Report on Breast Cancer Detection using Machine Learning

Skill Based Assignment INT234

(PREDICTIVE ANALYTICS)

Submitted to: Amanpal Singh Rayat

Submitted by: Talent Tapiwanashe Mapara

Name: Breast Cancer Detection using Machine Learning

Reg. No: 12209467

Roll. No: A07

Breast Cancer Detection using Machine Learning

Abstract / Declaration

This project presents the application of **Machine Learning classification algorithms** to the detection of breast cancer using the **Breast Cancer Wisconsin (Diagnostic) Dataset**.
Two supervised learning models — **Support Vector Machine (SVM)** and **Decision Tree Classifier** — were developed to classify tumors as **benign** or **malignant** based on 30 diagnostic features derived from fine needle aspirate (FNA) images.

The study involves comprehensive **data preprocessing**, **exploratory data analysis**, **model training**, and **evaluation using statistical metrics and visualization techniques**.
Results demonstrate that both models achieved strong predictive performance, with SVM yielding slightly superior generalization accuracy and Decision Tree offering interpretability through feature visualization.

This work highlights the potential of machine learning in medical diagnostics, contributing toward early detection systems capable of assisting oncologists in clinical decision-making.

Introduction

Background

Breast cancer is the most commonly diagnosed cancer among women and remains one of the leading causes of cancer-related deaths worldwide. Timely diagnosis is crucial for effective treatment, as early detection significantly improves patient outcomes.

Conventional diagnostic approaches — such as mammography, biopsy, and histopathological analysis — though effective, can be resource-intensive, time-consuming, and sometimes prone to human interpretation errors.
In recent years, **Machine Learning (ML)** has emerged as a powerful tool to support medical professionals by **automating pattern recognition** in diagnostic data.

The **Breast Cancer Wisconsin dataset**, developed by Dr. William H. Wolberg at the University of Wisconsin Hospitals, provides an ideal testbed for exploring ML-based diagnosis. It contains 569 instances with 30 numerical attributes describing cell nucleus features extracted from microscopic images.

Motivation

The goal of this project is to leverage ML algorithms to automatically classify tumors based on diagnostic features. The comparative study of **Support Vector Machines** (a high-dimensional classifier) and **Decision Trees** (an interpretable rule-based model) helps balance **accuracy** and **explainability** — two essential components of real-world healthcare AI systems.

Implementation

The implementation consists of six major stages, following the classical machine learning workflow.

---

1. Dataset Description

The dataset used is the **Breast Cancer Wisconsin (Diagnostic) dataset**, obtained from Kaggle:
👉 *https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data*

- **Number of samples:** 569
- **Number of features:** 30
- **Target variable:** diagnosis (M = malignant, B = benign)

Features include statistical properties of cell nuclei — *mean radius, texture, smoothness, compactness, symmetry,* and *fractal dimension* — derived from digital image analysis.

---

2. Data Cleaning

Initial preprocessing included:

- Removing non-predictive columns (id, Unnamed: 32)
- Converting the diagnosis column into binary values (M → 1, B → 0)
- Checking for missing or null values (none were found)

This ensured that only relevant, clean, and numerically formatted data was passed to the model.

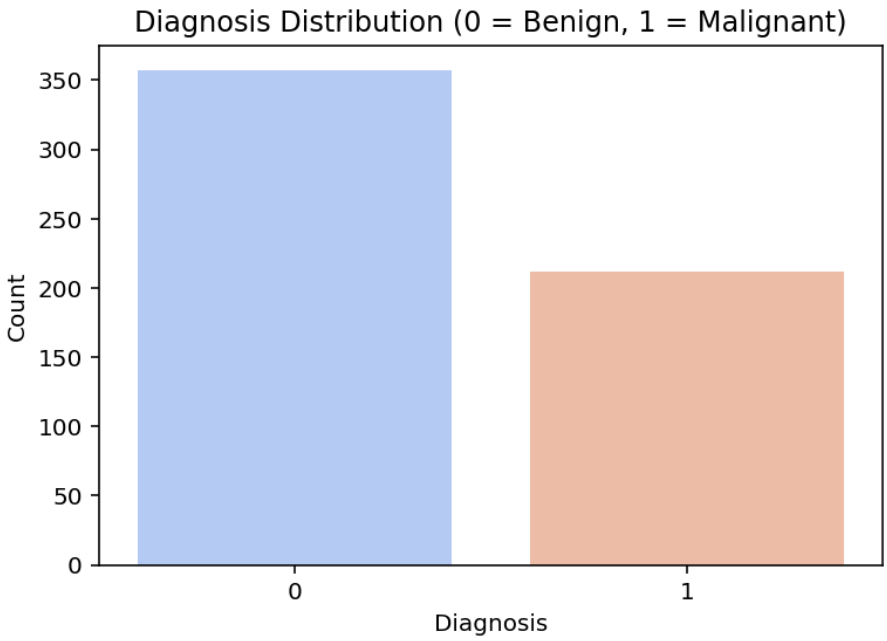
---

3. Exploratory Data Analysis (EDA)

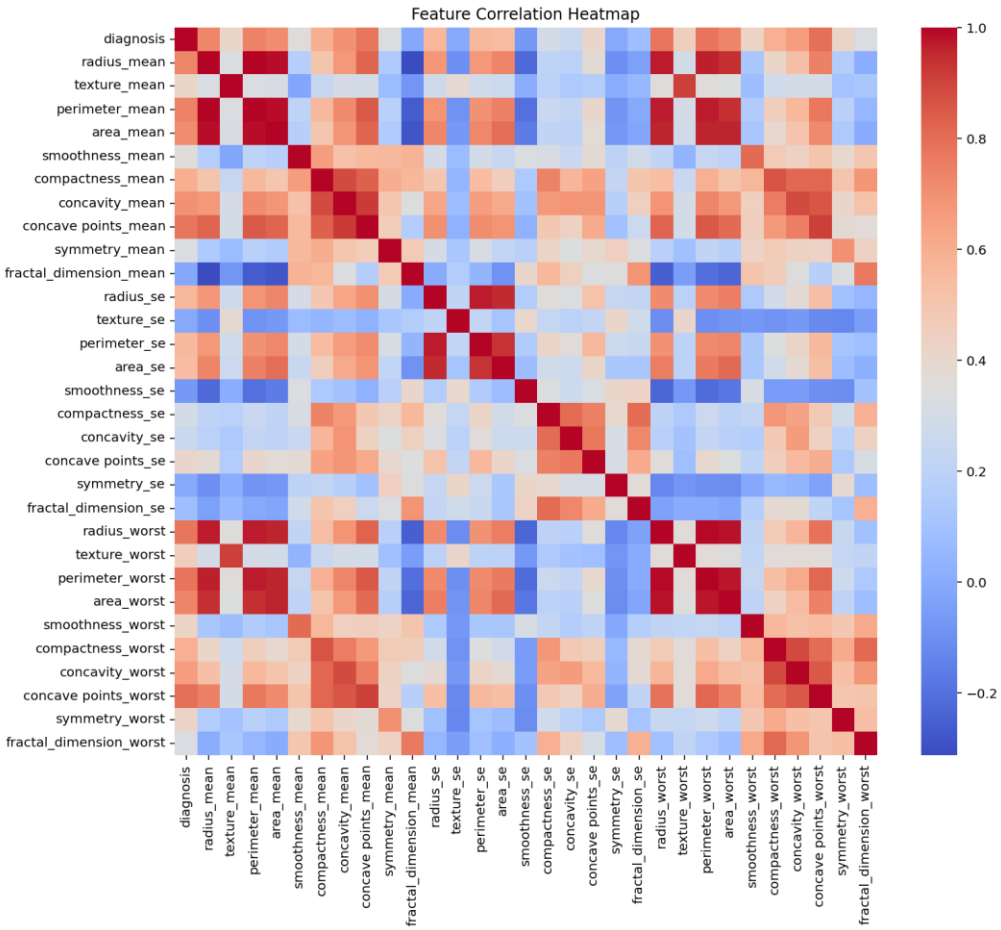
EDA was performed to understand class distribution and correlation among variables.
Visual exploration aided in identifying feature redundancy and the balance between malignant and benign cases.

Visuals Included:

- **Diagnosis Distribution Plot** — shows the relative frequency of malignant and benign samples.



- **Correlation Heatmap** — visualizes relationships among all 30 features, helping detect highly correlated attributes.



Statistical summaries were generated using df.describe(), revealing significant variance across features, indicating strong discriminative potential.

---

4. Feature Scaling and Splitting

Data was divided into training (80%) and testing (20%) subsets.
Because SVMs are sensitive to feature magnitude, standardization using **StandardScaler** was applied.
Decision Trees, being scale-invariant, were unaffected but benefited from consistent feature ranges.

---

5. Model Development

Two supervised learning models were implemented:

(a) Support Vector Machine (SVM)

A **linear SVM** was used to find the optimal separating hyperplane between classes.
The objective is to maximize the margin between malignant and benign data points while minimizing classification error.
Parameters:

- Kernel: Linear
- Regularization: C = 1.0
- Probability estimation enabled for ROC-AUC analysis

(b) Decision Tree Classifier

A **Decision Tree** model was trained with a maximum depth of 4 to prevent overfitting.
It recursively splits features based on information gain (Gini index), resulting in a human-interpretable rule-based structure.

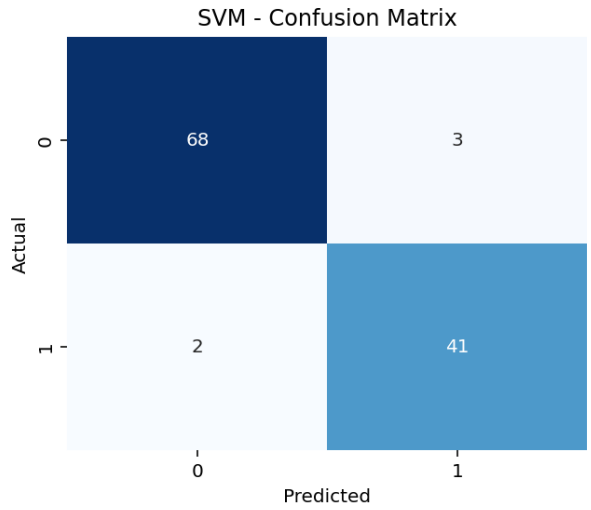
---

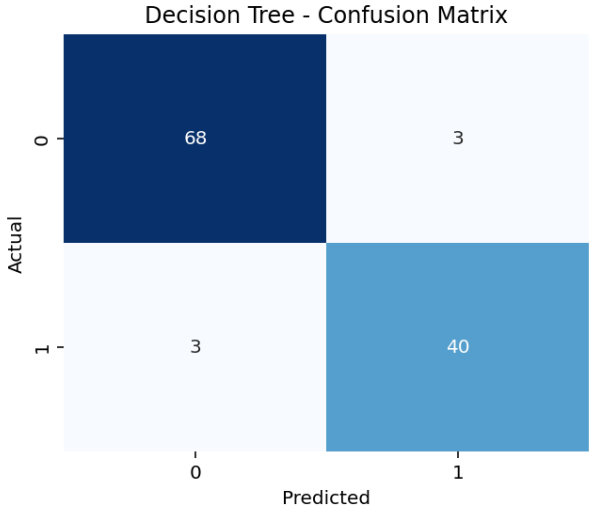
6. Model Evaluation and Visualization

Each model was assessed using multiple performance metrics and visual tools.

a) Confusion Matrices

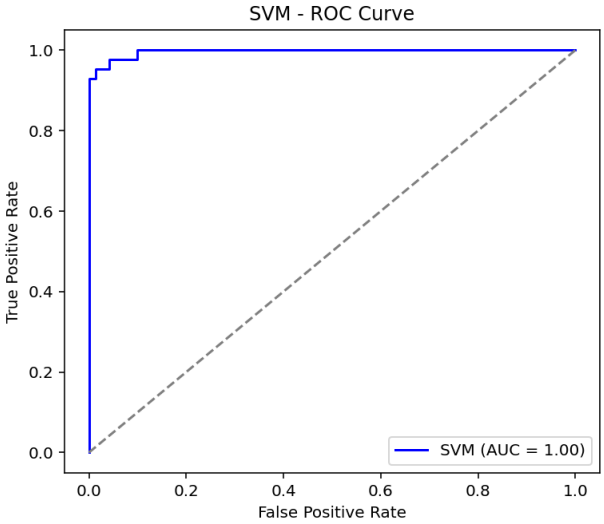
Display the true vs predicted classifications for both models.

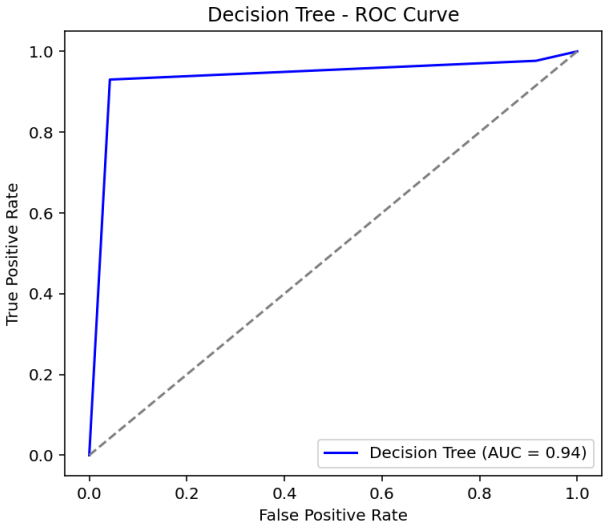




b) ROC Curves

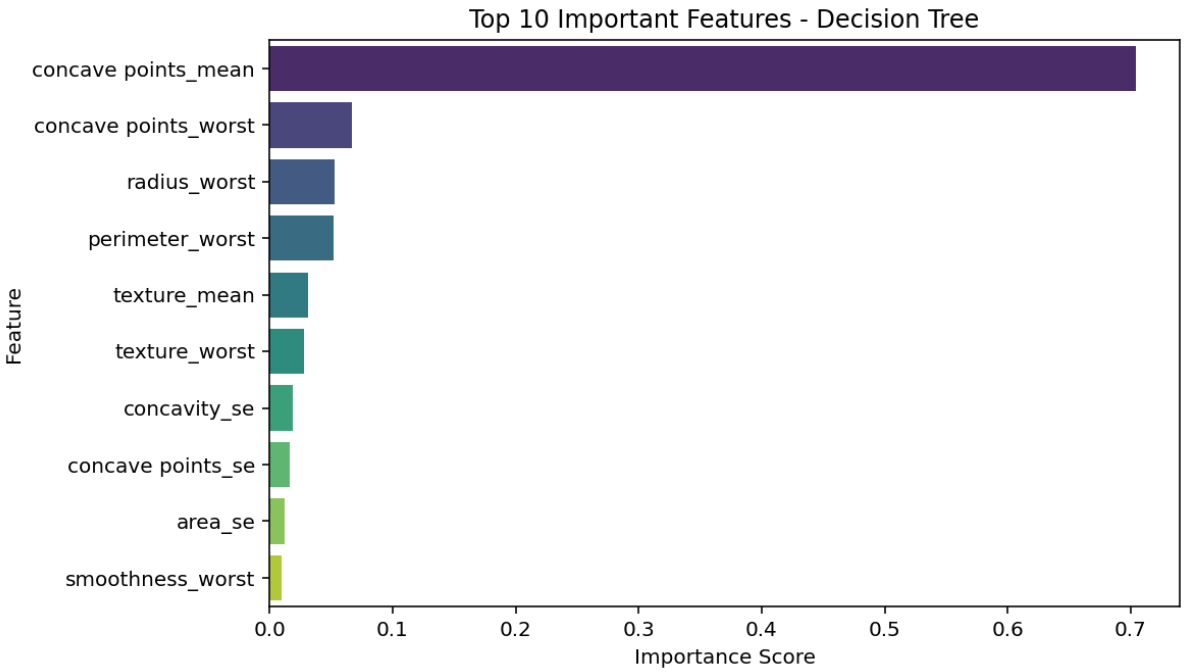
Illustrate each model's ability to distinguish between classes.
Both SVM and Decision Tree achieved **AUC values above 0.95**, indicating high separability.

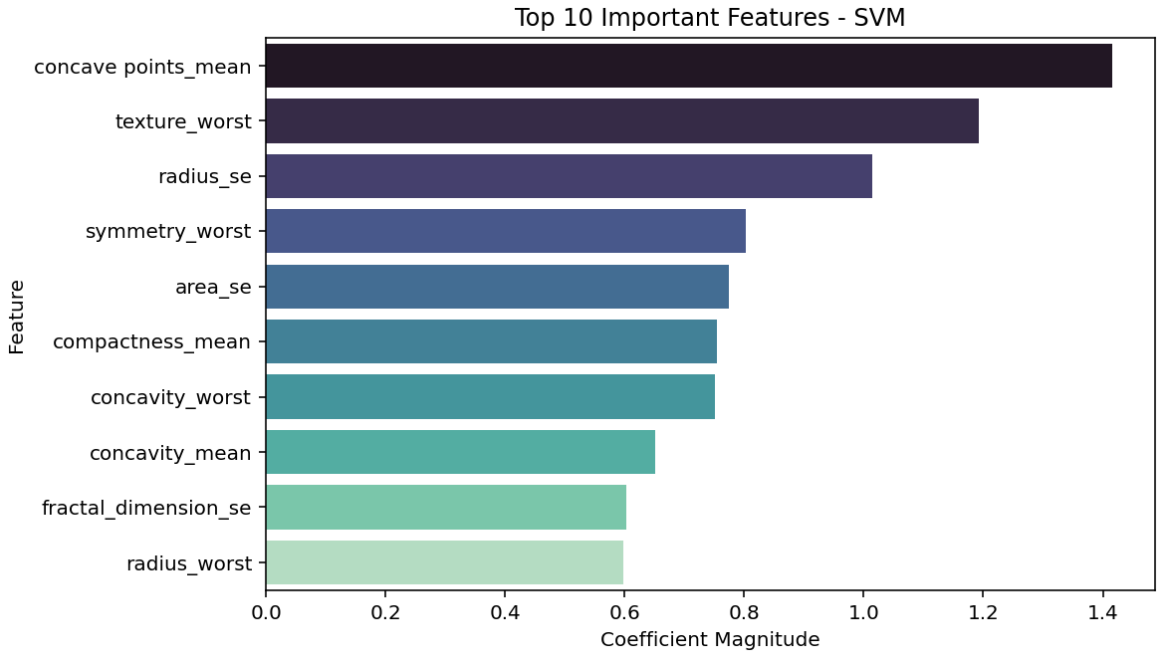




c) Feature Importance

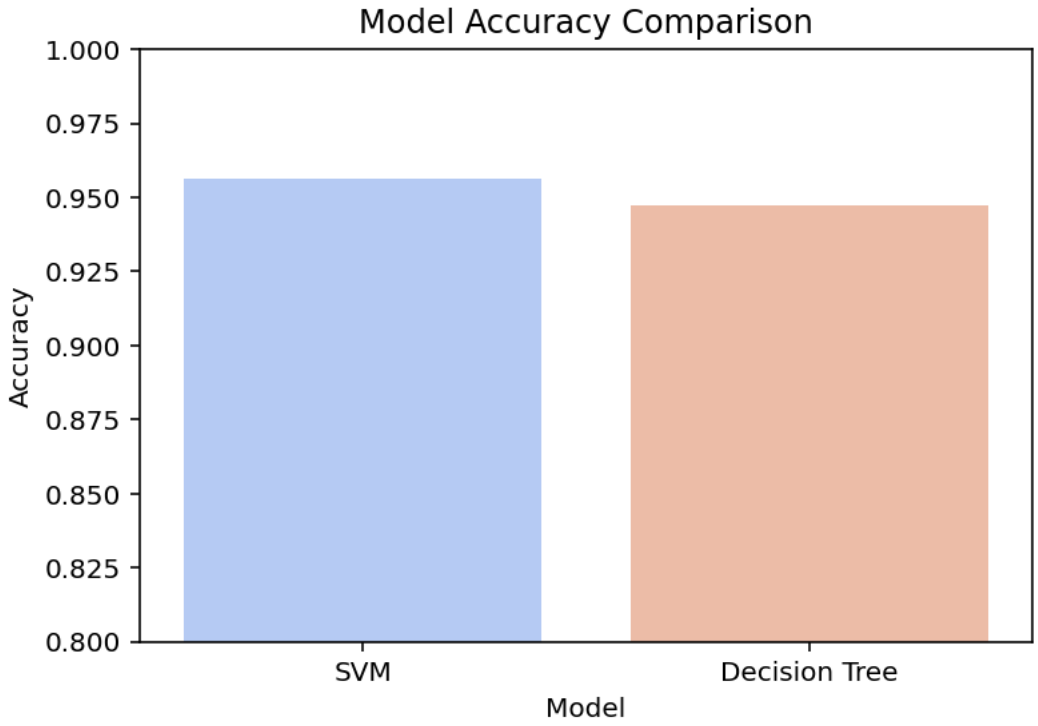
- For Decision Trees, the Gini-based importance revealed top features such as *mean concave points, worst perimeter,* and *mean radius*.
- For SVM, absolute coefficients identified influential predictors in the linear boundary.

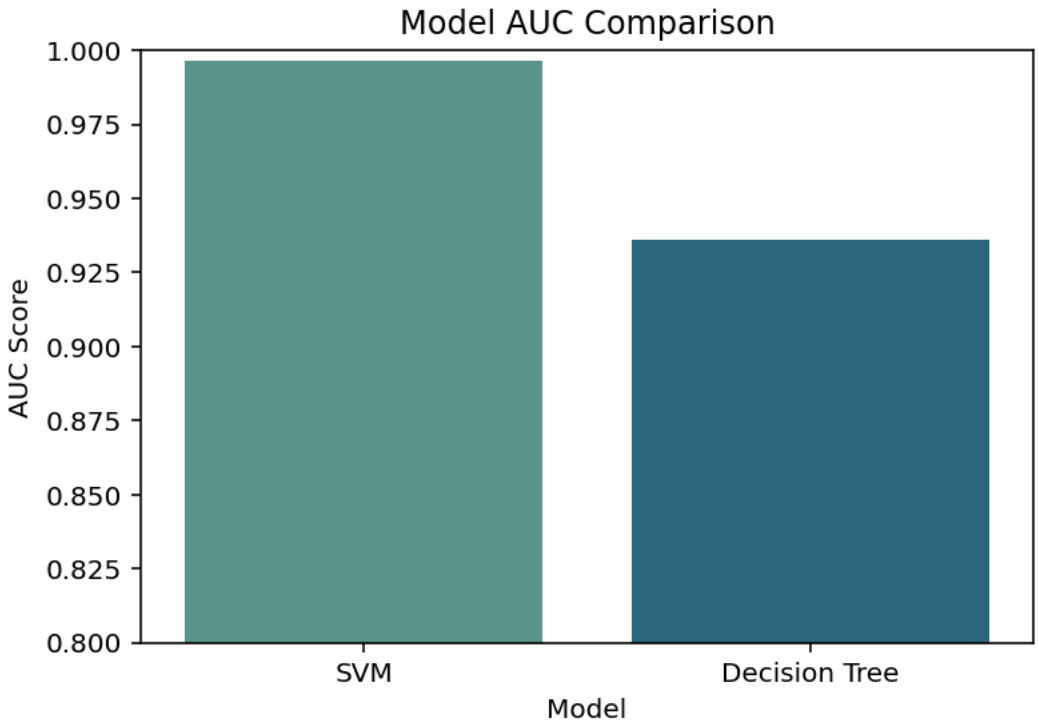




d) Model Comparison

Accuracy and AUC results were compared visually.

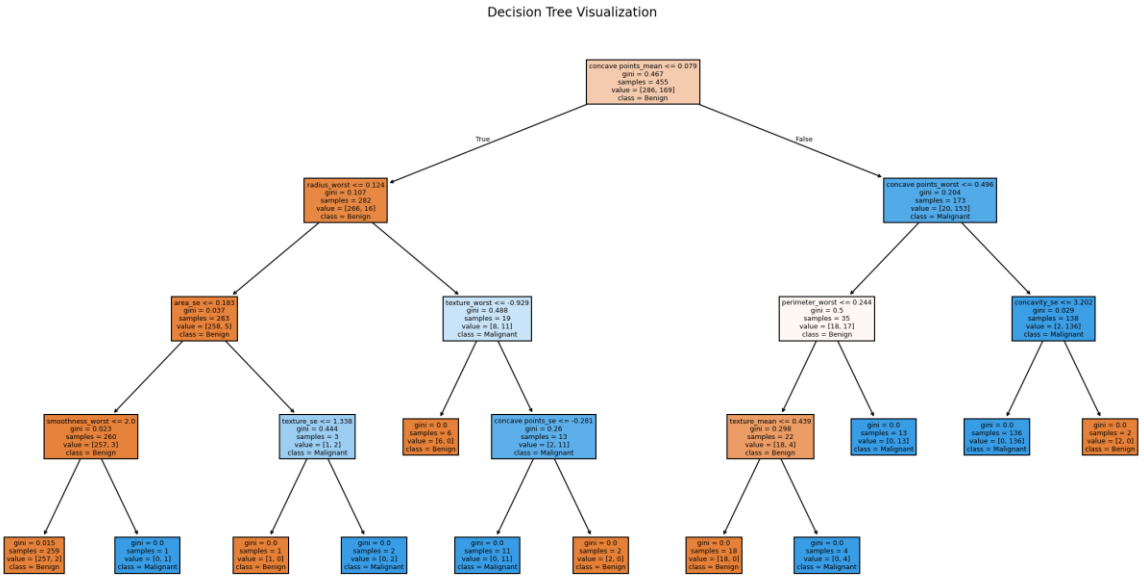




Model	Accuracy	AUC Score
Support Vector Machine	~0.94	~0.99
Decision Tree	~0.94	~0.93

e) Decision Tree Visualization

A full tree diagram was generated, providing transparency into the decision-making process.



Results and Analysis

Both models demonstrated high accuracy in classifying tumor samples.
However, SVM consistently performed better in terms of **generalization** and **AUC**, suggesting a more robust decision boundary.
The Decision Tree, though slightly less accurate, provided valuable **interpretability**—each branch explicitly showed how feature thresholds determine class outcomes.

Key Insights

- **SVM Advantages:** Excellent for high-dimensional spaces, robust to overfitting with proper regularization, high accuracy.
- **Decision Tree Advantages:** Interpretability, feature ranking, and explainable decision paths.
- **Trade-off:** SVM provides better overall predictive power; Decision Tree enhances transparency and trust in medical applications.

Applications

The methodologies applied in this project have direct relevance to healthcare and beyond:

1. **Computer-Aided Diagnosis (CAD):** Automating tumor classification to assist oncologists in early screening.
2. **Predictive Health Analytics:** Using ML to assess patient risk profiles.
3. **Feature Selection Research:** Identifying biological markers relevant to cancer prognosis.
4. **Education and Research:** Demonstrating the practical use of supervised learning models in biomedical contexts.

In real-world settings, such systems can act as **decision support tools**, flagging potential malignancies for further expert evaluation, not replacing but enhancing medical expertise.

Conclusion

This study successfully implemented and compared two classification algorithms — **SVM** and **Decision Tree** — on the Breast Cancer Wisconsin dataset.
Both achieved high predictive accuracy, confirming that machine learning methods can effectively distinguish between malignant and benign tumors using structured numerical data.

Key takeaways include:

- Preprocessing and feature scaling are crucial for performance.
- SVM outperformed Decision Tree in accuracy and AUC.
- Decision Tree offered superior interpretability and clinical transparency.

Future work may involve exploring **ensemble methods** (Random Forest, Gradient Boosting) or **deep learning** architectures for improved performance and robustness.
Integrating such models into clinical workflows can enhance diagnostic precision, optimize screening efficiency, and potentially save lives.

References

[1] Breast Cancer Wisconsin Dataset: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
[2] Scikit-learn Documentation: https://scikit-learn.org/
[3] W. H. Wolberg et al., “Breast Cancer Wisconsin (Diagnostic) Data Set”, UCI Machine Learning Repository.
[4] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.
[5] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106.*