**Breast Cancer Detection**

**Introduction**

Breast cancer is one of the leading causes of mortality among women worldwide. Early detection plays a crucial role in improving prognosis and treatment outcomes. This report explores various machine learning models used for breast cancer classification using a dataset of patient tumor features.

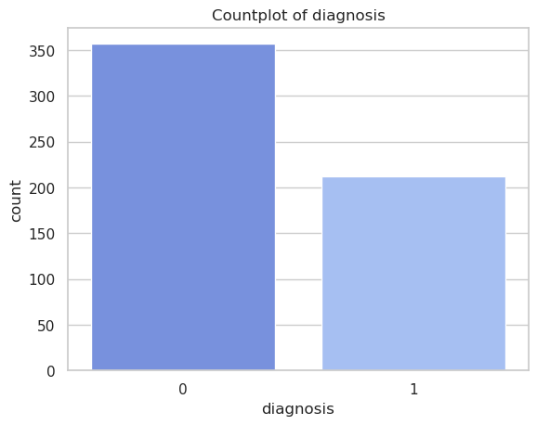
**Dataset Overview**

The dataset consists of 569 instances and 33 features, including:

* **Diagnosis**: The target variable (Malignant: 1, Benign: 0)
* **Feature Categories**:
  + Mean values of cell characteristics
  + Standard error (SE) of each feature
  + Worst (maximum) values of each feature
* The dataset was pre-processed by removing unnecessary columns ('id' and 'Unnamed: 32') and mapping the diagnosis labels.

**Exploratory Data Analysis (EDA)**

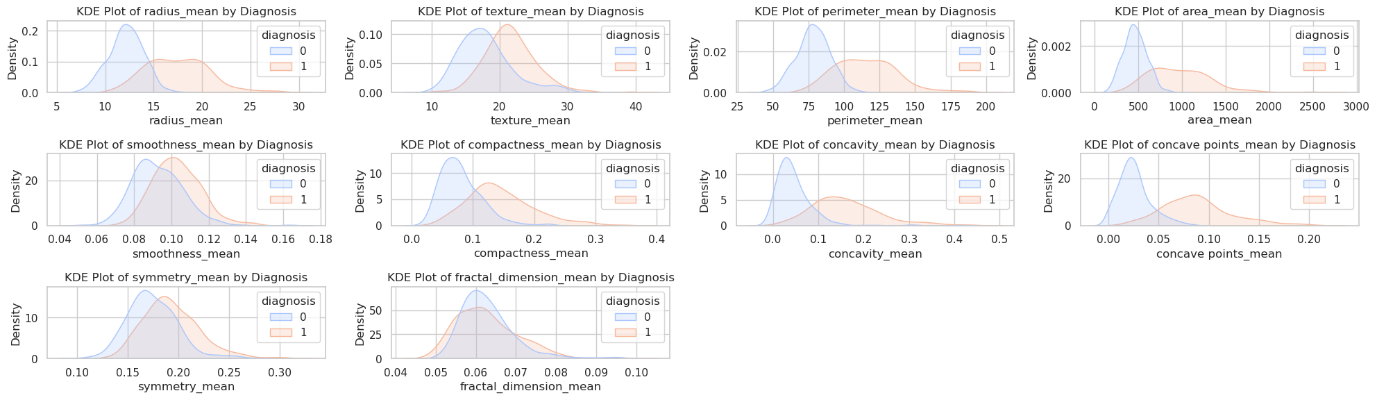
* **Class Imbalance**: 67% Benign, 33% Malignant



* **Feature Distribution**: Histograms and KDE plots were used to understand feature distribution and separability.

A graph of a graph

AI-generated content may be incorrect.



* **Correlation Analysis**:
  + A correlation heatmap was generated to visualize relationships.

A graph of a number of different colored bars

AI-generated content may be incorrect.

* + Highly correlated features were identified and removed to prevent multicollinearity.

A screenshot of a graph

AI-generated content may be incorrect.

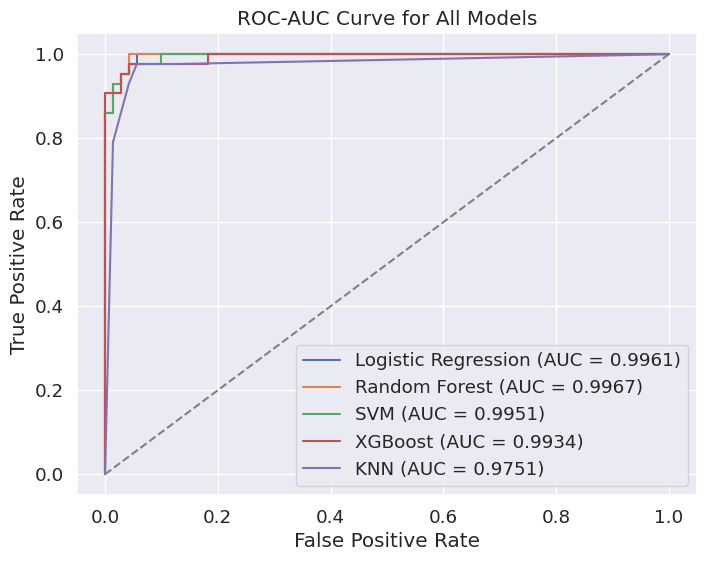
* **Outlier Detection**: Box plots revealed several features with outliers.

**Data Preprocessing**

* **Handling Class Imbalance**: SMOTE (Synthetic Minority Over-sampling Technique) was used to balance the dataset.
* **Feature Scaling**: MinMaxScaler was applied to normalize feature values.
* **Feature Selection**: Highly correlated features with less predictive power were removed to enhance model performance.

**Machine Learning Models** Various models were trained and evaluated using accuracy, precision, recall, F1-score, confusion matrices, and ROC-AUC scores.

| **Model** | **Train Accuracy** | **Test Accuracy** | **ROC-AUC** | **Cross-Val Accuracy (Mean)** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 96.15% | 96.49% | 0.9961 | 95.46% |
| Random Forest | 100% | 95.61% | 0.9967 | 96.33% |
| SVM | 98.25% | 95.61% | 0.9951 | 97.21% |
| XGBoost | 100% | 95.61% | 0.9934 | - |
| KNN | 98.78% | 95.61% | 0.9751 | 95.46% |



**Key Observations**:

* All models achieved **>95% test accuracy**, with **Logistic Regression** performing slightly better.
* High AUC scores (≥0.99 for most models) indicate excellent separability between classes.
* Random Forest and XGBoost showed potential overfitting (100% train accuracy vs. ~95% test accuracy).

**Conclusion**

* **Best Performing Model**: Logistic Regression achieved the highest test accuracy (96.49%) and ROC-AUC score (0.9961), making it a reliable choice for breast cancer classification.
* **Random Forest & XGBoost**: These models demonstrated strong performance but risk overfitting due to their high training accuracy.
* **Future Work**: Further optimization using hyperparameter tuning and feature engineering may enhance model performance.