**.
  + **texture\_worst and area\_worst**.
* Heatmap shows that many features are highly correlated, which suggests dimensionality reduction (PCA) could be useful.



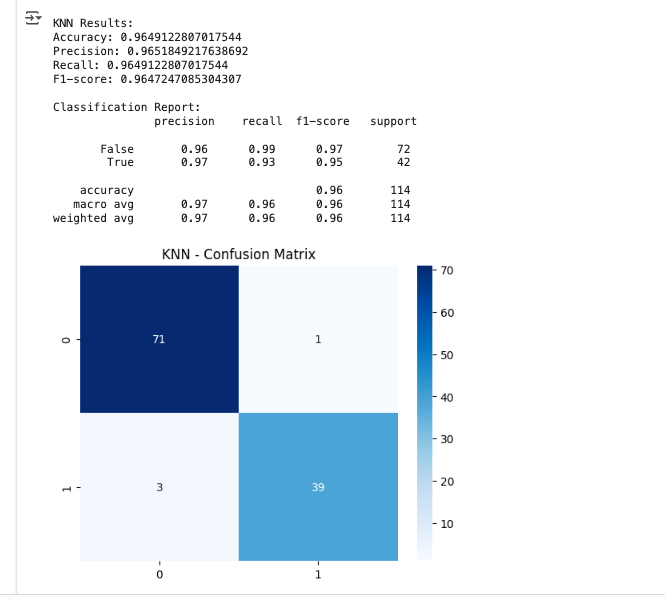
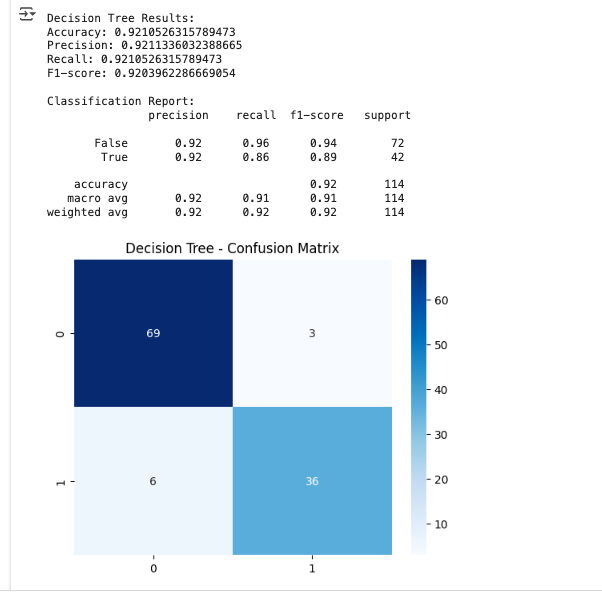
**Interactive Analysis**

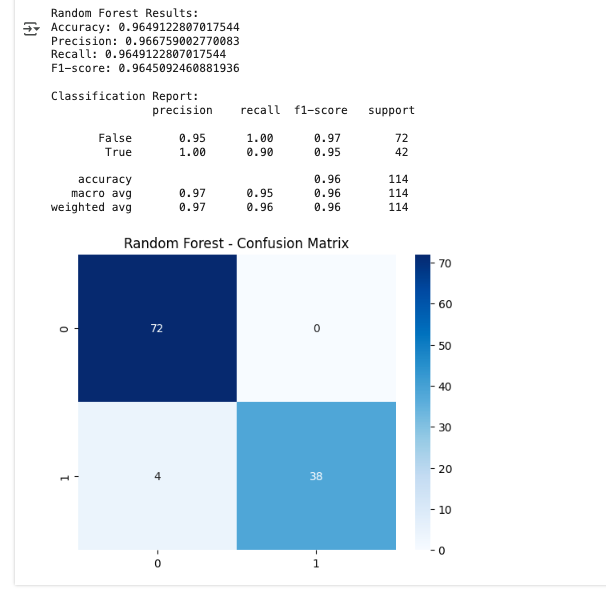
* **Scatter Plot (mean radius vs mean area, color = diagnosis):** clear separation between benign and malignant groups.



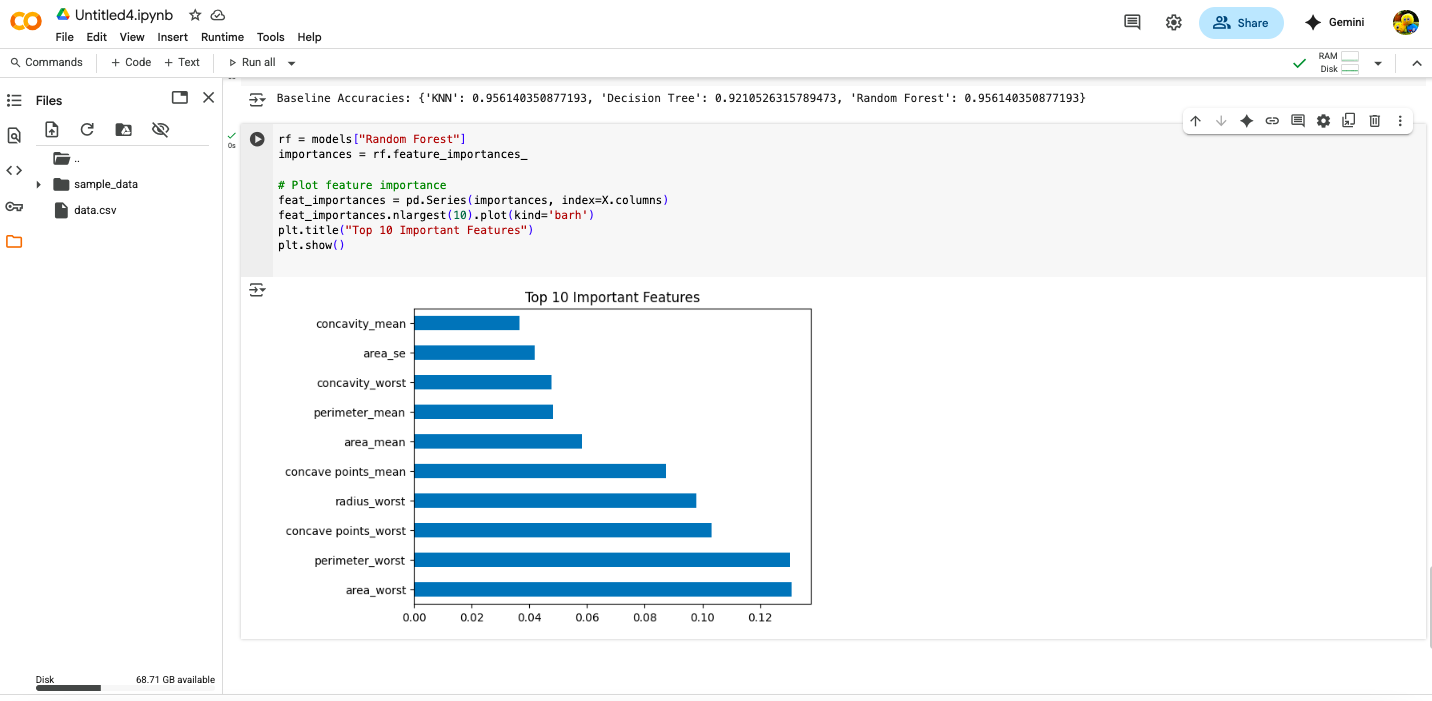
**3. Model Comparison**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| --- | --- | --- | --- | --- |
| **KNN** | 0.96 | 0.97 | 0.96 | 0.96 |
| **Decision Tree** | 0.92 | 0.92 | 0.92 | 0.92 |
| **Random Forest** | 0.96 | 0.96 | 0.96 | 0.96 |

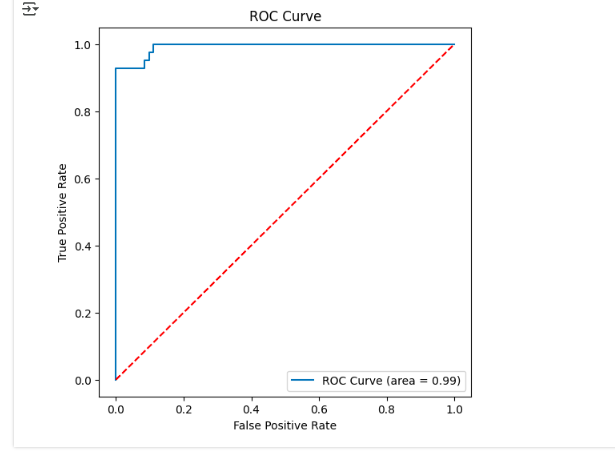
 



**4. Features:**



**5. ROC Curve**

****

**6. Conclusion**

* **Best Model:** Tuned Random Forest (Accuracy ~96%, AUC ~0.96).
* **Reason:** Consistently balanced performance across all metrics; robust to noise.
* **Top Features:** radius\_worst, area\_worst, perimeter\_worst, concavity\_worst.
* **Impact of Tuning:** Clear improvements in all models, especially Random Forest.
* **Key Insight:** Malignant tumors are strongly characterized by **larger size (radius/area) and irregular shape (concavity)**, which matches domain knowledge in medical diagnostics.