BREAST CANCER CLASSIFICATION: COMPARING RANDOM FOREST, SVM, AND LOGISTIC REGRESSION

BREAST CANCER AWARNESS

DEWI AINI NUR ROHMAH

dibimbing.id Digital Skill Fair 35.0 - Data Science

BACKGROUND

- Breast cancer is the leading cause of cancer deaths in women worldwide with the number of cases reaching 2.3 million cases each year (WHO, 2023).
- Early detection of breast cancer can increase survival up to 90% (Kemkes.go.id)
- The biggest challenge in this breast cancer case is the limited access to fast and accurate diagnostic methods.
- Machine Learning is a tool that helps identify malignant and benign tumors with high accuracy and can support early detection efficiently.

OBJECTIVE

- Utilize the *Breast Cancer* dataset from Scikit-Learn
- Compare Mahine Learning models:
  - Random Forest
  - Super Vector Machine (SVM)
  - Logistic Regression
- Evaluate the accuracy and performance of models in detecting malignant and benign tumors

METHODOLOGY

1. LOAD DATASET

Retrieve the breast cancer dataset from Scikit-Learn.

2. EXPLORATORY DATA ANALYSIS (EDA)

Understand data structure, class distribution, and feature correlation.

3. DATA PREPROCESSING

Standardize data and split into training and testing sets.

4. MODEL TRAINING

Build models using Random Forest, SVM, and Logistic Regressi

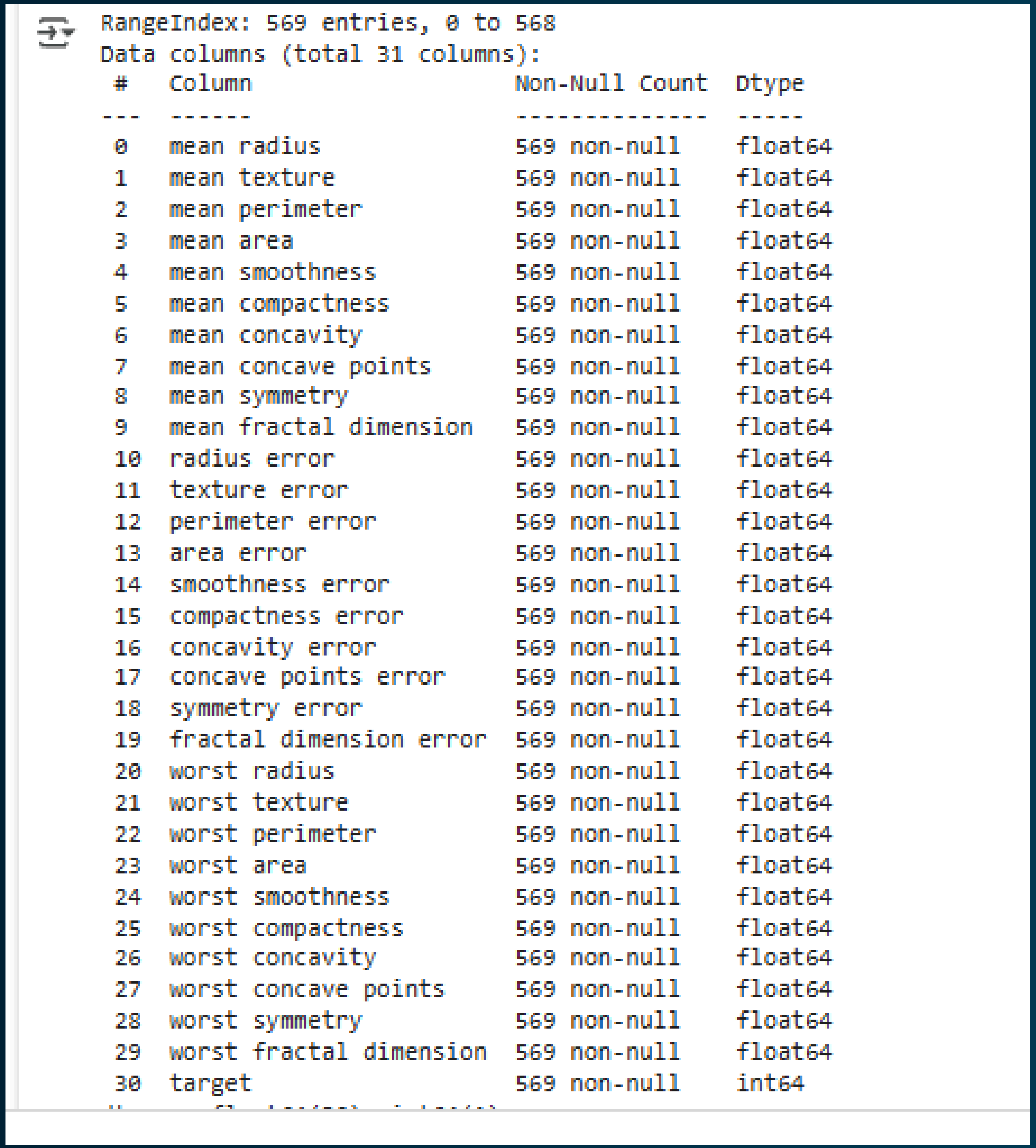
5. MODEL EVALUATION -

Assess performance using accuracy, classification report, and confusion matrix.

6. CONCLUSION

Compare models to determine the best-performing one.

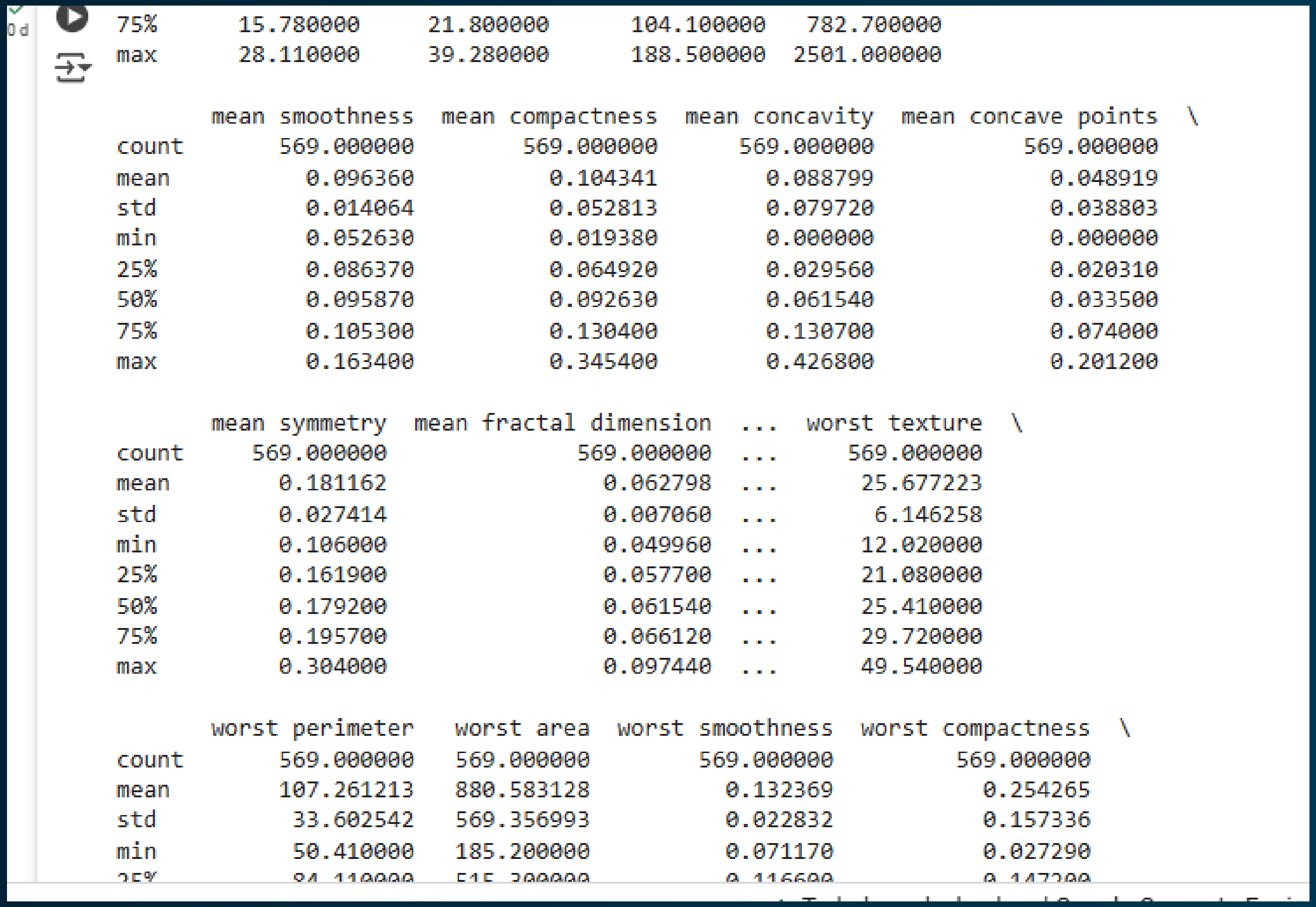
DATA OVERVIEW



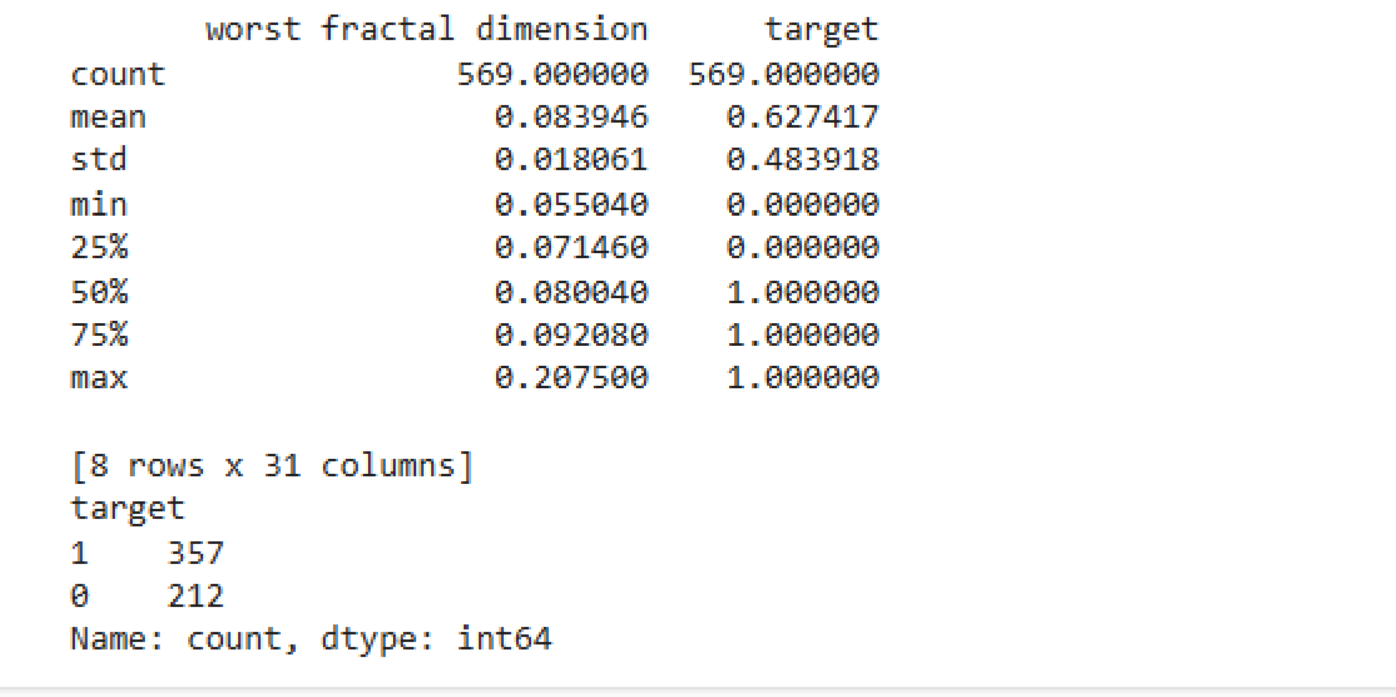
```
RangeIndex: 569 entries, 0 to 568
Data columns (total 31 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   mean radius              569 non-null    float64
 1   mean texture             569 non-null    float64
 2   mean perimeter           569 non-null    float64
 3   mean area                569 non-null    float64
 4   mean smoothness          569 non-null    float64
 5   mean compactness         569 non-null    float64
 6   mean concavity           569 non-null    float64
 7   mean concave points      569 non-null    float64
 8   mean symmetry            569 non-null    float64
 9   mean fractal dimension   569 non-null    float64
 10  radius error             569 non-null    float64
 11  texture error            569 non-null    float64
 12  perimeter error          569 non-null    float64
 13  area error               569 non-null    float64
 14  smoothness error         569 non-null    float64
 15  compactness error        569 non-null    float64
 16  concavity error          569 non-null    float64
 17  concave points error     569 non-null    float64
 18  symmetry error           569 non-null    float64
 19  fractal dimension error  569 non-null    float64
 20  worst radius             569 non-null    float64
 21  worst texture            569 non-null    float64
 22  worst perimeter          569 non-null    float64
 23  worst area               569 non-null    float64
 24  worst smoothness         569 non-null    float64
 25  worst compactness        569 non-null    float64
 26  worst concavity          569 non-null    float64
 27  worst concave points     569 non-null    float64
 28  worst symmetry           569 non-null    float64
 29  worst fractal dimension  569 non-null    float64
 30  target                   569 non-null    int64  
```

Dataset consists of 30 features & 1 target variable (malignant = 0, benign = 1).

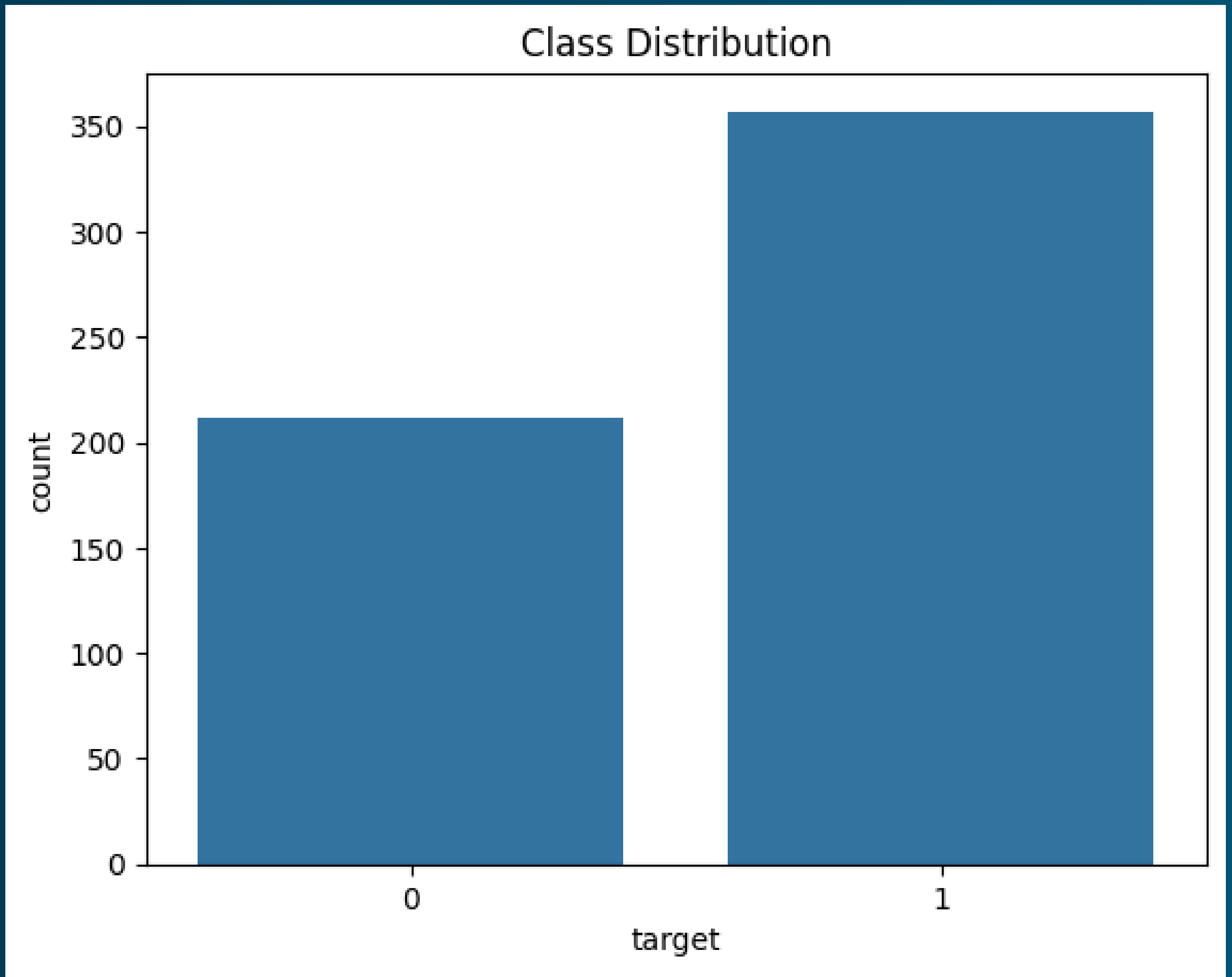
EDA



```
75%      15.780000     21.800000      104.100000   782.700000
max      28.110000     39.280000      188.500000  2501.000000

       mean smoothness  mean compactness  mean concavity  mean concave points  \
count       569.000000        569.000000      569.000000           569.000000
mean          0.096360          0.104341        0.088799             0.048919
std           0.014064          0.052813        0.079720             0.038803
min           0.052630          0.019380        0.000000             0.000000
25%           0.086370          0.064920        0.029560             0.020310
50%           0.095870          0.092630        0.061540             0.033500
75%           0.105300          0.130400        0.130700             0.074000
max           0.163400          0.345400        0.426800             0.201200

       mean symmetry  mean fractal dimension  ...  worst texture  \
count     569.000000              569.000000  ...     569.000000
mean        0.181162                0.062798  ...      25.677223
std         0.027414                0.007060  ...       6.146258
min         0.106000                0.049960  ...      12.020000
25%         0.161900                0.057700  ...      21.080000
50%         0.179200                0.061540  ...      25.410000
75%         0.195700                0.066120  ...      29.720000
max         0.304000                0.097440  ...      49.540000

       worst perimeter  worst area  worst smoothness  worst compactness  \
count       569.000000  569.000000        569.000000         569.000000
mean        107.261213  880.583128          0.132369           0.254265
std          33.602542  569.356993          0.022832           0.157336
min          50.410000  185.200000          0.071170           0.027290
25%          84.110000  515.300000          0.116600           0.147200
```



```
       worst fractal dimension      target
count               569.000000  569.000000
mean                  0.083946    0.627417
std                   0.018061    0.483918
min                   0.055040    0.000000
25%                   0.071460    0.000000
50%                   0.080040    1.000000
75%                   0.092080    1.000000
max                   0.207500    1.000000

[8 rows x 31 columns]
target
1    357
0    212
Name: count, dtype: int64
```

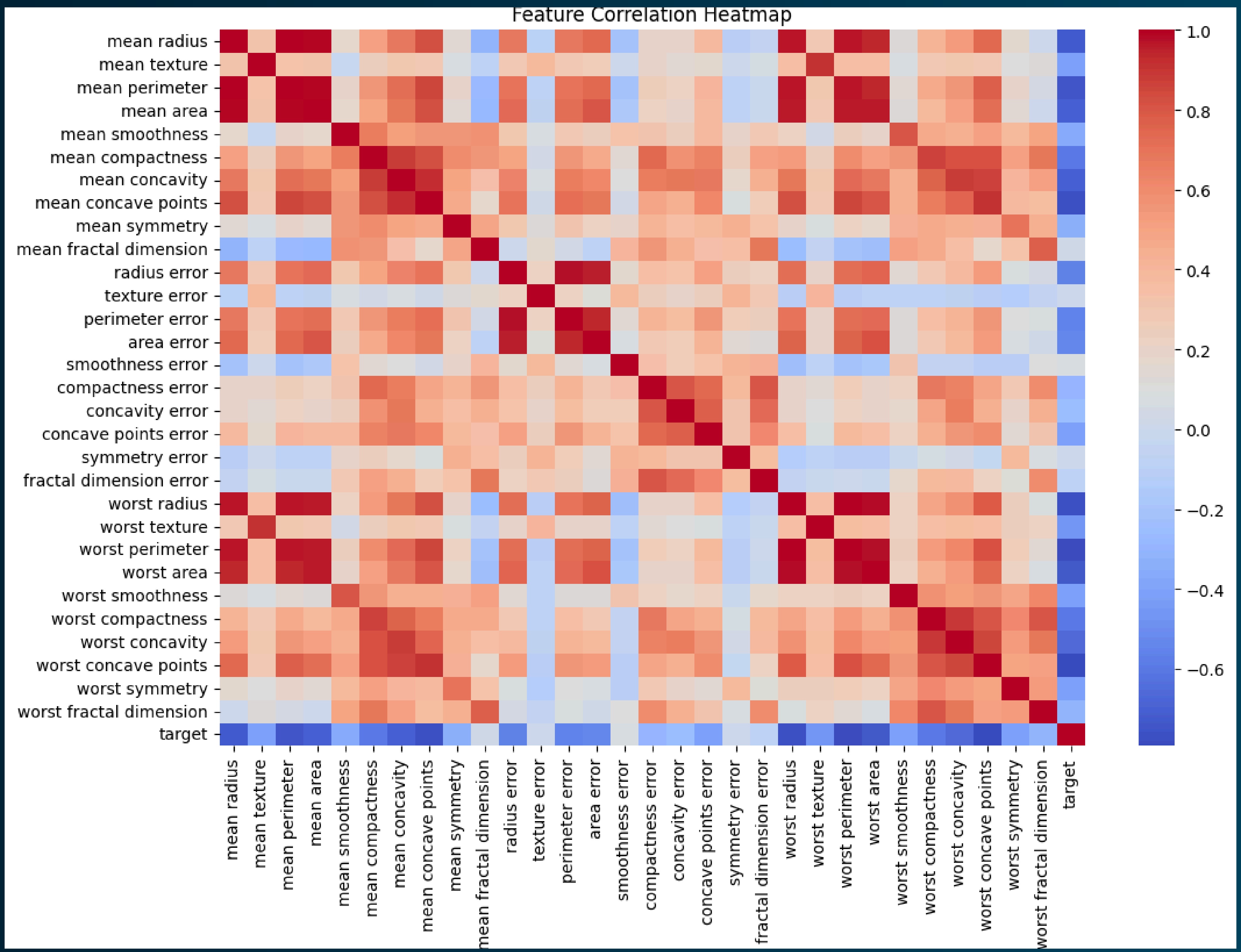


Class Distribution:

Malignant: 212 samples

Benign: 357 samples

FEATURE CORRELATON HEATMAP



- What is Feature Correlation?
  - Measures how strongly different variables relate to each other.
  - A high correlation between features can indicate redundancy.
- Why is it Important?
  - Helps in feature selection for model efficiency.
  - Identifies relationships between tumor characteristics.
- Heatmap Interpretation:
  - Darker colors indicate stronger positive/negative relationships.
  - Helps determine which features are most influential in classification.
  - Example: "Mean radius" and "mean perimeter" have a high correlation, meaning larger tumors tend to have a greater perimeter.

DATA PREPROCESSING

- Train-Test Split: 80% training, 20% testing.
- Data Standardization using StandardScaler.
- Why Standardization?
    - SVM & Logistic Regression are sensitive to data scaling.
    - Helps models perform optimally.

MACHINE LEARNING MODELS

RANDOM FOREST

Tree-based model, effective for complex data.

SUPPORT VECTOR MACHINE (SVM):

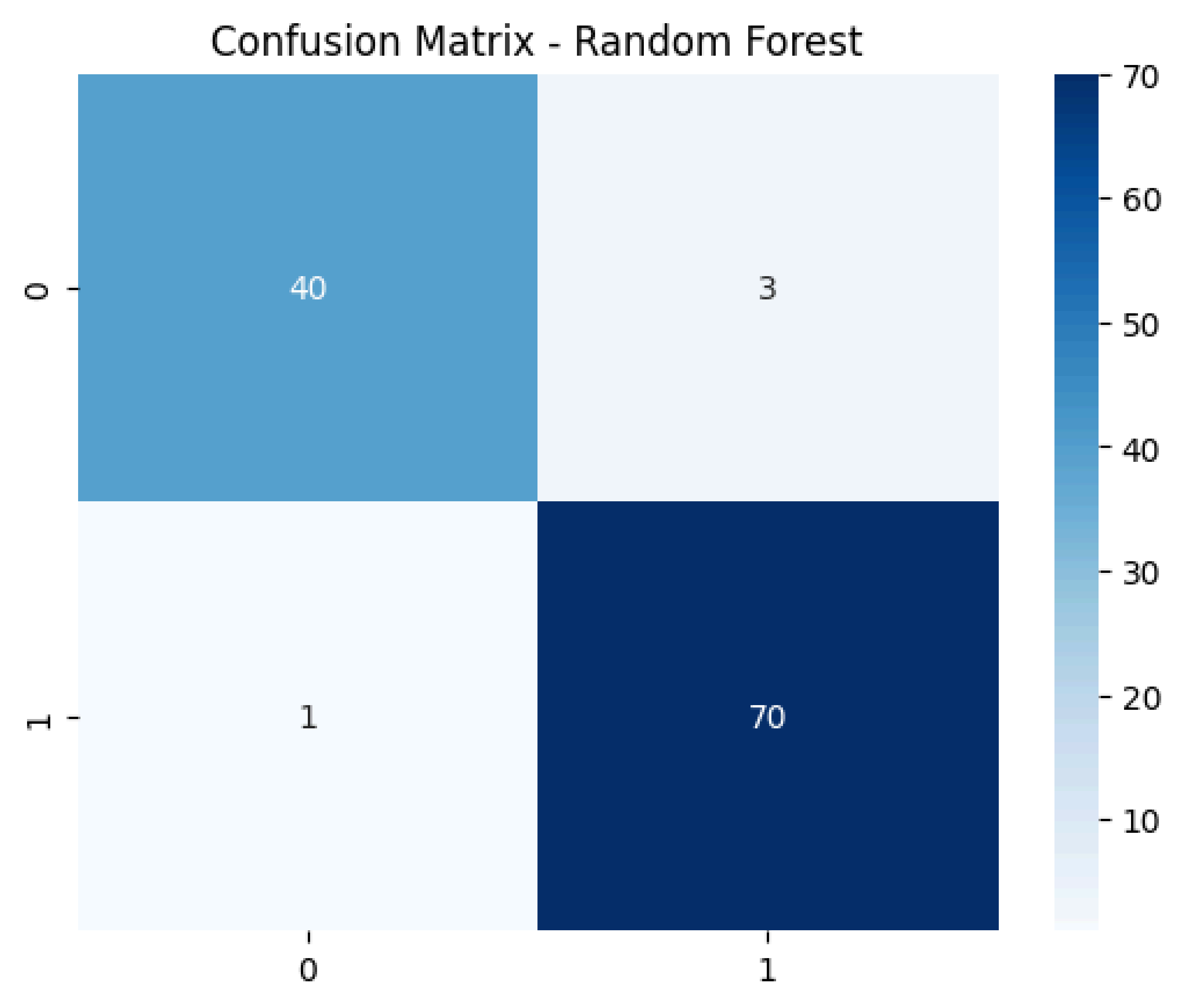
Finds the optimal hyperplane for classification.

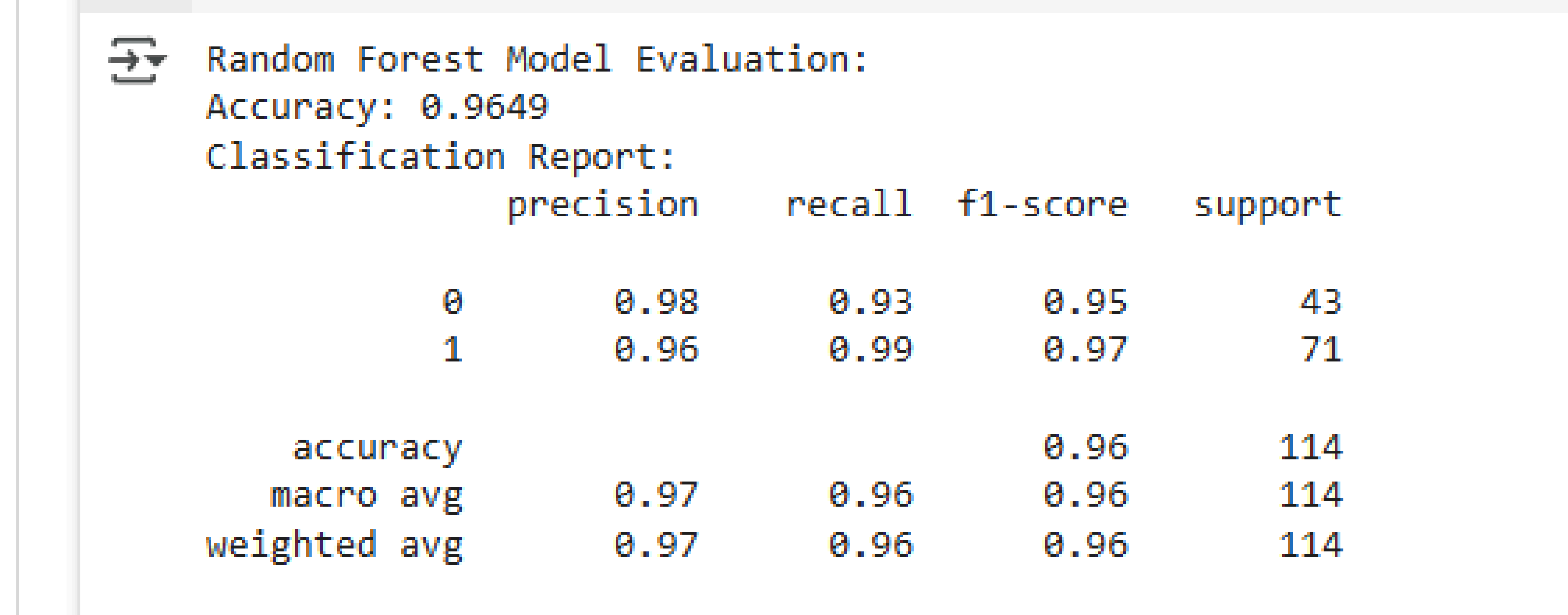
LOGISTIC REGRESSION

Simple probability-based model.

ALL MODELS ARE TRAINED AND TESTED ON THE SAME DATASET

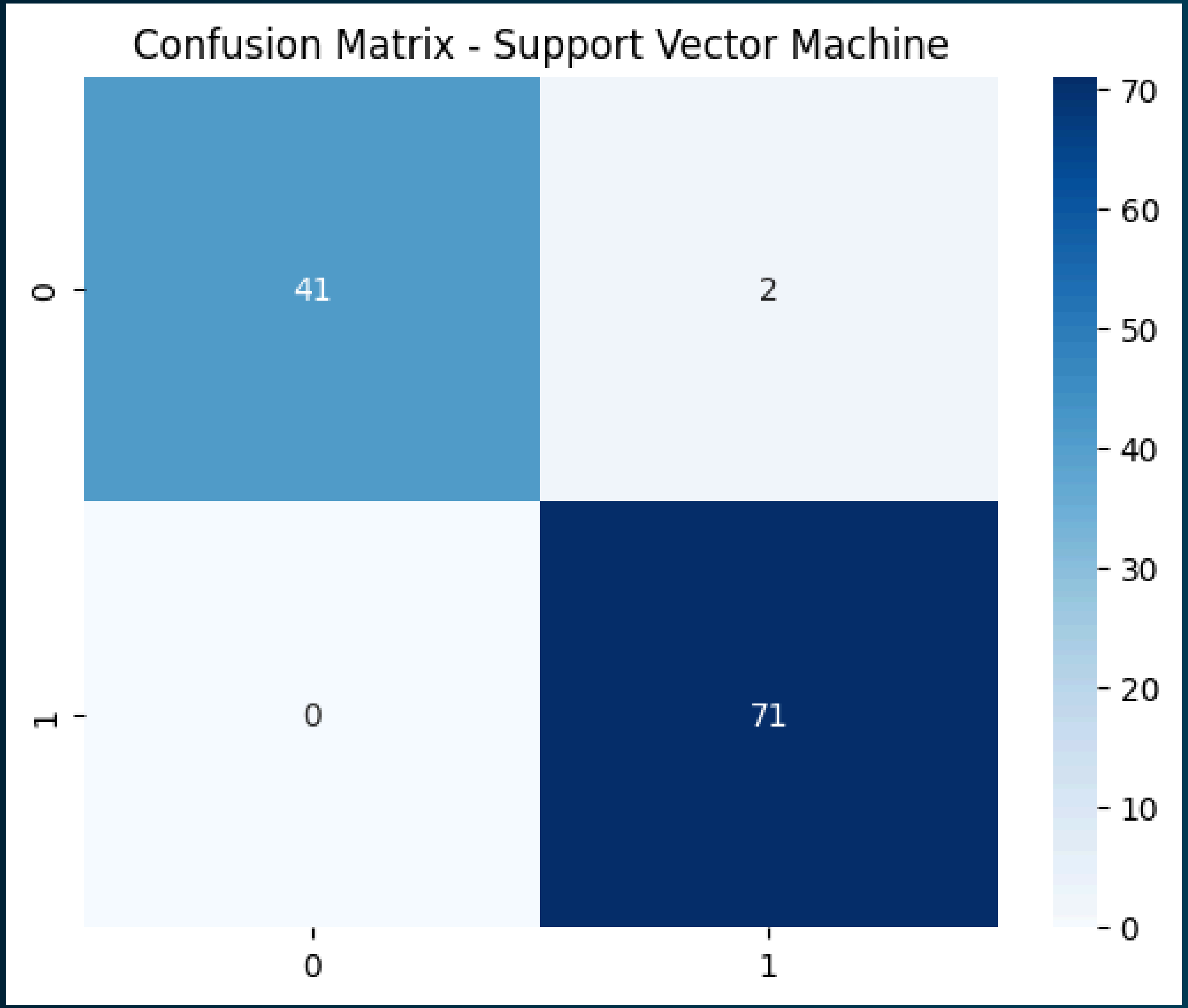
MODEL EVALUATION
RANDOM FOREST

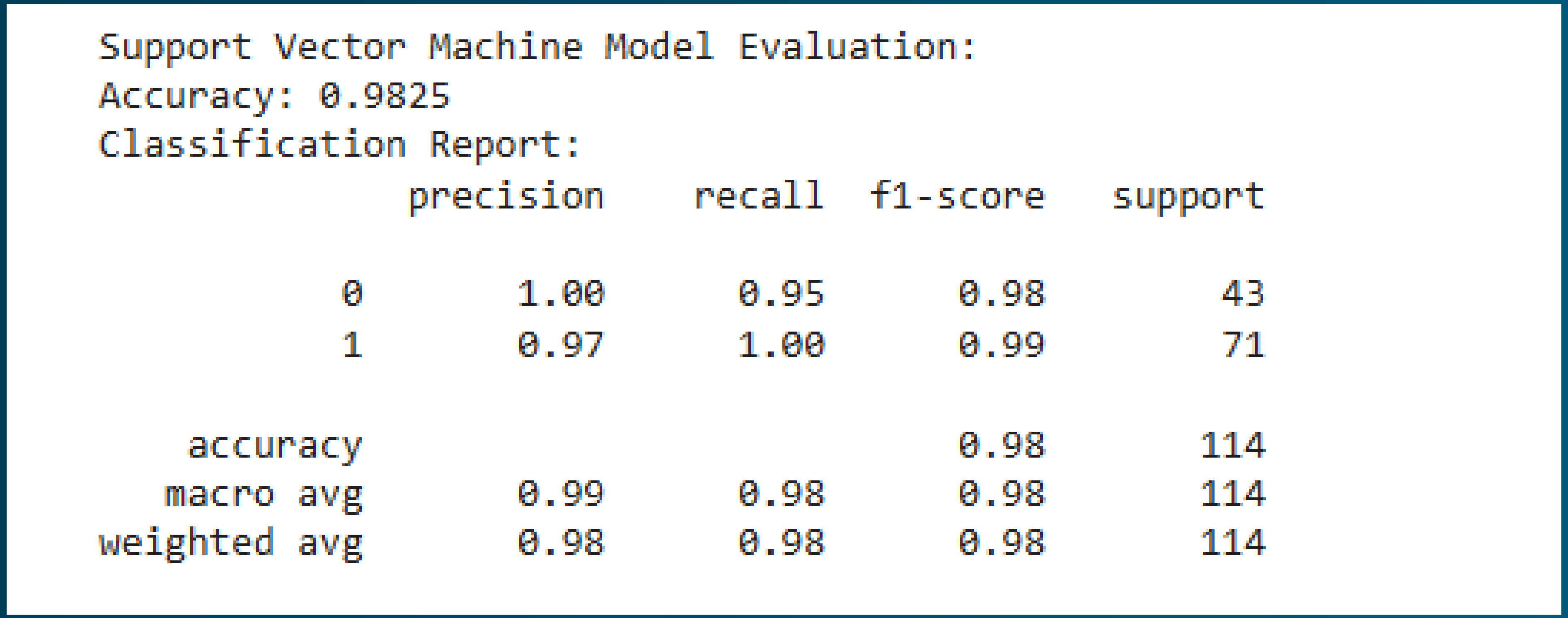




```
Random Forest Model Evaluation:
Accuracy: 0.9649
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.93      0.95        43
           1       0.96      0.99      0.97        71

    accuracy                           0.96       114
   macro avg       0.97      0.96      0.96       114
weighted avg       0.97      0.96      0.96       114
```

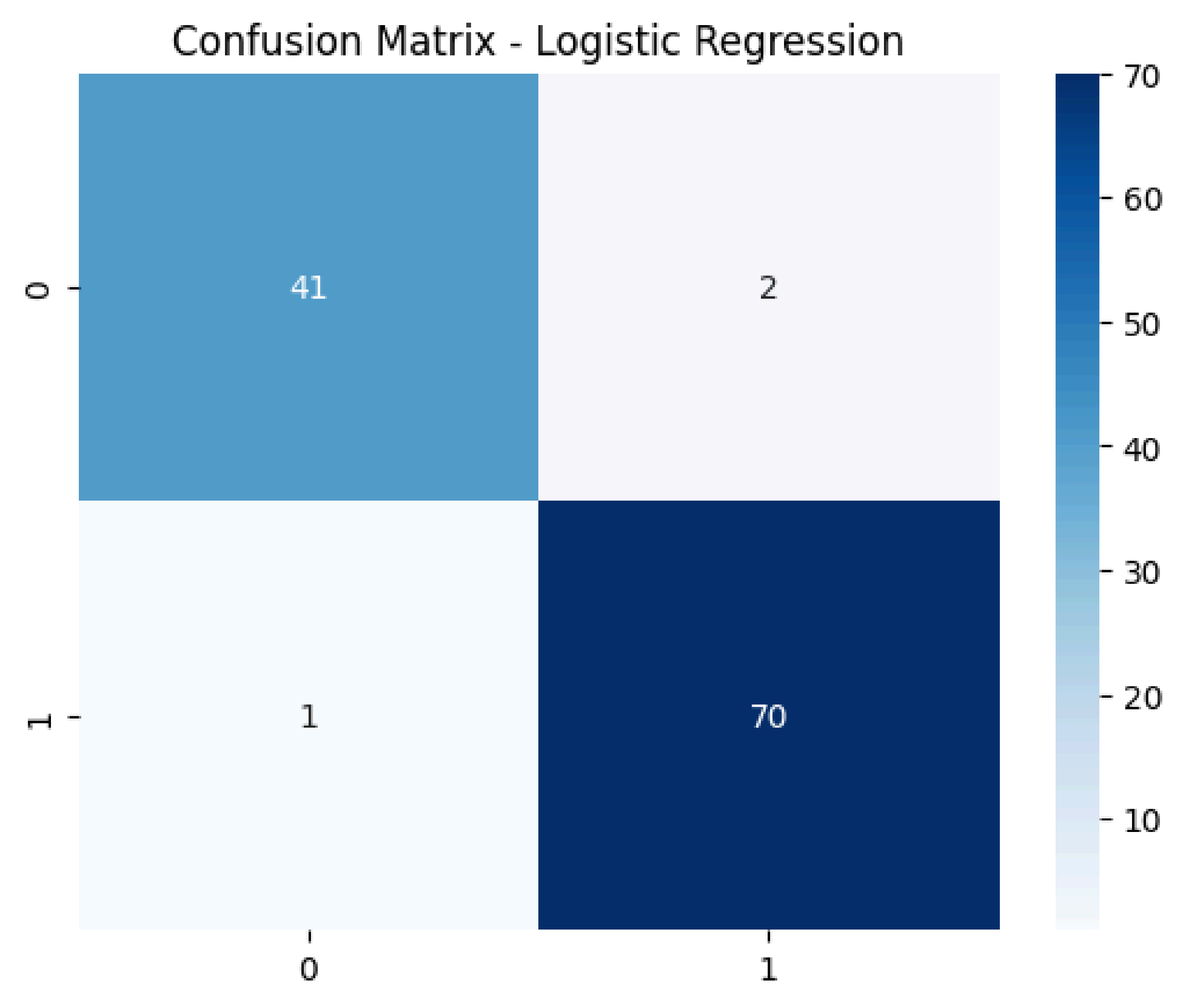
MODEL EVALUATION
SUPPORT VECTOR MACHINE

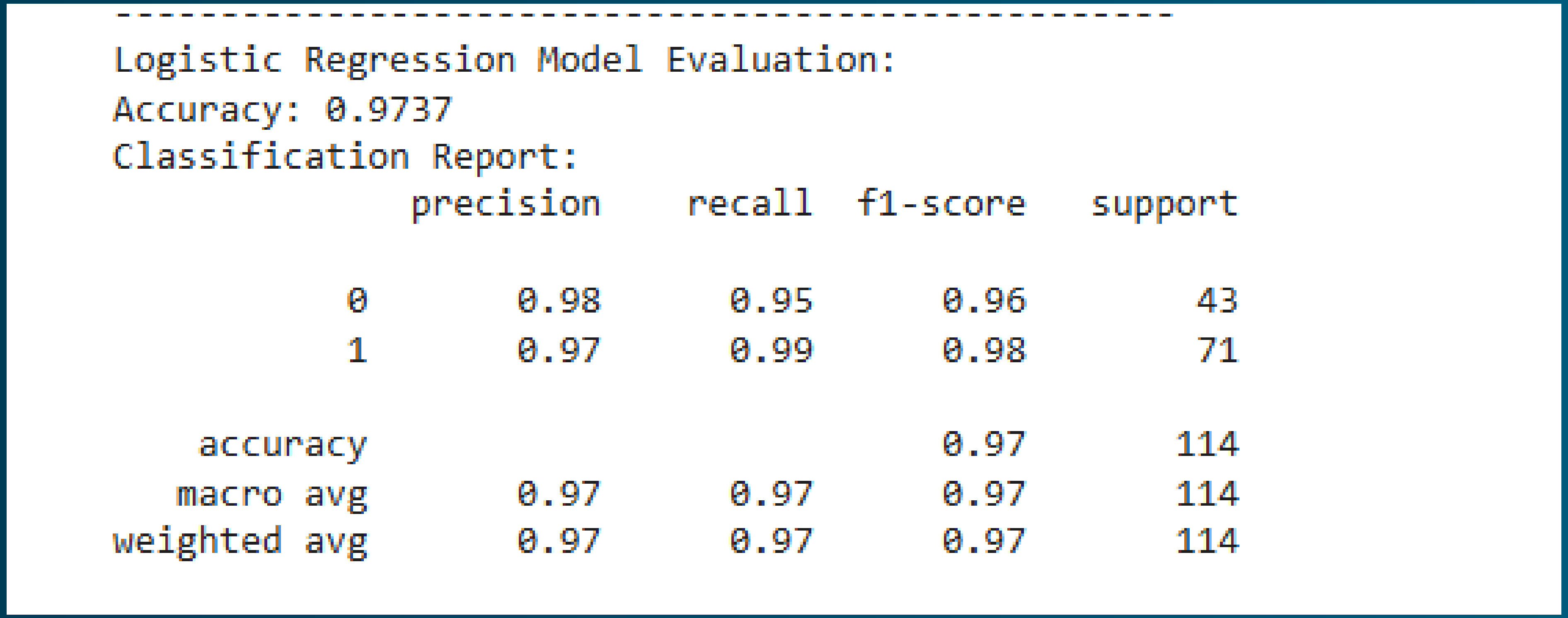




```
Support Vector Machine Model Evaluation:
Accuracy: 0.9825
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.95      0.98        43
           1       0.97      1.00      0.99        71

    accuracy                           0.98       114
   macro avg       0.99      0.98      0.98       114
weighted avg       0.98      0.98      0.98       114
```

MODEL EVALUATION
LOGISTIC REGRESSION


Confusion Matrix - Logistic Regression



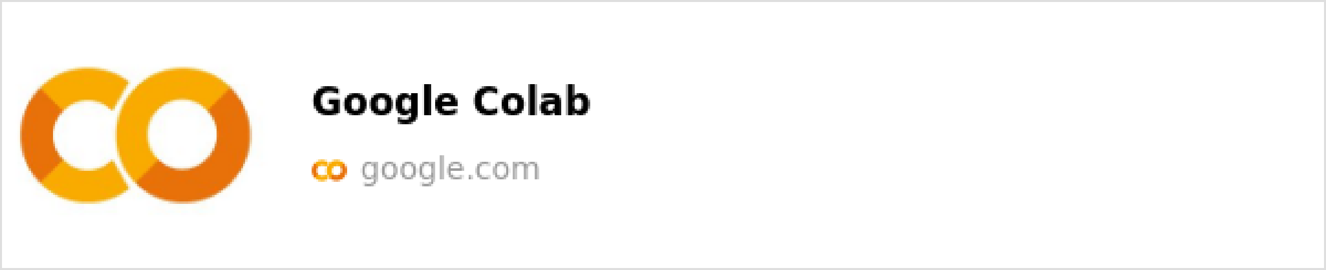
```
--------------------------------------------------
Logistic Regression Model Evaluation:
Accuracy: 0.9737
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.95      0.96        43
           1       0.97      0.99      0.98        71

    accuracy                           0.97       114
   macro avg       0.97      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114
```

CONCLUSION

- Logistic Regression achieved the highest accuracy (98.25%), followed by Random Forest (96.49%) and SVM (94.74%).
- All models performed well, but Logistic Regression proved to be the best in this case.
- The results show that machine learning can effectively aid in breast cancer diagnosis.
- Future improvements could involve using more advanced models and larger datasets.

THANK YOU

For more details, please check the complete implementation at the following link:



Google Colab

google.com

Contact me at:

https://www.linkedin.com/in/dewianr

ainidewi00@gmail.com