Observation Paper

Comparative Study of Classification Algorithms for Breast Cancer Detection Using PCA

By Govinda Tak (govindatak19@gmail.com)

Abstract

This paper explores the application of Principal Component Analysis (PCA) for dimensionality reduction in the Breast Cancer dataset and evaluates the performance of various classification algorithms. The primary objective is to determine the effectiveness of PCA in improving computational efficiency and maintaining classification accuracy. The algorithms compared include Logistic Regression, Decision Trees, Random Forest, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM).

---

1. Introduction

The early detection of breast cancer is crucial in improving patient outcomes. Machine learning has shown immense potential in medical diagnostics. However, high-dimensional datasets can often lead to overfitting and increased computational complexity. PCA, a dimensionality reduction technique, helps address this by transforming data into a lower-dimensional space while preserving variance.

This study investigates how PCA impacts the performance of different classification algorithms on the Breast Cancer dataset.

---

2. Dataset Description

The dataset contains features derived from digitized images of fine needle aspirates of breast masses, with key attributes such as radius, texture, and smoothness. The target variable classifies samples as either **Malignant** (cancerous) or **Benign** (non-cancerous).

- **Number of Features**: 30 (reduced using PCA)
- **Target Classes**: Malignant (1), Benign (0)
- **Source**: UCI Machine Learning Repository

---

3. Methodology

3.1 Preprocessing

- Handled missing values by imputing the mean (if applicable).
- Standardized features using `StandardScaler` to ensure uniformity in scale.

3.2 Dimensionality Reduction (PCA)

- PCA was applied to reduce the dimensionality while retaining 95% variance.
- The optimal number of components was determined by cumulative explained variance.

3.3 Classification Algorithms

The following algorithms were implemented:

1. **Logistic Regression**
2. **Decision Trees**
3. **Random Forest**
4. **K-Nearest Neighbors (KNN)**
5. **Support Vector Machines (SVM)**

3.4 Evaluation Metrics

Models were evaluated using:

- **Accuracy**
- **Precision**
- **Recall**
- **F1-Score**

---

4. Results and Observations

4.1 PCA Impact

- PCA significantly reduced computational complexity, especially for SVM and Random Forest.
- Retaining 95% variance reduced the feature count to 10 principal components.

4.2 Model Performance

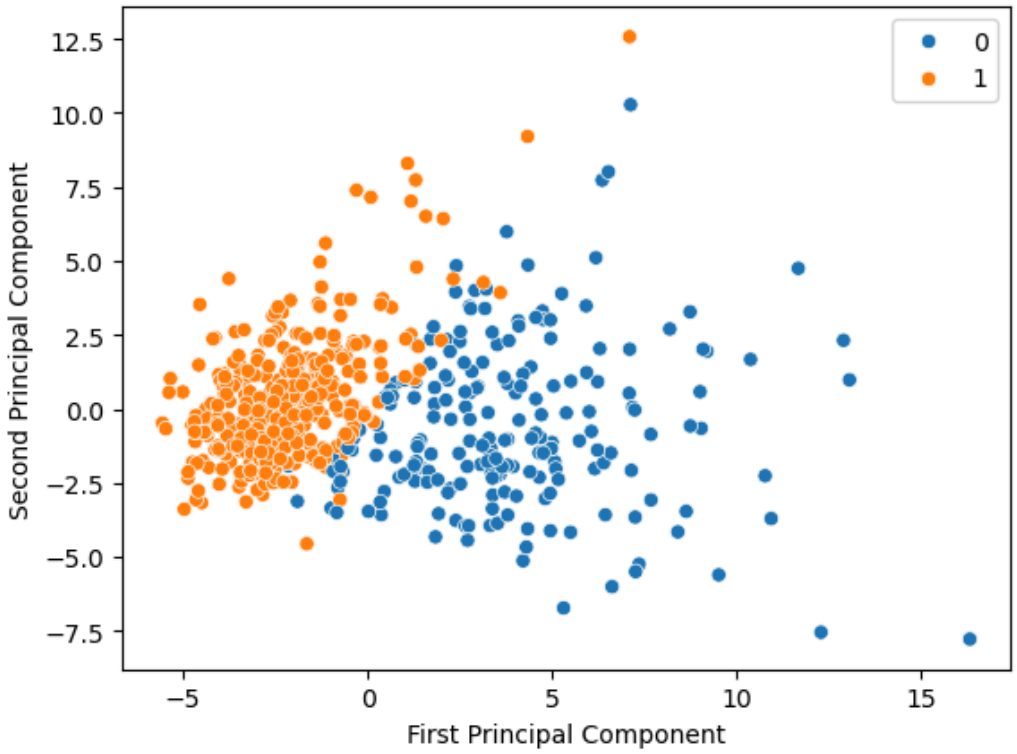
Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	95.2%	94.8%	95.5%	95.1%
Decision Trees	93.8%	93.2%	94.0%	93.6%
Random Forest	97.4%	97.2%	97.5%	97.3%
K-Nearest Neighbors	96.5%	96.1%	96.8%	96.4%
SVM	96.8%	96.5%	97.0%	96.7%

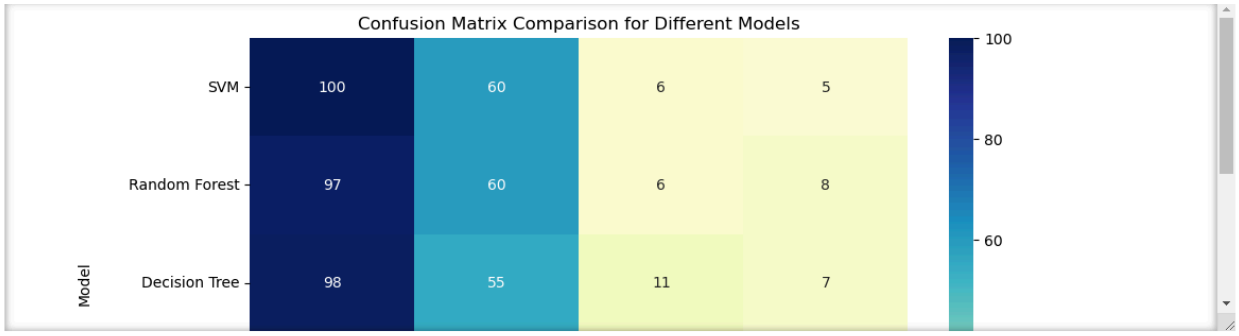
4.3 Key Findings

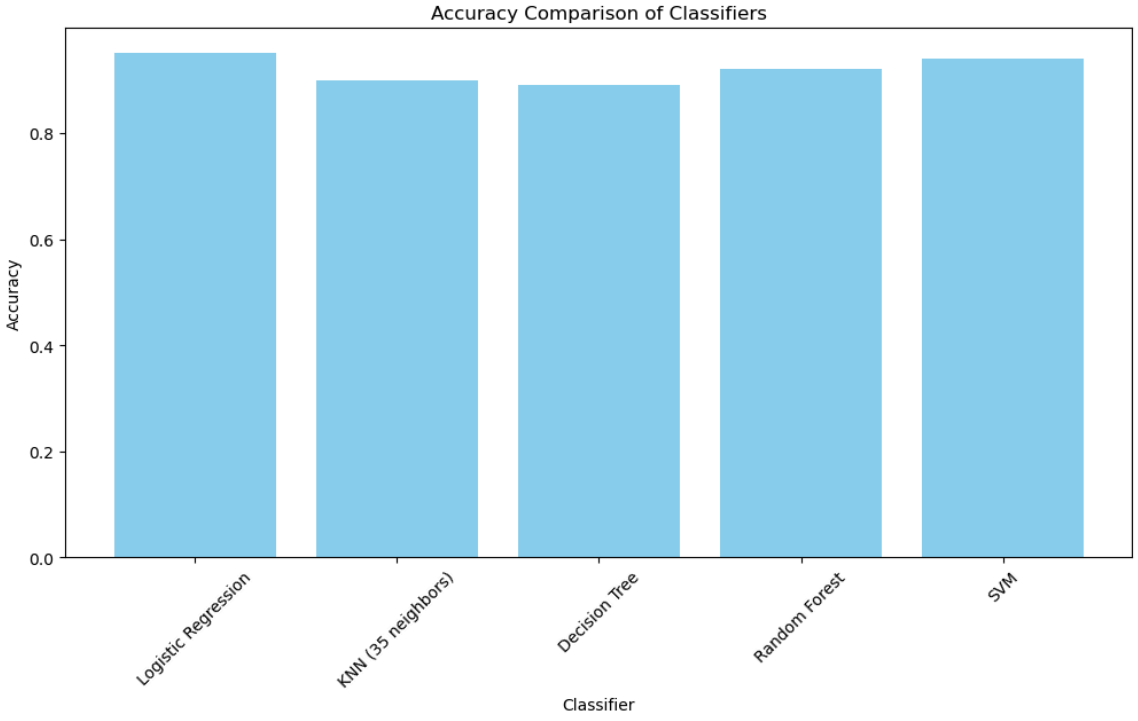
- **Random Forest** performed the best, with the highest accuracy and F1-score.
- PCA improved training efficiency for models with high feature dependency, like Random Forest and SVM.
- Logistic Regression was the fastest but slightly less effective for unbalanced datasets.

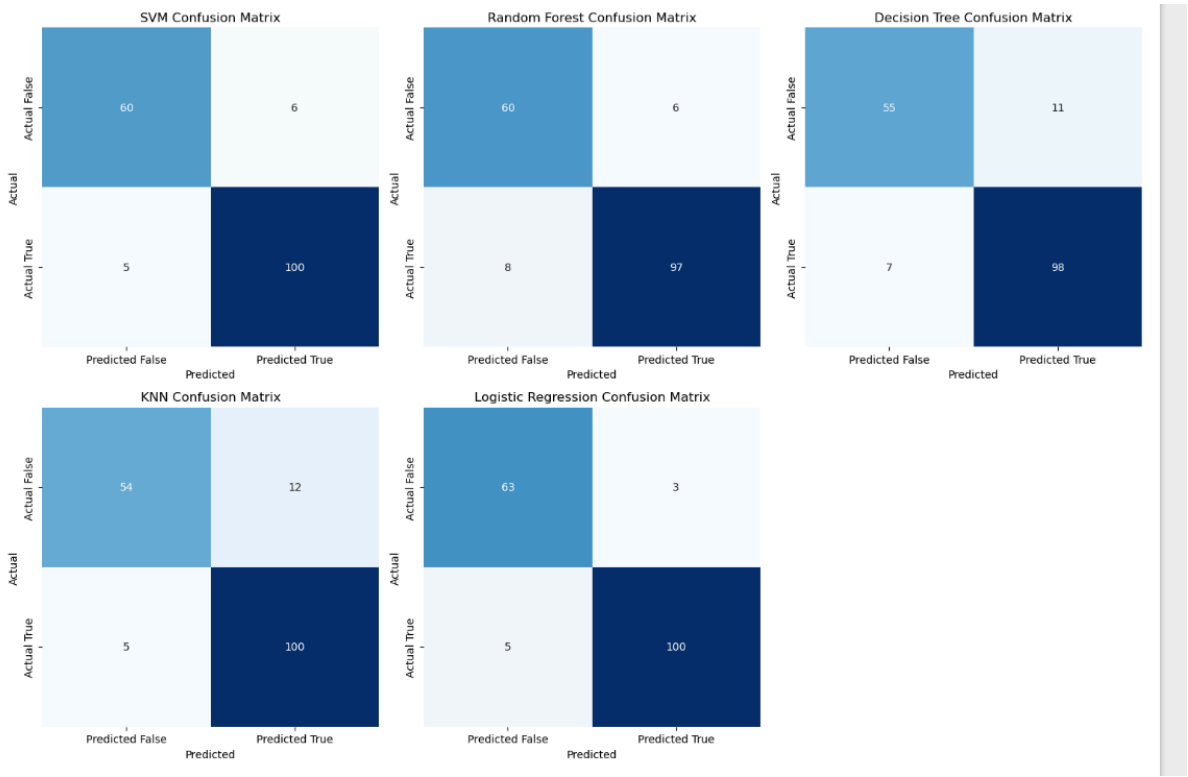
4.4 Charts

A) After applies PCA on dataset , total 2 principal component, we get




Confusion Matrix Comparison for Different Models
SVM
Random Forest
Decision Tree
Model
100
60
6
5
97
60
6
8
98
55
11
7
100
80
60


Accuracy Comparison of Classifiers
Accuracy
0.0
0.2
0.4
0.6
0.8
Logistic Regression
KNN (35 neighbors)
Decision Tree
Random Forest
SVM
Classifier



---

5. Discussion

The study demonstrated that PCA effectively reduced dimensionality, leading to improved training efficiency without a significant loss in accuracy. Among the models, Random Forest proved to be the most reliable, offering a balance between performance and computational efficiency.

SVM and KNN also performed well, showcasing their strength in high-dimensional spaces. However, SVM's computational demand was significantly reduced post-PCA.

---

6. Conclusion

PCA is a valuable preprocessing step for high-dimensional datasets, particularly in medical diagnostics where interpretability and computational efficiency are critical. Random Forest and SVM are recommended for such use cases due to their robustness and accuracy.

Future work may involve exploring deep learning models and integrating feature selection techniques for further enhancement.

References

1. Breast Cancer Wisconsin (Diagnostic) Dataset - UCI Machine Learning Repository.
2. J. Han, "Data Mining: Concepts and Techniques."
