First Milestone Project Report:

Breast Cancer Prediction

By, Lakshmi Narayanan P

---

1. Project Objective:

The primary goal of this project is to develop and evaluate machine learning models to accurately classify tumors as either malignant (M) or benign (B) based on a set of provided measurements from the Cancer_Data.csv dataset.

2. Data Management and Preparation:

This section outlines the steps taken to prepare the data for model training.

2.1. Data Loading and Initial Exploration:

- The dataset Cancer_Data.csv was loaded using the pandas library.
- Initial exploration included:
  - Viewing the first few rows (raw_data.head()).
  - Assessing DataFrame structure, data types, and null values (raw_data.info()).
  - Generating descriptive statistics (raw_data.describe()).
  - Checking for missing values (raw_data.isnull().sum()).
  - Analyzing the distribution of the target variable 'diagnosis' (raw_data['diagnosis'].value_counts()).

2.2. Data Cleaning:

- The 'Unnamed: 32' column (all null values) and the 'id' column (identifier) were removed from the dataset.
- The cleaned dataset was stored in a DataFrame named Data.

2.3. Data Visualization:

- seaborn and matplotlib libraries were utilized for visualizations.
- **Distribution of Diagnosis**: A count plot was generated to show the number of malignant vs. benign tumors.
- **Feature Relationships**: A pair plot was created to explore relationships between 'radius_mean', 'texture_mean', 'perimeter_mean', and 'area_mean', differentiated by 'diagnosis'.

2.4. Data Preprocessing:

- **Target Variable Conversion**: The 'diagnosis' column was numerically encoded: Malignant ('M') to 1 and Benign ('B') to 0.
- **Feature and Label Splitting**: The data was separated into features (X) and the target label (y).
- **Train-Test Split**: The dataset was divided into training (80%) and testing (20%) sets with random_state=42 for reproducibility.
- **Feature Scaling**: StandardScaler was applied to standardize the features by removing the mean and scaling to unit variance. The scaler was fitted on the training data and applied to both training and testing sets.

3. Model Training and Evaluation with Original Features:

Three machine learning models were trained using all processed features. Their performance metrics on the test set are detailed below.

Model	Accuracy (%)	Precision (0-Benign)	Recall (0-Benign)	F1-score (0-Benign)	Precision (1-Malignant)	Recall (1-Malignant)	F1-score (1-Malignant)	TP	FP	FN	TN
Logistic Regression	97.37	0.97	0.99	0.98	0.98	0.95	0.96	41	1	2	70
Random Forest	96.49	0.96	0.99	0.97	0.98	0.93	0.95	40	1	3	70
SVC	98.25	0.97	1.00	0.99	1.00	0.95	0.98	41	0	2	71

- *TP: True Positives (Correctly predicted Malignant),*
- *FP: False Positives (Incorrectly predicted Malignant),*
- *FN: False Negatives (Incorrectly predicted Benign),*
- *TN: True Negatives (Correctly predicted Benign)*

4. Feature Selection and Model Re-evaluation:

A feature selection step was performed based on correlation with the target variable.

4.1. Feature Selection Criteria:

- Features with an absolute correlation value greater than 0.70 with the 'diagnosis' variable were selected.
- **Selected Features**: 'radius_mean', 'perimeter_mean', 'area_mean', 'concave points_mean', 'radius_worst', 'perimeter_worst', 'area_worst', and 'concave points_worst'.
- The models were retrained and evaluated using this subset of 8 features. *Note: The notebook did not explicitly show rescaling of this new feature subset before retraining the models listed below.*

Below are their detailed performance metrics on the test set using selected features:

Model (Selected Features)	Accuracy (%)	Precision (0-Benign)	Recall (0-Benign)	F1-score (0-Benign)	Precision (1-Malignant)	Recall (1-Malignant)	F1-score (1-Malignant)	TP	FP	FN	TN
Logistic Regression	99.12	0.99	1.00	0.99	1.00	0.98	0.99	42	0	1	71
Random Forest	95.61	0.96	0.97	0.97	0.95	0.93	0.94	40	2	3	69
SVC	94.74	0.92	1.00	0.96	1.00	0.86	0.93	37	0	6	71

A convergence warning was noted for Logistic Regression with new features, indicating that increasing max_iter or ensuring proper scaling of the new feature subset might be beneficial.

5. Model Performance Comparison:

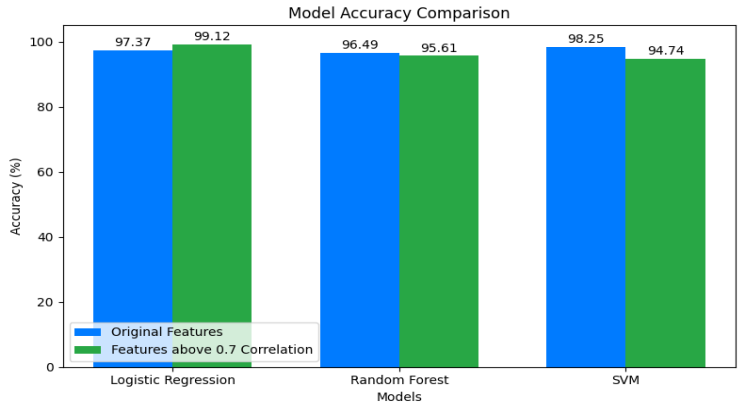
The following table and subsequent chart summarize the accuracy of each model with both the original (all preprocessed) features and the selected (highly correlated) features.

5.1. Accuracy Comparison Table:

Model	Accuracy with Original Features (%)	Accuracy with Selected Features (>0.7 Corr.) (%)	Change in Accuracy (%)
Logistic Regression	97.37	99.12	+1.75
Random Forest	96.49	95.61	-0.88
SVC	98.25	94.74	-3.51

5.2. Visual Comparison:

A bar chart was generated in the notebook to visually compare these accuracies, illustrating the performance changes with different feature sets.



6. Summary and Conclusion:

The project systematically approached the breast cancer prediction task from data loading to model evaluation and feature selection.

- **With Original Features**: The Support Vector Classifier (SVC) yielded the highest accuracy at 98.25%. Logistic Regression was a close second (97.37%), while Random Forest was slightly behind (96.49%).
- **With Selected Features**: Using a subset of 8 features (those with >70% correlation with 'diagnosis'), Logistic Regression's accuracy improved to 99.12%, making it the top-performing model in this scenario. This suggests that for this specific model, a more targeted feature set was advantageous. In contrast, both RandomForestClassifier (95.61%) and SVC (94.74%) saw a decrease in performance with the reduced feature set.
- **Overall**: The Logistic Regression model, when combined with the feature selection strategy of using highly correlated features, achieved the highest accuracy (99.12%) with only one misclassification on the test set. This indicates its effectiveness for this dataset, especially with careful feature selection. The SVC model showed strong performance with a comprehensive feature set.

The variation in model performance with different feature sets highlights that feature engineering and selection are crucial steps and their impact can be model-dependent. Further work could involve more sophisticated feature selection methods, rigorous hyperparameter optimization for all models, and ensuring the scaled version of the selected features is used for re-training to maintain consistency.