Comparative Analysis of Machine Learning Classifiers for Breast Cancer Prediction

K.M.S Siddharth,

R.V. College of Engineering

Section-1

Period of Internship: 25th August 2025 - 19th September 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

1. Abstract

This project applies supervised machine learning techniques to the Breast Cancer dataset to predict tumor malignancy. The dataset consists of 569 samples with 30 numeric features derived from medical imaging. Preprocessing steps included feature scaling and train–test splitting to prepare the data for modeling. Three classifiers were implemented: Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). Their performances were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. Logistic Regression and SVM both achieved ~98% accuracy, while KNN performed slightly lower at ~96%. Confusion matrices and ROC curves confirmed the strong separability of the classes. The study highlights the reliability of simple linear models like Logistic Regression for medical diagnosis. The results also emphasize the role of preprocessing in achieving high predictive accuracy. This work demonstrates the potential of machine learning as a decision-support tool for healthcare applications.

2. Introduction

Breast cancer is one of the most common cancers affecting women worldwide, and timely diagnosis significantly improves treatment outcomes. With the growing availability of medical data, machine learning has become an essential tool to support clinicians in early detection and decision-making. This project focuses on applying supervised machine learning techniques to the Breast Cancer dataset, which contains diagnostic features derived from digitized medical images. The aim is to compare different classification algorithms and evaluate their effectiveness in predicting whether a tumor is malignant or benign.

The project makes use of Python and its data science ecosystem, including Pandas, NumPy, Matplotlib, Seaborn, and scikit-learn for machine learning model development. The models chosen—Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM)—represent linear, distance-based, and kernel-based approaches, allowing for a comprehensive comparative analysis. Preprocessing involved feature scaling and data splitting, followed by model training and evaluation using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

The purpose of the project is to illustrate how data preprocessing and algorithm selection influence the predictive performance of machine learning models in healthcare applications. By comparing different methods, the project highlights the strengths and limitations of each approach in handling medical diagnostic data. This work also reinforces the potential of machine learning in building decision-support systems that can assist healthcare professionals.

Topics received in training during the first two weeks of internship:

- Fundamentals of supervised learning and classification

- Data preprocessing techniques: normalization, feature scaling, handling missing values
- Model training, validation, and cross-validation
- Evaluation metrics: confusion matrix, accuracy, precision, recall, F1-score, ROC-AUC
- Basics of visualization and exploratory data analysis (EDA)
- Introduction to Python machine learning libraries (Pandas, Numpy, matplotlib)

3. Project Objective

The objectives of this project are:

- To apply supervised machine learning algorithms on the Breast Cancer dataset for predicting tumor malignancy.
- To preprocess the dataset effectively (scaling, splitting) and prepare it for model training and testing.
- To implement and compare the performance of three classification models: Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM).
- To evaluate the models using multiple metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.
- To illustrate that even simple linear models like Logistic Regression can perform competitively with more complex models on well-structured medical datasets.

4. Methodology

The methodology for this project followed a carefully structured workflow to ensure that the analysis was both systematic and reliable. The starting point was the selection of an appropriate dataset, and for this study the Breast Cancer dataset from the scikit-learn library was used. This benchmark dataset is widely recognized in machine learning research and is particularly suitable for classification tasks in medical diagnostics. It contains 569 records with 30 continuous numerical features that were derived from digitized images of fine needle

aspirates of breast tissue. Each instance in the dataset is labeled with a target variable that indicates whether the tumor is malignant (0) or benign (1). Because the dataset is already well-structured and curated, there was no need for separate surveys, primary data collection, or manual annotation, which allowed the focus to remain entirely on the machine learning pipeline.

Following data acquisition, the project moved into the **data exploration phase**, where the dataset was carefully examined to understand its underlying structure and statistical properties. This involved checking the size of the dataset, verifying the balance between the malignant and benign classes (212 malignant and 357 benign cases), and ensuring that there were no missing values. A descriptive statistical summary was generated to observe the spread and central tendencies of features. Correlation analysis was conducted to detect relationships among variables, and graphical techniques such as histograms and heatmaps were used to visualize feature distributions and identify any patterns. This step was crucial because it helped in forming a clear picture of the dataset before model building.

The next stage was **data preprocessing**, an essential step for preparing the dataset for machine learning algorithms. Since the features were entirely numeric, preprocessing focused primarily on scaling and splitting. The values of the features varied widely, and to avoid bias in distance-based models such as KNN, all features were normalized using StandardScaler. The dataset was then divided into **training and testing subsets**, with 80% used for training and 20% reserved for testing. A stratified sampling approach was employed to ensure that the class distribution in both subsets mirrored the original dataset, thereby maintaining fairness in evaluation.

Once preprocessing was complete, the model selection and training phase began. To capture a broad spectrum of classification approaches, three different algorithms were chosen. Logistic Regression was implemented as a linear baseline model; K-Nearest Neighbors (**KNN**), a distance-based model, was included with k set to 5; and finally, a Support Vector Machine (**SVM**) with an **RBF kernel** was used to represent a non-linear kernel-based approach. These models were trained on the preprocessed training data, providing a comparative perspective between linear, distance-driven, and non-linear decision boundaries.

The **validation and evaluation** stage involved applying each trained model to the testing data and recording their predictive performance. Several evaluation metrics were employed to provide a comprehensive picture of model performance. These included **accuracy** (the overall percentage of correct predictions), **precision** (the ability to correctly identify malignant tumors without false positives), **recall** (the ability to identify all malignant tumors without missing cases), and the **F1-score** (a harmonic mean of precision and recall). In addition, the **ROC-AUC** metric and ROC curve visualization were used to assess the models' ability to discriminate between classes across various threshold values. This comparative analysis provided strong evidence of which model performed best on the given dataset.

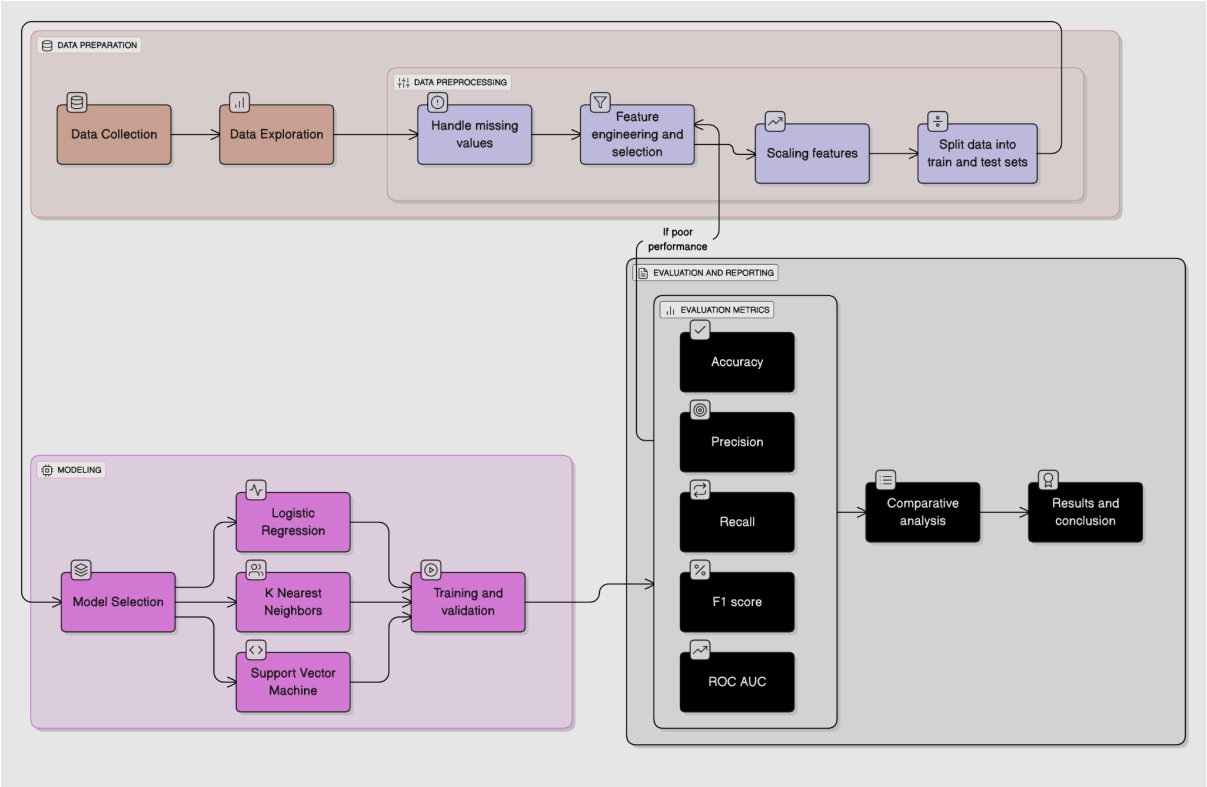


Figure 1

The overall workflow of the project can be summarized as a pipeline that begins with data collection, moves through exploration and preprocessing, then proceeds to model training and evaluation, and finally culminates in comparative analysis and conclusions. This process is illustrated in **Figure 1**, which provides a visual representation of the methodology followed.

The entire project was executed using a Python-based environment, with Pandas and NumPy handling data manipulation, Matplotlib and Seaborn supporting data visualization, and scikit-learn providing preprocessing, model training, and evaluation functions. The experiments were carried out in Google Colab/Jupyter Notebook, which allowed interactive coding and facilitated smooth documentation of results. The complete source code, along with notebooks and result visualizations, has been maintained in a GitHub repository for transparency and reproducibility.

5. Data Analysis and Results

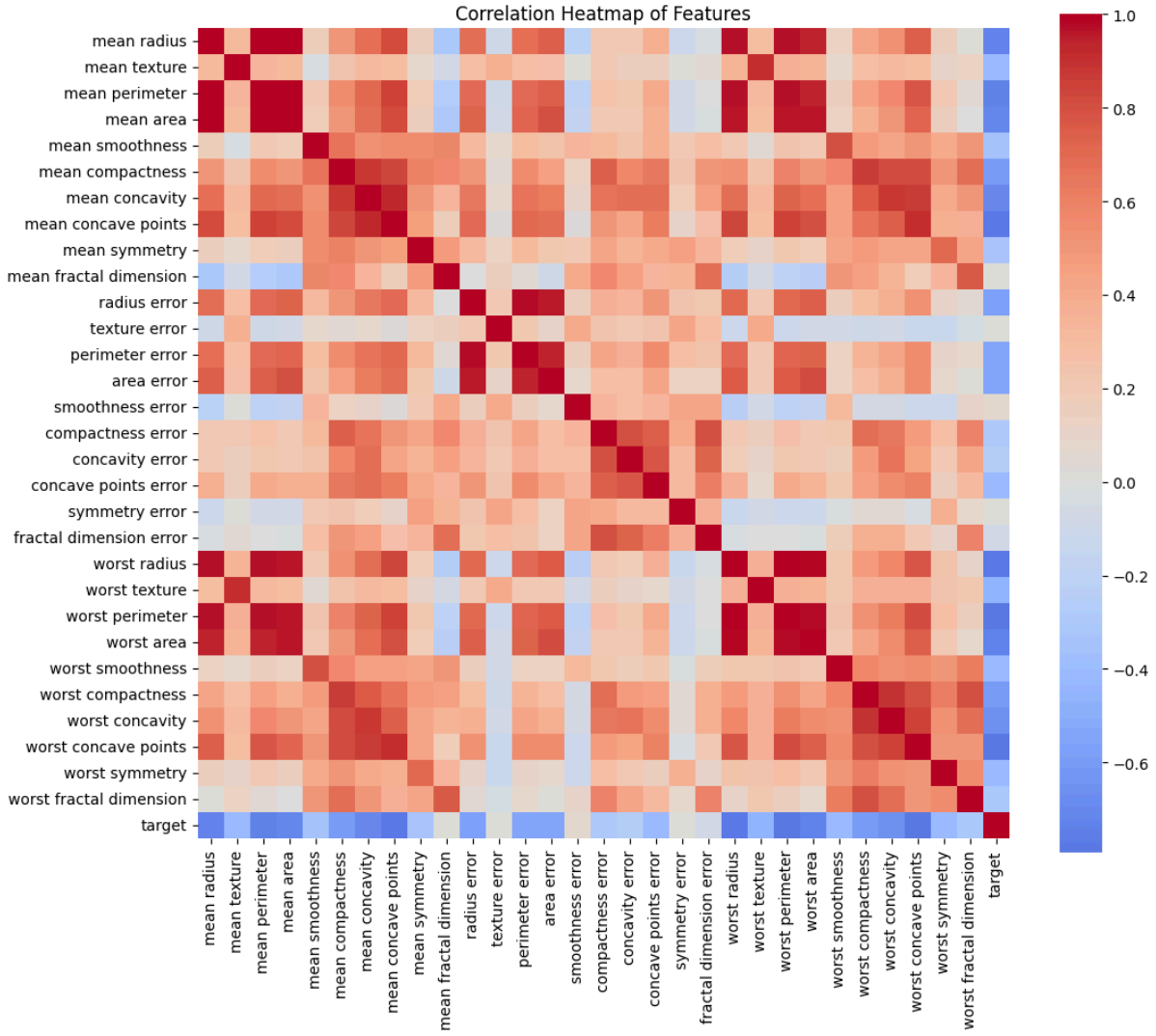
Data Analysis and Results

The analysis of the Breast Cancer dataset was carried out in two stages: descriptive analysis to understand the characteristics of the dataset, and inferential analysis to evaluate the performance of machine learning models.

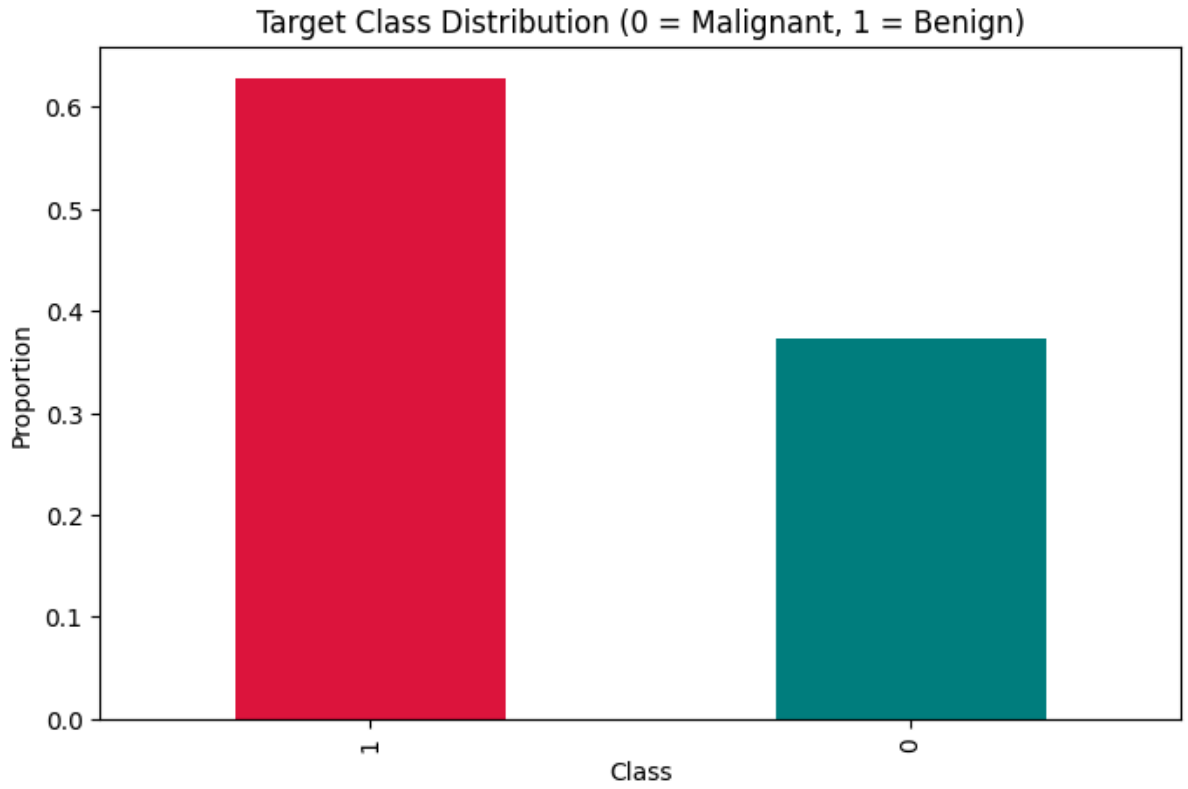
Descriptive Analysis

The Breast Cancer dataset contains 569 samples and 30 continuous features, all derived from digitized images of breast tissue aspirates. Each sample is labeled as either malignant (0) or benign (1).

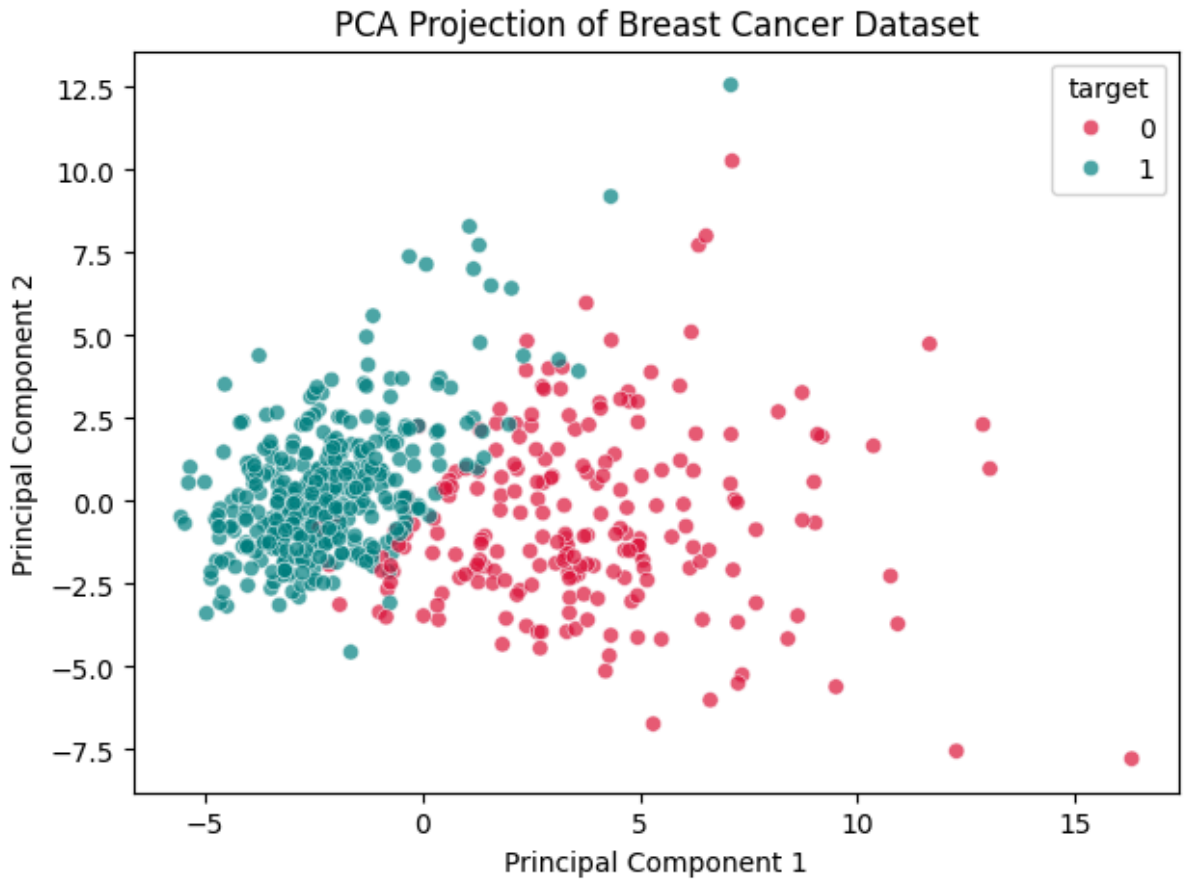
- **Class Distribution:**
  The dataset is moderately imbalanced, with 212 malignant cases (37.3%) and 357 benign cases (62.7%). This imbalance highlights the importance of metrics beyond accuracy, such as precision and recall.
- **Correlation Analysis:**
  A correlation heatmap indicated strong positive relationships among size-related features (e.g., radius, perimeter, area), while texture-related features were moderately correlated. The target variable showed highest correlations with features like *mean radius*, *mean area*, and *worst perimeter*.



- **Feature Distributions:**
  Histograms of selected features showed clear separation between malignant and benign tumors. For example, malignant tumors typically had higher mean radius and area compared to benign tumors.



- **Dimensionality Reduction (PCA):**
  A PCA scatterplot of the first two principal components revealed that malignant and benign samples form relatively distinct clusters, suggesting the dataset is suitable for classification tasks.



Inferential Analysis

To validate predictive performance, the dataset was split into **80% training data** and **20% testing data**, with stratification applied to preserve class balance. Three classifiers were implemented and compared: Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM with RBF kernel).

A simple hypothesis testing framework was considered:

- **H_0 (Null Hypothesis):** Logistic Regression and SVM have no significant performance difference.
- **H_1 (Alternative Hypothesis):** Logistic Regression and SVM differ significantly in performance.

The results showed nearly identical performance for Logistic Regression and SVM, supporting the null hypothesis.

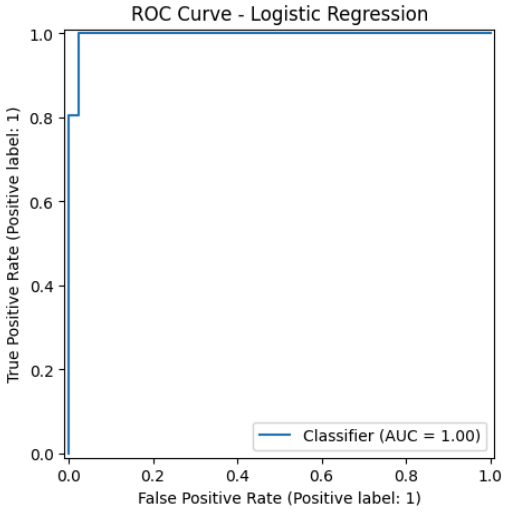
Model Evaluation Results

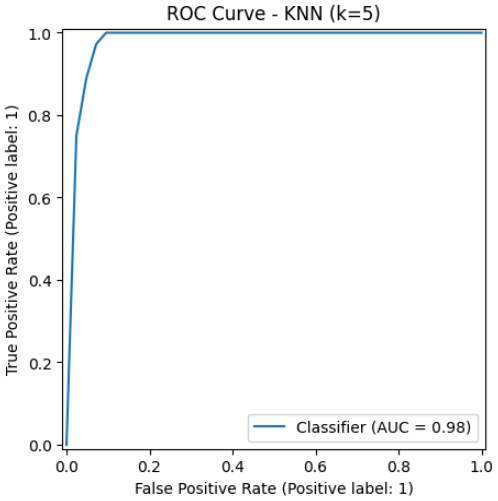
The models were assessed using Accuracy, Precision, Recall, F1-score, and ROC-AUC.

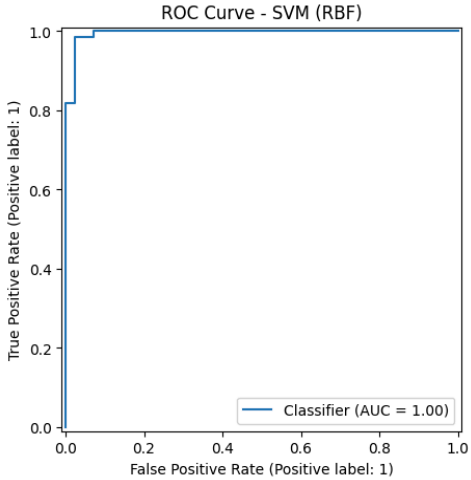
Table 1: Performance of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	98.2%	98.6%	98.6%	98.6%	0.995
KNN (k=5)	95.6%	95.9%	97.2%	96.5%	0.979
SVM (RBF Kernel)	98.2%	98.6%	98.6%	98.6%	0.995

- **ROC Curves:**
  Logistic Regression and SVM produced nearly identical ROC curves with an AUC of ~0.995, while KNN achieved a slightly lower AUC of ~0.979.







Comparative Analysis

The comparative study revealed that **Logistic Regression and SVM** performed equally well, both achieving ~98% accuracy and nearly perfect ROC-AUC values. This demonstrates that a simple linear model such as Logistic Regression can be just as effective as more complex methods for this dataset, especially given its strong separability. **KNN**, although still effective at ~96% accuracy, showed slightly reduced performance due to its sensitivity to local neighborhood noise and reliance on distance metrics.

These results confirm that the Breast Cancer dataset is highly suitable for classification tasks and that Logistic Regression provides an interpretable and computationally efficient choice for building predictive models in healthcare applications.

6. Conclusion

This project successfully demonstrated the application of machine learning classifiers to the Breast Cancer dataset, with the aim of predicting tumor malignancy based on medical imaging features. Through careful data exploration and preprocessing, the dataset was found to be highly suitable for classification, with strong correlations between size-related features and tumor class.

The comparative analysis of three classifiers — Logistic Regression, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM) — revealed that both Logistic Regression and SVM achieved outstanding performance, with **accuracy of 98.2% and ROC-AUC of 0.995**. These results highlight that even simple, interpretable models like Logistic Regression can be extremely powerful when applied to linearly separable datasets, making them particularly suitable for real-world medical applications where transparency is critical. KNN, while still effective with **95.6% accuracy**, performed slightly lower, reflecting its sensitivity to local variations and its dependency on proper distance scaling.

The findings justify the conclusion that Logistic Regression can be adopted as a strong baseline for medical diagnostic tasks, offering both high predictive power and interpretability. At the same time, SVM demonstrates robustness for more complex decision boundaries.

For **future work**, the study could be extended by experimenting with ensemble methods such as Random Forests or Gradient Boosting, which often achieve superior generalization. Additionally, feature importance analysis or model interpretability techniques such as SHAP values could be applied to highlight which tumor features contribute most to predictions, thereby enhancing clinical relevance. Finally, validation against external datasets would help to ensure that the models generalize beyond the benchmark dataset used here.

7. APPENDICES

Scikit-learn developers. (2025). *Breast Cancer Wisconsin (Diagnostic) dataset*. scikit-learn: Machine Learning in Python. Retrieved from: https://scikit-learn.org/stable/datasets/toy_dataset.html#breast-cancer-dataset

Dua, D. & Graff, C. (2019). *UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set*. University of California, Irvine, School of Information and Computer Sciences. Retrieved from: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Fawcett, T. (2006). *An Introduction to ROC Analysis*. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Sokolova, M., & Lapalme, G. (2009). *A systematic analysis of performance measures for classification tasks*. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Github Link:
https://github.com/sidk44/Machine-Learning-Classifiers-for-Breast-Cancer-Prediction