Project Title: DATA ANALYSIS ON BREAST CANCER PREDICTION

Project Dataset details:

The main aim of predicting breast cancer using a dataset is to develop models that can reliably determine whether a tumor is BENIGN which means non-cancerous or MALIGNANT which means cancerous. Therefore, by examining patterns in patient data, including factors like cell properties, tumor size, and other clinical details, these models can assist doctors in making more informed diagnostic choices.

when analyzing a breast cancer dataset, it is very important to choose appropriate dependent and independent variables, as well as the right algorithm for building predictive models.

Choice of Dependent and Independent Variables:

In this project the dependent variable is the one variable that needs to be predicted or that is the outcome. Here, it is represented as B for benign and M for malignant.

Here, independent variables are the main factors used to make prediction in this dataset which include 10 factors listed below:

1. **Perimeter**: Perimeter is the total distance around the nucleus.
2. **Radius:** Radius is the average distance from the center of the nucleus to its edge point.
3. **Texture**: Texture measures the variation in gray-scale values within the nucleus.
4. **Area**: Area represents the size of the nucleus.
5. **Smoothness**: Smoothness indicates how much the radius lengths vary locally.
6. **Compactness**: Compactness is calculated by taking the square of the perimeter, and then dividing it by the area, and after that subtracting 1.
7. **Concavity**: Concavity describes how pronounced the inward curves are along the edge of the nucleus
8. **Concave Points:** Concave Points counts the number of inward curves along the contour of the nucleus
9. **Symmetry:** Symmetry assesses how balanced the nucleus is.
10. **Fractal Dimension:** Fractal Dimension feature captures the complex outline of the nucleus, drawing on principles from fractal geometry.

Selection of Algorithms:

After thorough group discussion, it was concluded that the following algorithms will be utilized for data analysis. The three algorithms used in this project are:

1. Support Vector Machine (SVM)
2. Decision Tree
3. Random Forest

Reason for choosing SVM:

- Breast cancer datasets usually include a wide range if, features, and Support Vector Machine are well suited for handling large numbers of them, making them more effective for complex datasets.
- SVM is specifically designed to perform tasks like determining whether a tumor is benign or malignant, which makes it an ideal choice for breast cancer classification.
- The method focuses on identifying a clear boundary between the two classes B or M, which helps to prevent overfitting and allows it to perform well on new unseen data.

Reason for choosing Decision Tree:

- Decision tree provide a clear and visual method for understanding how decisions progress, making them easy for medical professionals to interpret.
- With the help of decision tree, they can effectively manage both numerical and categorical data, which is ideal for breast cancer datasets that typically contain a combination of these data types.
- Moreover, decision trees are capable of identifying complex patterns in data without depending on linearity assumptions.

Reason for choosing Random Forest:

- Random forest works by combining multiple decision trees, that helps to decrease the number of errors and increase the accuracy of prediction.
- This algorithm mainly good at understanding complex relationships between various features and their outcomes, and it can point out which features in the given dataset are very important.

Data Preparation for Breast Cancer Prediction:

Data preparation is an important step in the breast cancer prediction process, ensuring that the dataset is clean, organized, and ready for model development. As a team, we undertook the following steps for effective data preparation.

- Data Cleaning:

As part of the data preparation process, we performed the following steps to ensure the quality and integrity of the dataset:

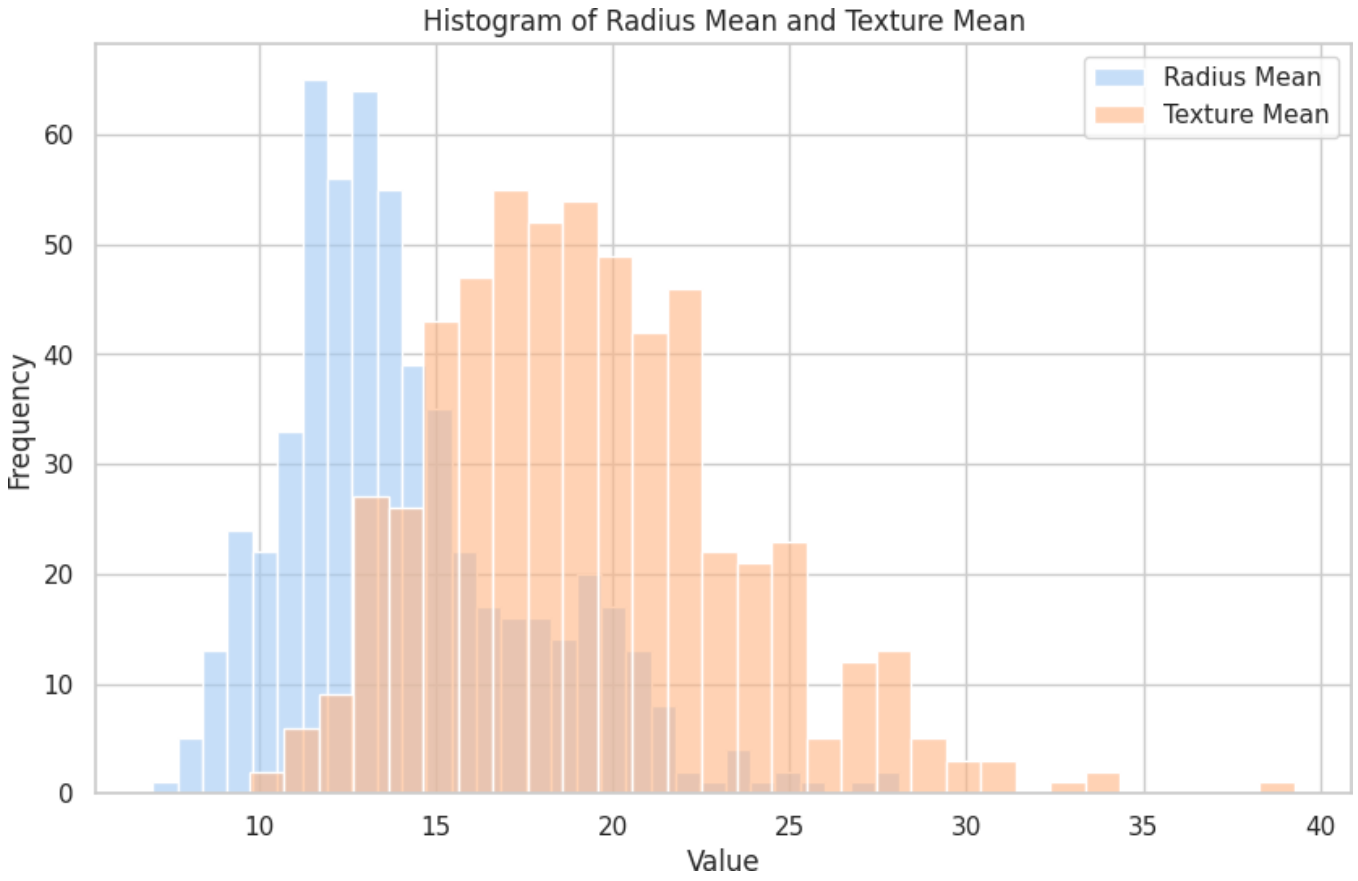
Handling Missing Values:

We began by checking for missing values in the dataset using the ***isnull()*** method provided by Pandas. This function helps identify any missing data across the columns of the dataset. After running this, we found that there were **no missing values** in any of the columns. Therefore, no removal of missing data was required.

Visualization of Dataset:-

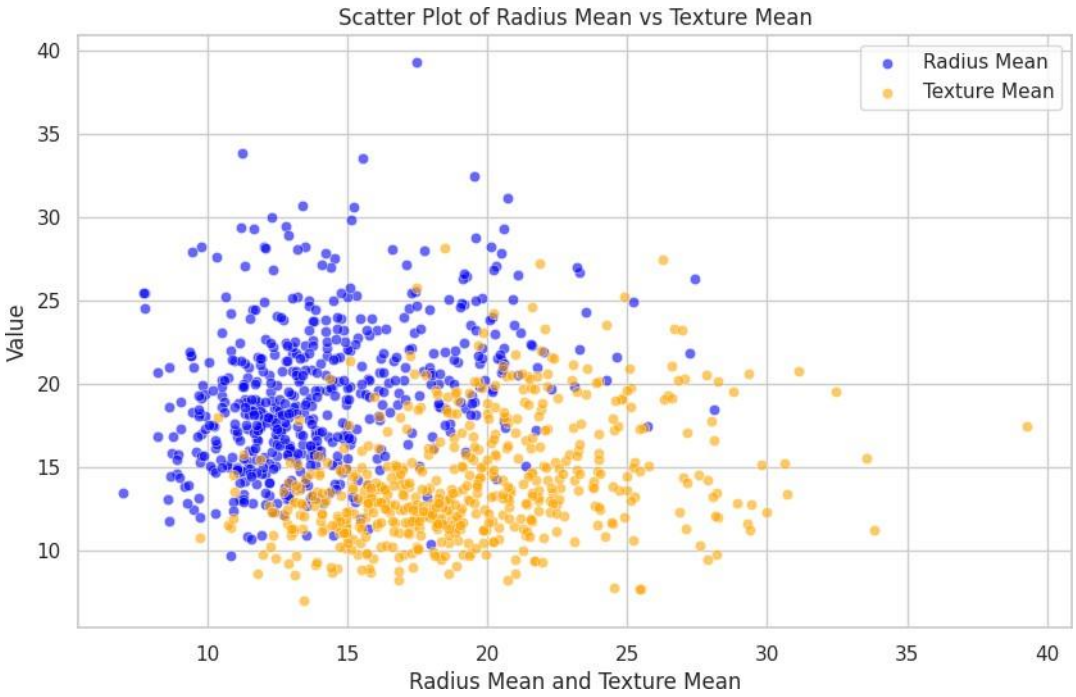
- Histogram

The analysis of histograms for the "radius mean" and "texture mean" features from the breast cancer dataset provides valuable insights into their distributions, which are essential for understanding tumor characteristics. The histogram for "radius mean" usually displays a right-skewed pattern, indicating that most tumors tend to be smaller, with only a few larger tumors present. However, the "texture mean" histogram shows variability in texture traits, hinting at a range of tumor types. These visualizations not only illustrate the distribution and frequency of these features but also serve as a basis for more in-depth statistical analysis and predictive modeling, which can support the early detection and classification of breast cancer.



- Scatter plot:-

The scatter plot that examines the "radius mean" and "texture mean" features from the breast cancer dataset offers a visual insight into how these two important characteristics relate to each other. Each point on the plot represents a tumor, with its location determined by its "radius mean" on the x-axis and "texture mean" on the y-axis. This visualization helps in spotting potential correlations. Moreover, the scatter plot may uncover clusters or patterns within the data, highlighting different tumor types or behaviors, which can be important for creating predictive models for breast cancer diagnosis.



Feature selection:

The original dataset had 32 columns, but after data cleaning and features reduction, it became clear that several columns had similar information. To tackle this issue, we created a correlation matrix to point out highly correlated features. Following this analysis, we removed the similar features, which reduced the dataset to 21 columns. This process not only streamlined the data but also enhanced the efficiency of the modeling stages that followed. The columns that were removed are given below:

‘id’

'perimeter_mean'

'area_mean'

'concave points_mean'

'perimeter_se'

'area_se'

'radius_worst'

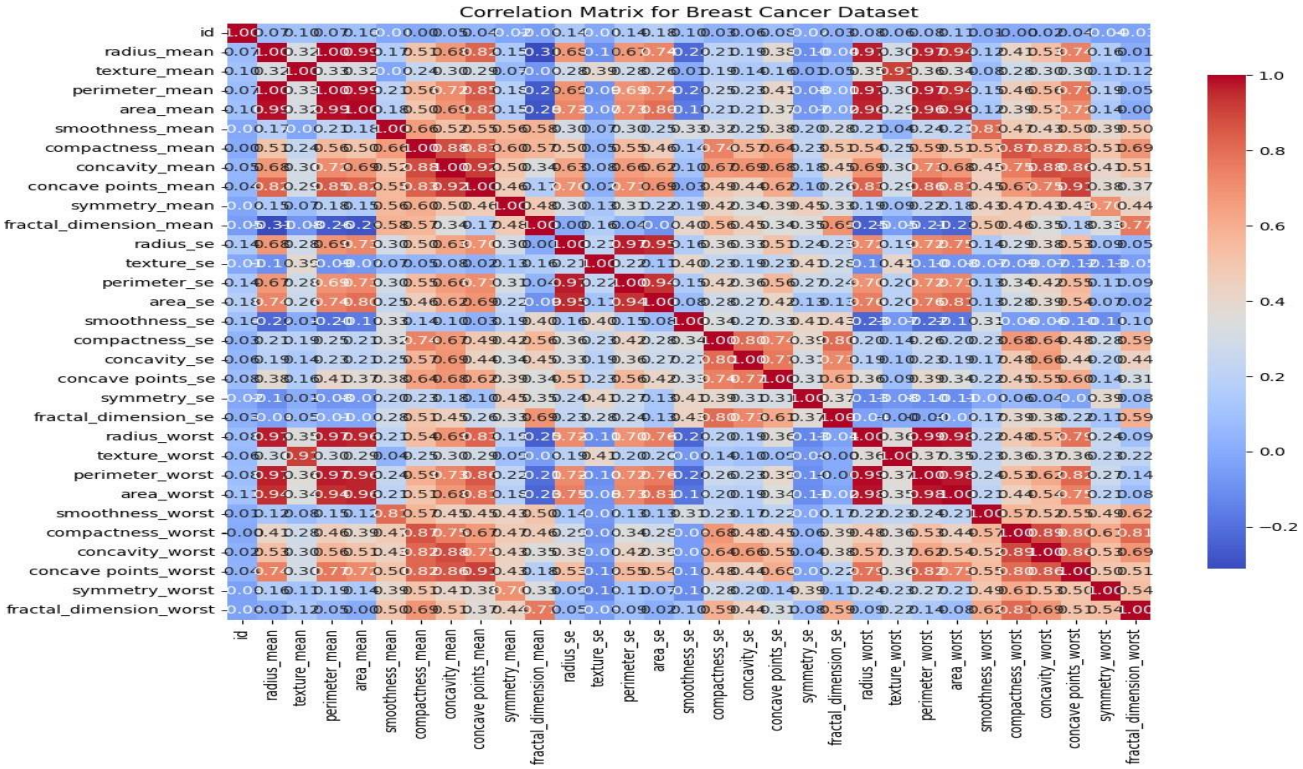
'texture_worst'

'perimeter_worst'

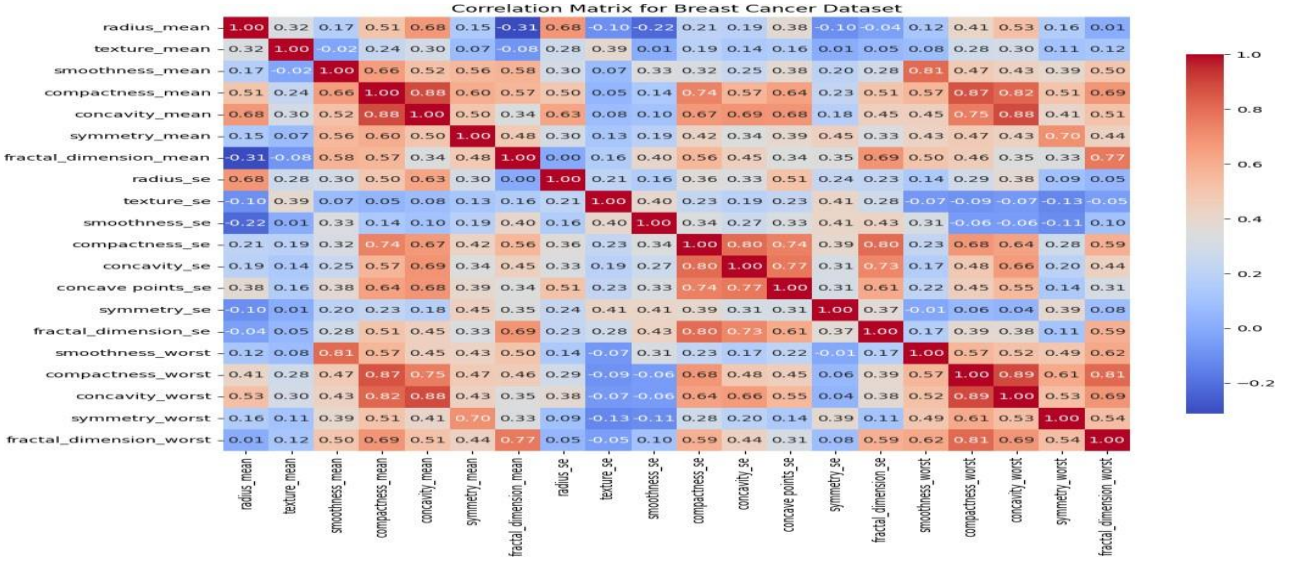
'area_worst'

'concave points_worst'

Correlation Matrix before elimination:



Correlation Matrix after elimination:



Model Development and Evaluation:

The project focused on breast cancer prediction, the team employed three algorithms: Support Vector Machine (SVM), Decision Tree, and Random Forest. These algorithms were selected due to their demonstrated effectiveness in classification tasks and their ability to provide valuable insights from the dataset.

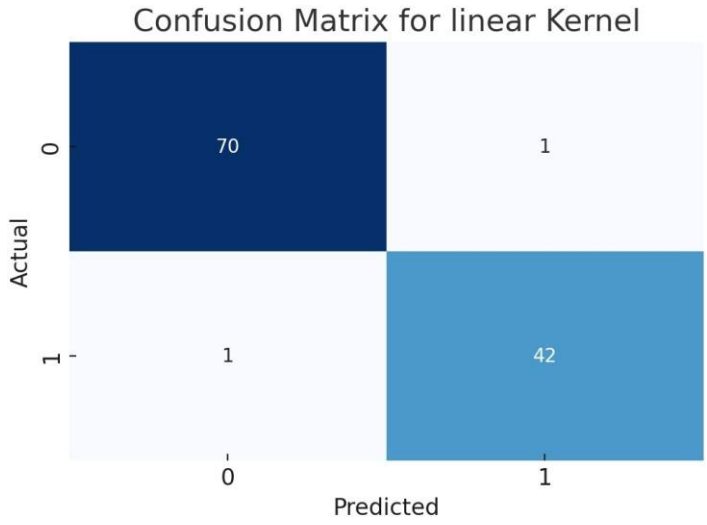
1. **SUPPORT VECTOR MACHINE:**

SVM is a powerful classification algorithm that is commonly employed in breast cancer prediction because it can effectively manage high-dimensional data and establish precise decision boundaries, even in complex situations. The primary aim of utilizing SVM in breast cancer prediction is to determine whether a tumor is BENIGN which means non-cancerous or MALIGNANT which means cancerous by analyzing different features derived from medical

Kernel Selection:

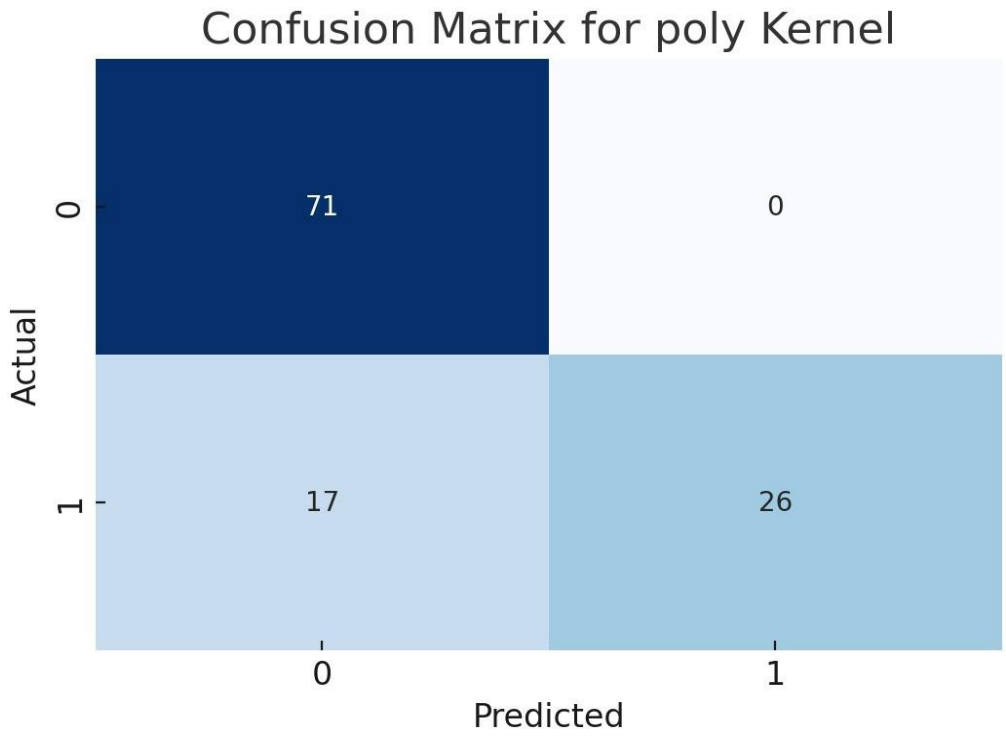
Selecting the appropriate **kernel** in SVM (Support Vector Machine) is crucial for achieving optimal performance in predicting breast cancer .The kernel function determines how the input data is transformed into a higher-dimensional space to make it easier to separate the classes (benign or malignant tumors). The right kernel depends on the nature of your dataset and the relationships between the features.

Matrix for Linear Kernel:



Kernel Type	Class	Precision	Recall	F1-Score	Support
SVM with linear kernel	0	0.99	0.99	0.99	71
	1	0.98	0.98	0.98	43
Accuracy				0.98	114
Macro avg		0.98	0.98	0.98	114
Weighted avg		0.98	0.98	0.98	114

Matrix for Poly Kernel:



Kernel Type	Class	Precision	Recall	F1-Score	Support
SVM with Poly kernel	0	0.81	1.00	0.89	71
	1	1.00	0.60	0.75	43
Accuracy				0.85	114
Macro avg		0.90	0.80	0.82	114
Weighted avg		0.88	0.85	0.84	114

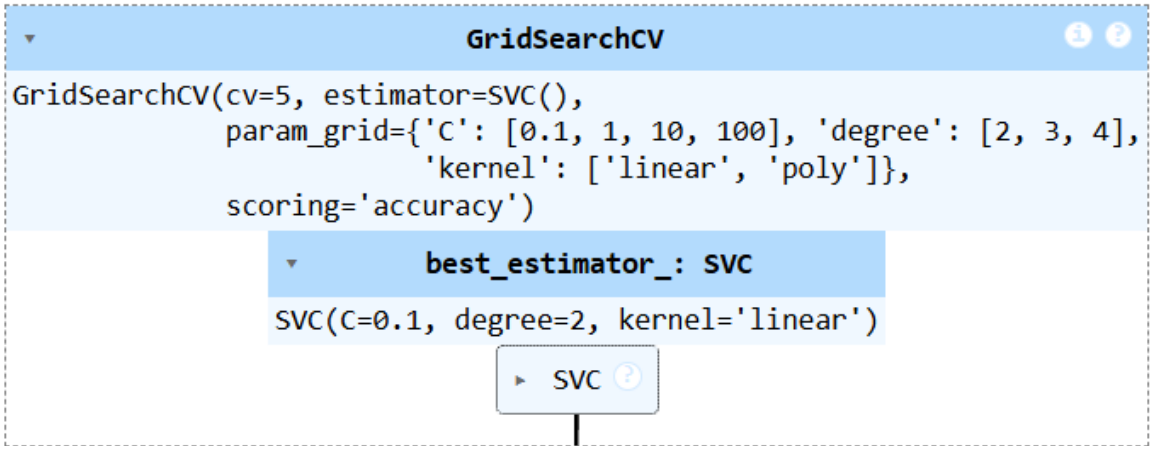
Comparing both these kernels, for breast cancer dataset prediction, SVM with a linear kernel is the better choice due to its superior overall accuracy and balanced precision and recall scores.

GridSearchCV tool is used in this project for hyperparameter tuning. In our project, it will assist in finding the best combination of hyperparameters for your SVM model, ensuring optimal performance in predicting breast cancer outcomes.

```
# Output the best parameters found by GridSearchCV
print("Best parameters found: ", grid_search.best_params_)
print("Best cross-validation score: {:.2f}".format(grid_search.best_score_))
```

```
Best parameters found:  {'C': 0.1, 'degree': 2, 'kernel': 'linear'}
Best cross-validation score: 0.97
```

```
# Set up GridSearchCV with SVM
grid_search = GridSearchCV(svm, param_grid, cv=5, scoring='accuracy')
grid_search.fit(X_train_scaled, y_train)
```

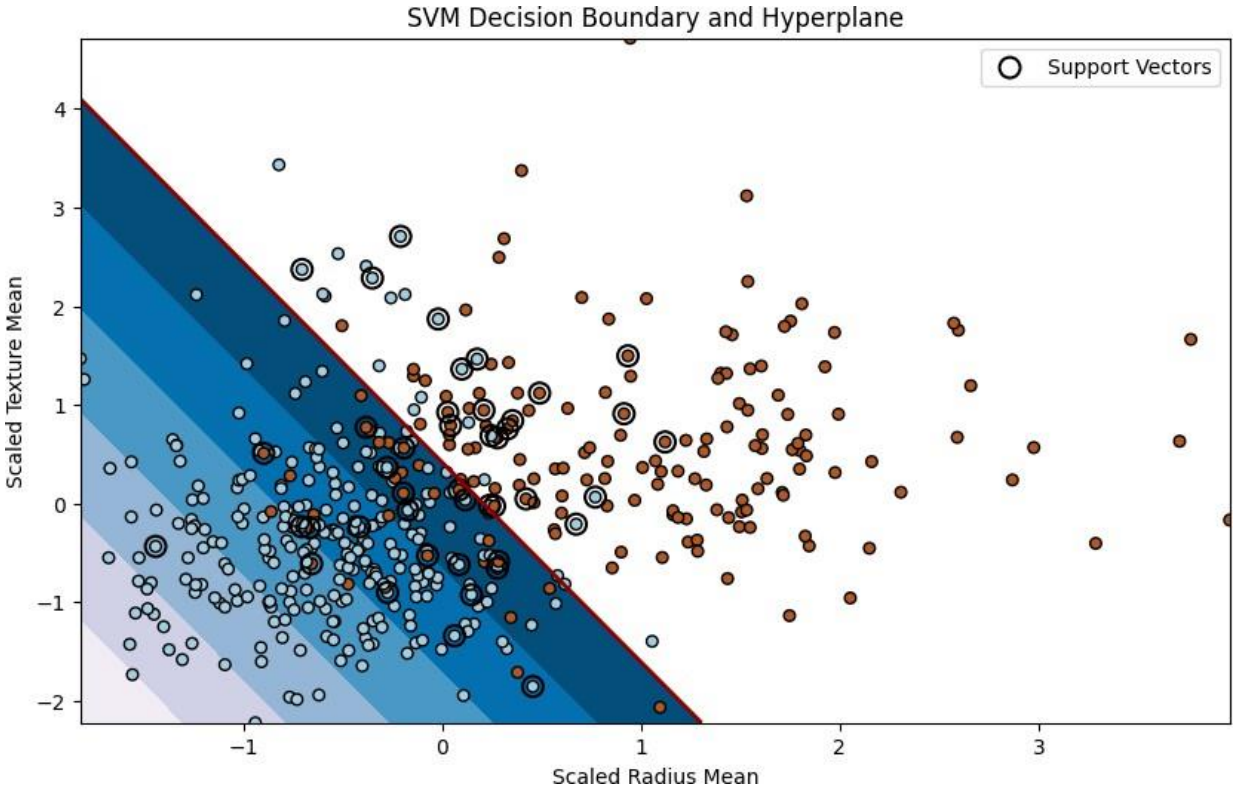


In this case, various combinations of the regularization parameter C (0.1, 1, 10, 100), polynomial degrees (2, 3, 4), and kernel types ('linear', 'poly') are evaluated. The accuracy metric is used to measure the model's performance on the test data. The best estimator found was SVC(C=0.1, degree=2, kernel='linear'), indicating that the combination of a lower regularization strength and a linear kernel provided the optimal balance between bias and variance for the classification task.

Hyperplane for SVM:

In this project, the hyperplane plays an important role in the Support Vector Classifier (SVC) model, serving as the line that differentiates between various classes based on the features *radius_mean* and *texture_mean*. The radius_mean indicates the average size of the tumor, while texture_mean indicates the smoothness or roughness of the tumor's surface. When we visualize these two features on a graph, the hyperplane allows to understand how the model separates benign tumors from malignant ones. It establishes a boundary that maximizes the distance between the nearest points of each class, enabling the model to make precise predictions on new data. Understanding the concept of this hyperplane is crucial for comprehending how the model functions and assessing its effectiveness in diagnosing breast cancer.

Visualization of Hyperplane:



2. DECISION TREE:

The **Decision Tree** is a machine learning model used for classification tasks. It works by splitting the data into smaller subsets based on the values of the features, forming a tree-like structure.

The model achieved an accuracy score of **97.36%**, indicating that it correctly classified approximately 97 out of 100 samples. Here is a comprehensive analysis of the results and their implications:

✓ High Accuracy (97.36%):

The accuracy score in this algorithm indicates that the model is efficient at differentiating between benign and malignant samples. Majority of the predictions made by the model were accurate and it shows accuracy more than before pruning.

✓ Performance Insights from the Confusion Matrix:

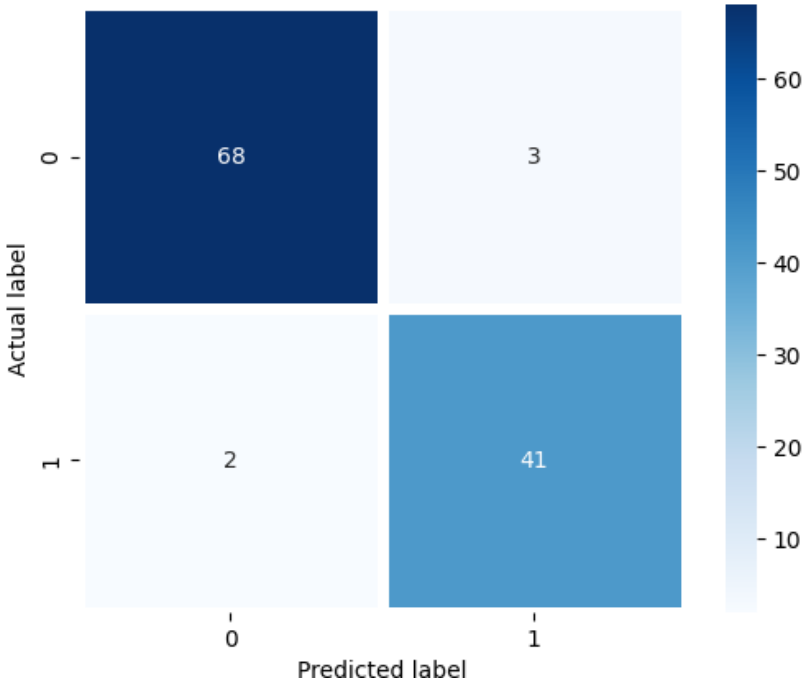
- The model correctly identified **70 benign** samples and **41 malignant** samples.
- There are 3 **misclassifications**:
- **1 False Positives**: Benign samples incorrectly classified as malignant. This could lead to unnecessary further testing or treatment.
- **2 False Negative**: A malignant sample incorrectly classified as benign. This is a more serious issue, as it could result in missing a cancer diagnosis. However, the low number of false negatives suggests the model remains fairly reliable.

3. Use of Entropy for Splitting:

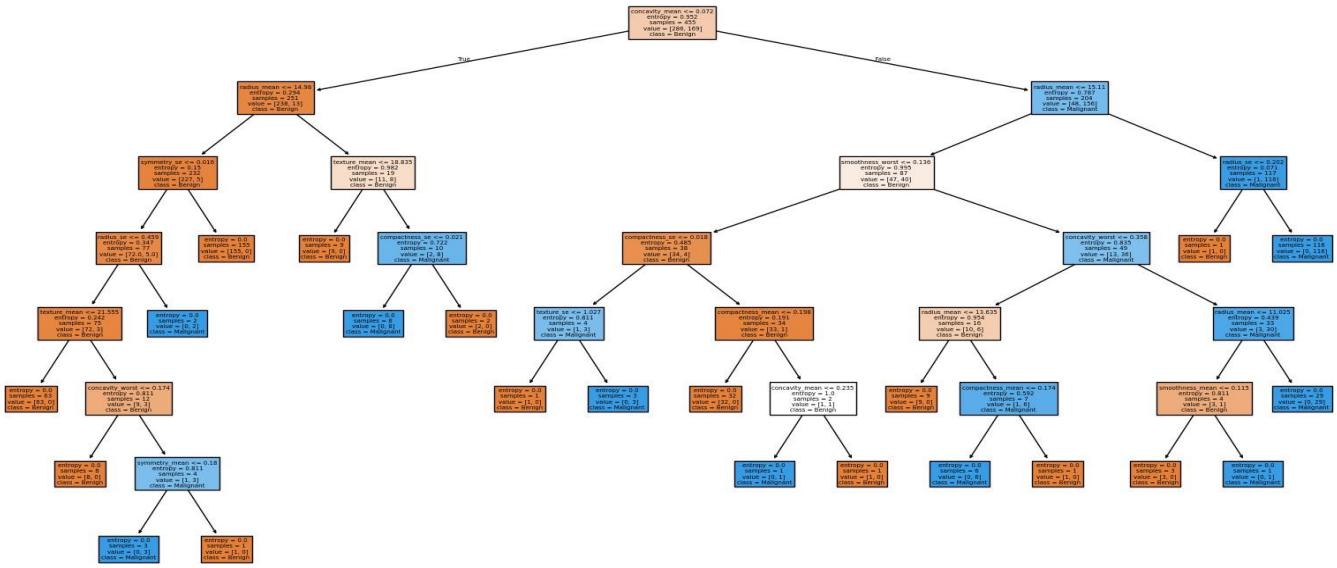
- The Decision Tree model used **entropy** as a criterion to determine the best way to split the data at each node.
- Entropy helps in minimizing disorder) and forming more homogeneous groups, leading to more informed and accurate decision-making by the model. The choice of entropy aligns well with the high performance, ensuring that each split in the tree maximizes the information gain, even after pruning.

Classification Report before pruning:				
	Precision	Recall	F1-Score	Support
Benign	0.97	0.96	0.96	71
Malignant	0.93	0.95	0.94	43
Accuracy			0.96	114
Macro Avg	0.95	0.96	0.95	114
Weighted Avg	0.96	0.96	0.96	114

Before Pruning Accuracy Score: 0.956140350877193



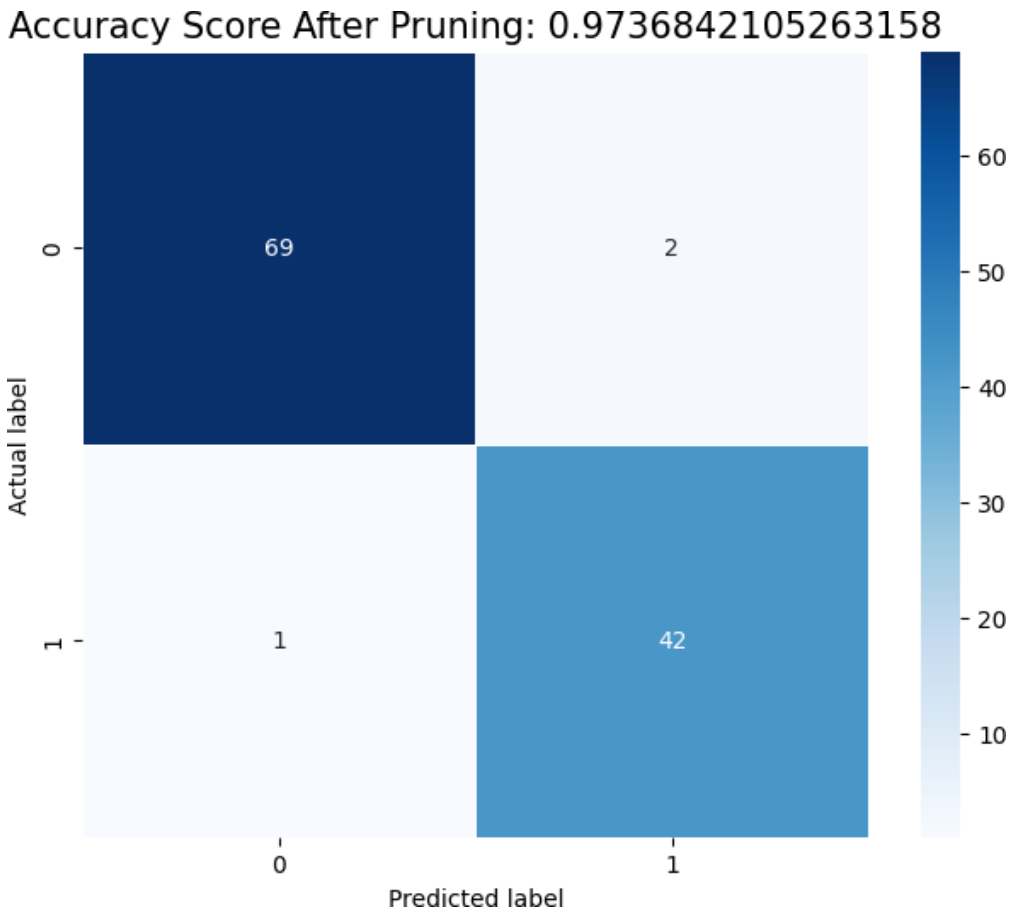
Decision Tree Before Pruning

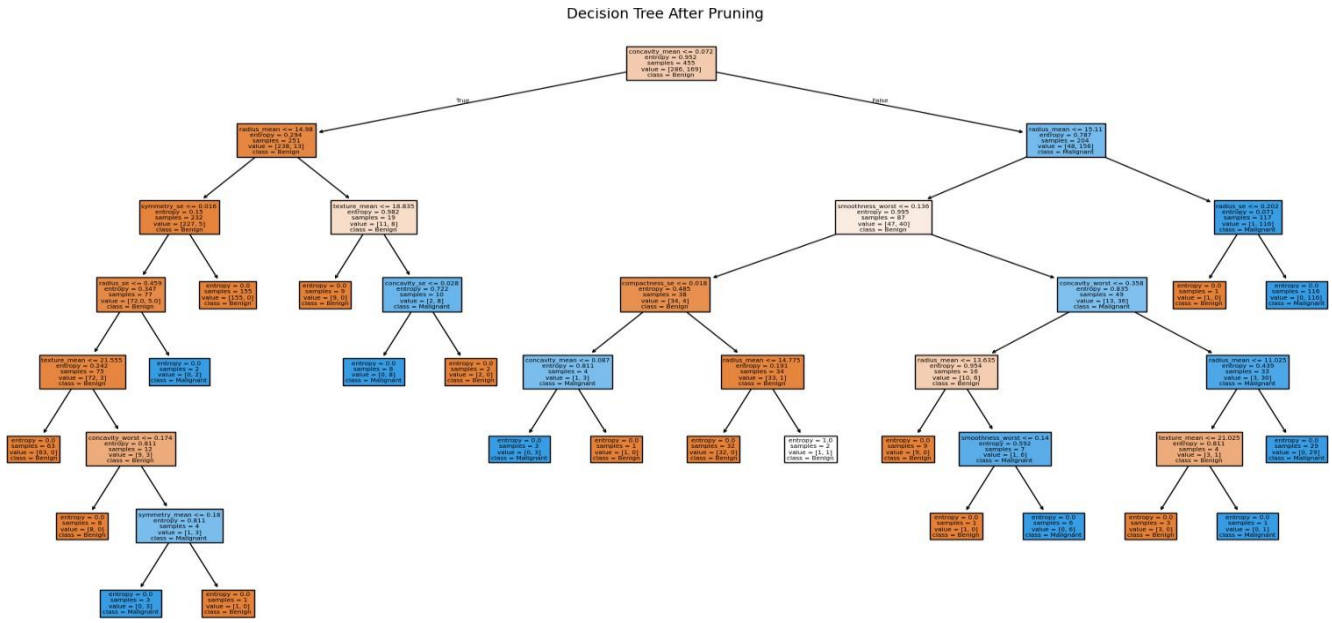


Updated Decision Tree Classifier Report After Pruning:

- The Decision Tree classifier, after pruning, achieved an accuracy of **97.36%** on the test set, indicating strong performance in classifying benign and malignant samples.

Classification Report after pruning:				
	Precision	Recall	F1-Score	Support
Benign	0.99	0.97	0.98	71
Malignant	0.95	0.98	0.97	43
Accuracy			0.97	114
Macro Avg	0.97	0.97	0.97	114
Weighted Avg	0.97	0.97	0.97	114





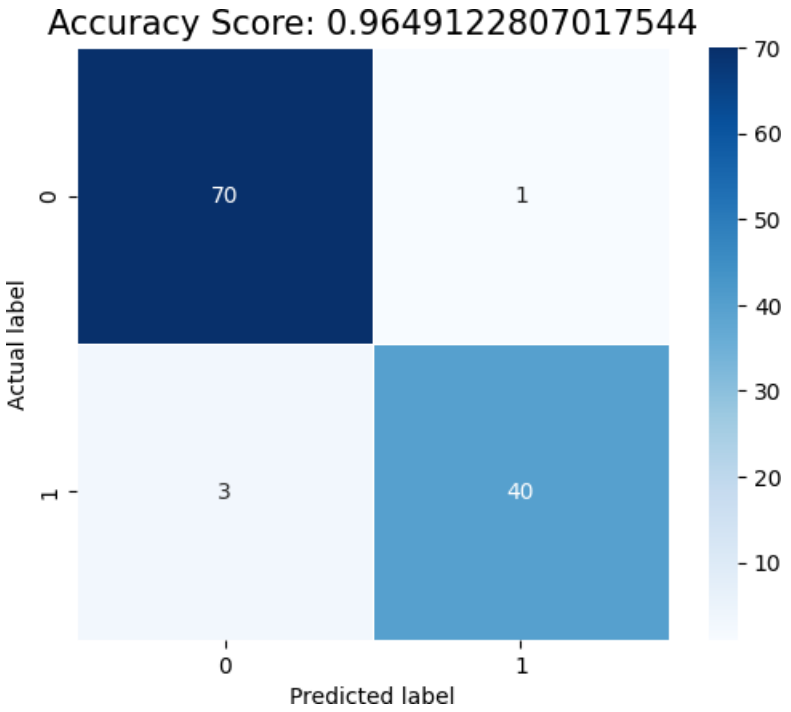
3. RANDOM FOREST:

In this project, we utilize the Random Forest Classifier to predict whether tumors are benign or malignant based on various features extracted from medical imaging data. The primary goal is to develop an accurate predictive model that aids in early cancer diagnosis.

The Random Forest Classifier achieved excellent performance metrics, which are summarized in the table below:

Classification Report	
Accuracy	96.49%
Precision	96.52%
Recall	96.49%
F1 Score	96.47%

Confusion Matrix :



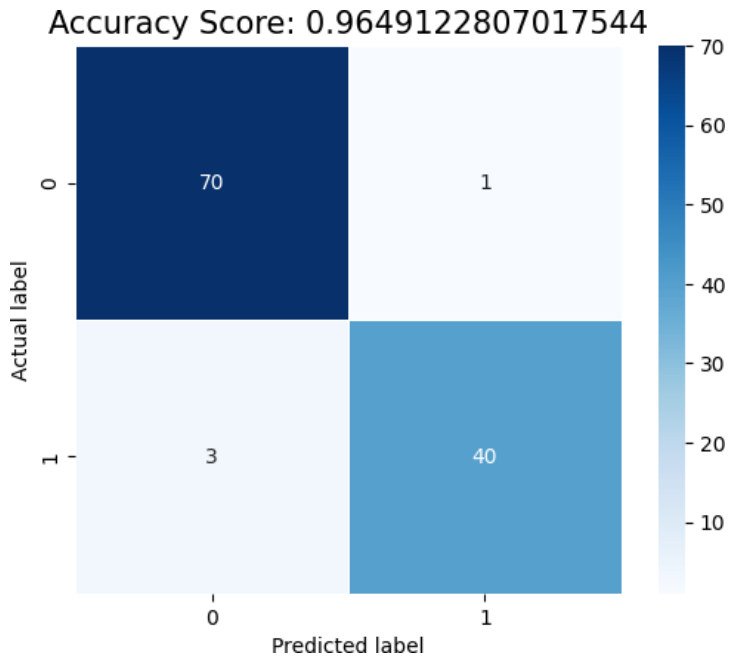
Hyperparameter Tuning:

Hyperparameter tuning was conducted to enhance the accuracy and robustness of the Random Forest Classifier used in predicting whether breast tumors are malignant or benign. The tuning process involved fitting the model through 5-fold cross-validation, evaluating 216 different hyperparameter combinations, resulting in a total of 1080 fits.

- **max_depth**: None (unrestricted depth)
- **max_features**: 'sqrt' (using the square root of the number of features)
- **min_samples_leaf**: 1 (minimum samples required to be at a leaf node)
- **min_samples_split**: 2 (minimum samples required to split an internal node)
- **n_estimators**: 15 (number of trees in the forest)

The best model achieved the following performance metrics, consistent with the initial model:

Best Model Accuracy	96.49%
Best Model Precision	96.52%
Best Model Recall	96.49%
Best Model F1 Score	96.47%



RESULT:

The final results of the predictive models indicate the accuracy of three different algorithms used for breast cancer classification based on the features analyzed: Support Vector Machine (SVM), Decision Tree, and Random Forest.

1. **Support Vector Machine (SVM) - 0.9824 Accuracy**:

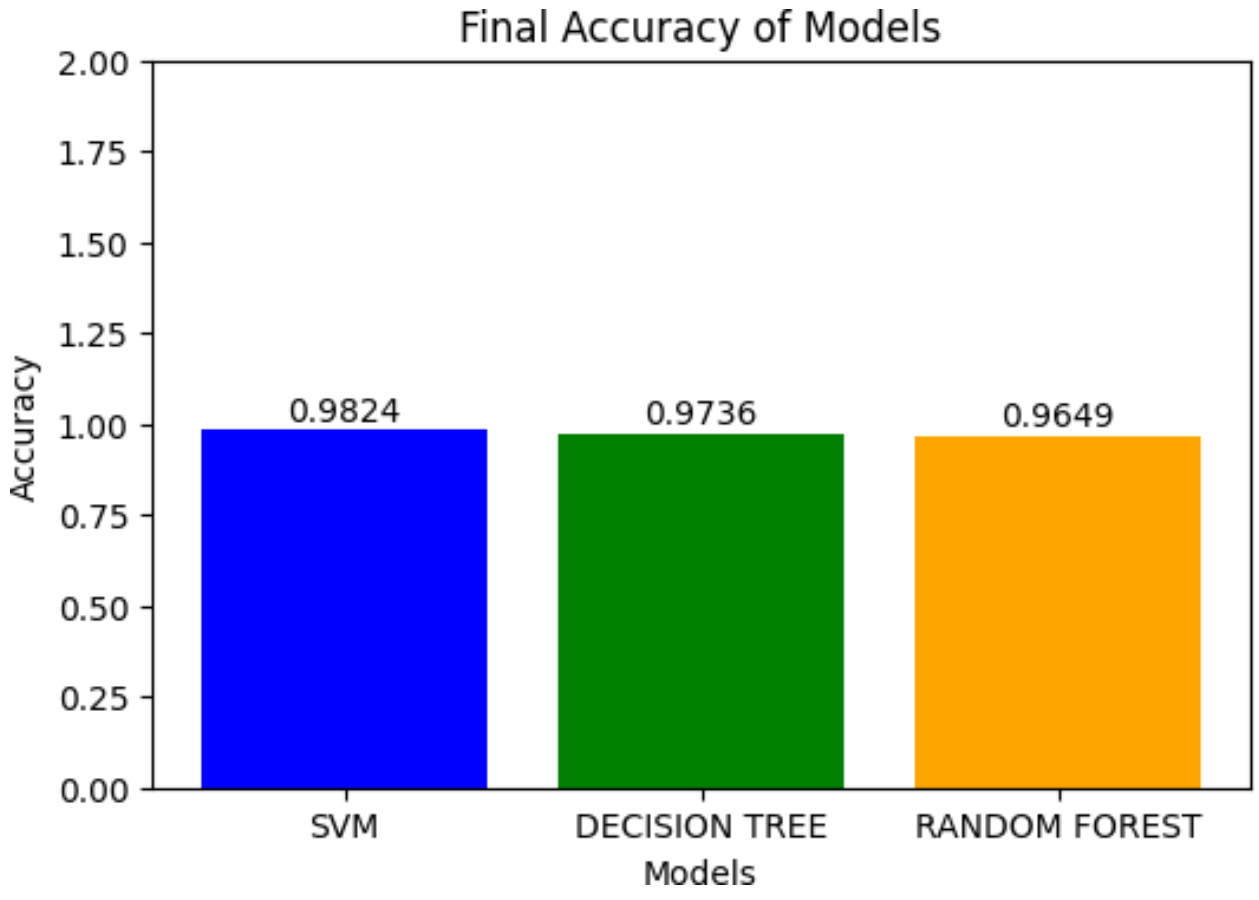
   The SVM model demonstrated an impressive accuracy of 98.24%. This high accuracy indicates that SVM is especially effective at differentiating between benign and malignant tumors based on the given features. SVM performs well with high-dimensional data and is resistant to overfitting, which could account for its outstanding performance in this scenario.

2. **Decision Tree - 0.9736 Accuracy**:

   The Decision Tree model achieved an accuracy of 97.36%, which is impressive but slightly lower than the SVM's performance. Decision Trees are known for their intuitive nature and ease of interpretation; however, they can occasionally overfit the training data, potentially affecting their performance on new, unseen data when compared to the SVM in this case.

3. **Random Forest - 0.9649 Accuracy**:

The Random Forest model achieved an accuracy of 96.49%. This ensemble method, which combines multiple Decision Trees to improve overall performance and reduce the risk of overfitting, still performed well, though it was the least accurate among the three models.



COLLAB LINK:

https://colab.research.google.com/drive/1YrjUA94eOlMInQuzMwwdI7C834vgQqcz?usp=sharing

REFERENCE :

https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic