BREAST CANCER CLASSIFICATION WITH MACHINE LEARNING ALGORITHMS

TABLE OF CONTENTS

- About Project
- Goals
- Dataset Overview
- Exploration Data Analysis
- Models Selection
- Result
- Best Model Analysis
- Conclusion

DATE
26/02/2025

ABOUT ME

Naftalia Sophiana Purba

Data Scientist | Data Analyst | Machine Learning Enthusiast

Bali, Indonesia

Undergraduate student of Mathematics from Udayana University with a deep interest in Data Analytics, Data Science, and Machine Learning

TOOLS AND LIBRARY

NumPy

pandas

matplotlib

seaborn

scikit learn

ABOUT PROJECT

This project aims to build a machine learning model to classify breast tumors as benign or malignant using the scikit-learn dataset. The tested models include **SVM, Logistic Regression, Random Forest, KNN, and Naïve Bayes**, compared to determine the best performance in breast cancer classification.

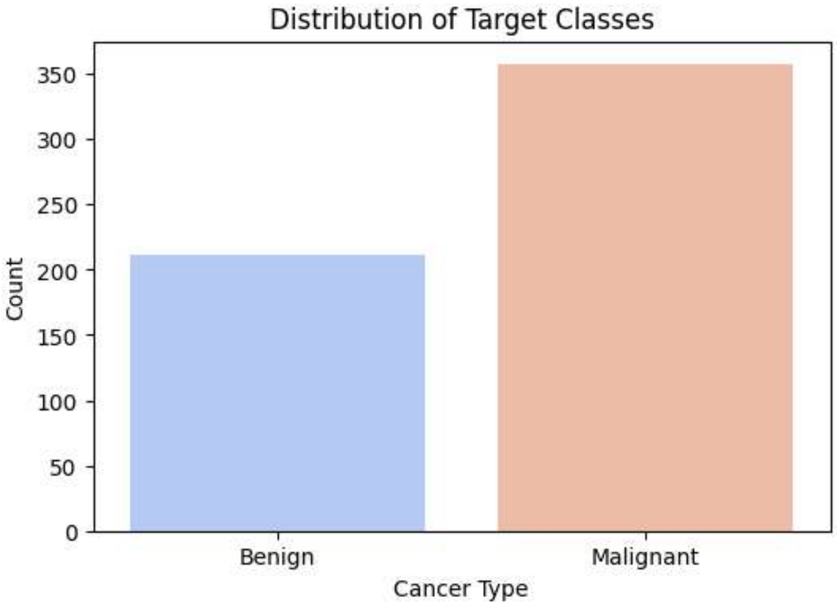
GOALS

The primary objective of this project is to develop an accurate and efficient machine learning model for breast cancer classification using the Scikit-learn dataset. By applying data preprocessing, feature analysis, and evaluating multiple classification algorithms, the project aims to enhance predictive performance and support early cancer diagnosis to assist medical professionals in making informed decisions.

DATASET OVERVIEW

• Source	: Scikit-learn
• Number of Instances	: 569
• Number of Features	: 30
• Feature Types	: Numerical
• Traget Classes	: 0 = Benign (No cancerous tumor) 1 = Malignant (Cancerous tumor)

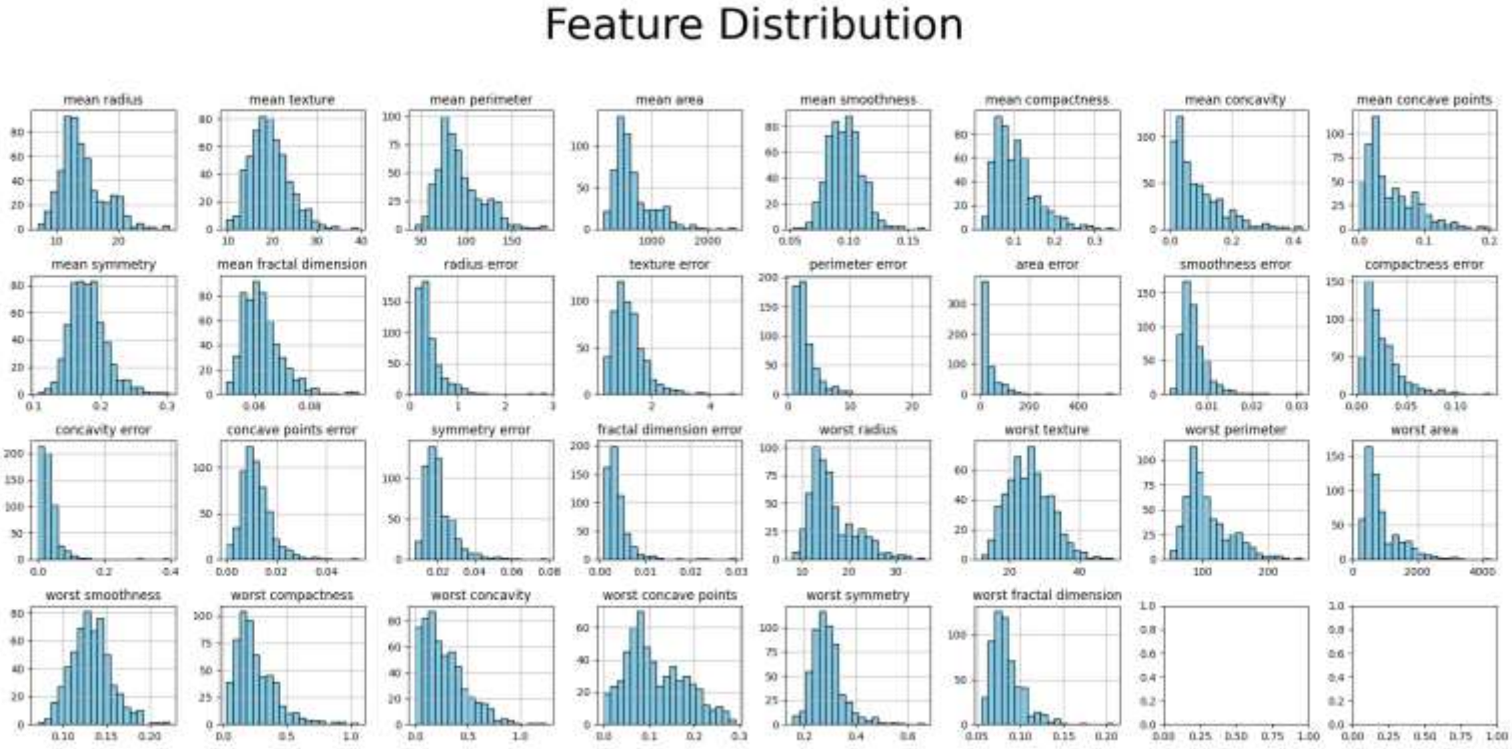
EXPLORATION DATA ANALYSIS



The class distribution in the graph shows that the number of **malignant (cancerous) cases** is higher than **benign (non-cancerous) cases**. This provides insight that breast cancer is more frequently detected as malignant in this dataset.

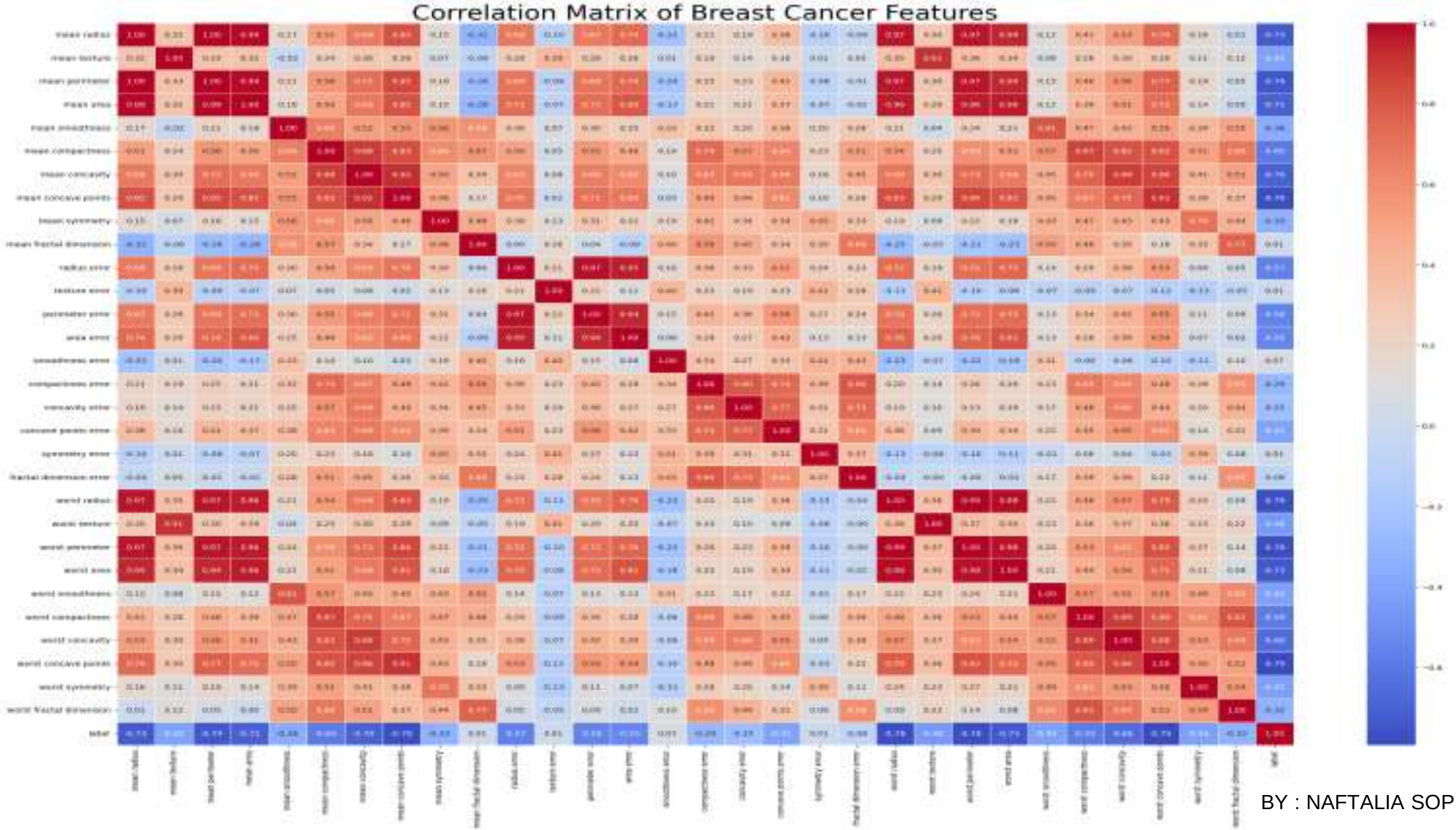
BY : NAFTALIA SOPHIANA PURBA

EXPLORATION DATA ANALYSIS



BY : NAFTALIA SOPHIANA PURBA

EXPLORATION DATA ANALYSIS

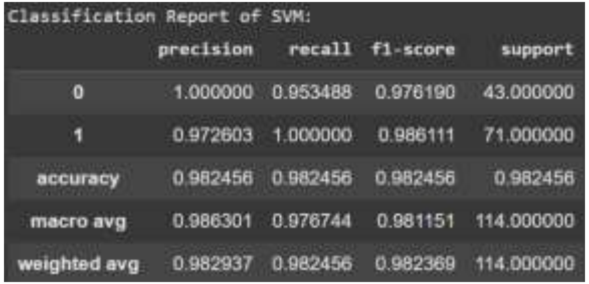


BY : NAFTALIA SOPHIANA PURBA

MODEL SELECTION

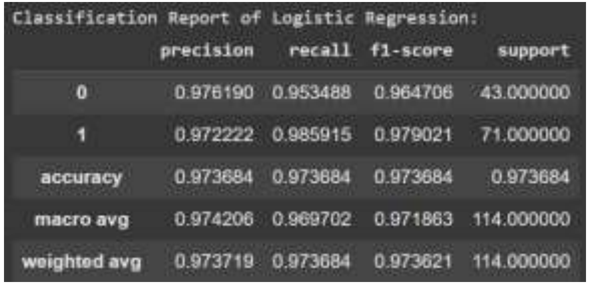
- Support Vector Machine (SVM)
- Logistic Regression
- Random Forest
- K-Nearesr Neighbour
- Naive Bayes

RESULT



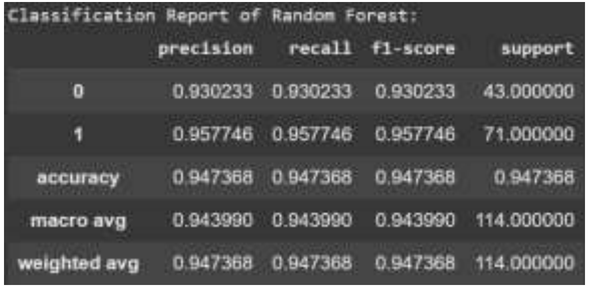
Classification Report of SVM:

	precision	recall	f1-score	support
0	1.000000	0.953488	0.976190	43.000000
1	0.972603	1.000000	0.986111	71.000000
accuracy	0.982456	0.982456	0.982456	0.982456
macro avg	0.986301	0.976744	0.981151	114.000000
weighted avg	0.982937	0.982456	0.982369	114.000000



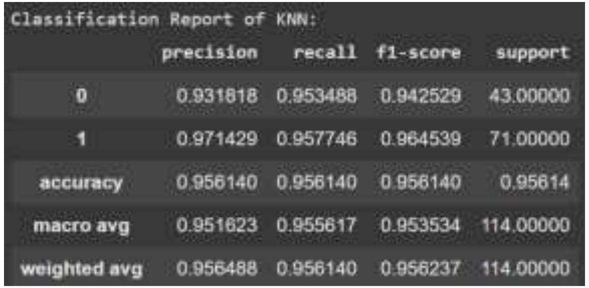
Classification Report of Logistic Regression:

	precision	recall	f1-score	support
0	0.976190	0.953488	0.964706	43.000000
1	0.972222	0.985915	0.979021	71.000000
accuracy	0.973684	0.973684	0.973684	0.973684
macro avg	0.974206	0.969702	0.971863	114.000000
weighted avg	0.973719	0.973684	0.973621	114.000000



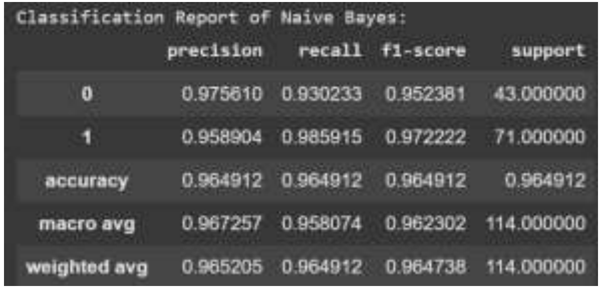
Classification Report of Random Forest:

	precision	recall	f1-score	support
0	0.930233	0.930233	0.930233	43.000000
1	0.957746	0.957746	0.957746	71.000000
accuracy	0.947368	0.947368	0.947368	0.947368
macro avg	0.943990	0.943990	0.943990	114.000000
weighted avg	0.947368	0.947368	0.947368	114.000000



Classification Report of KNN:

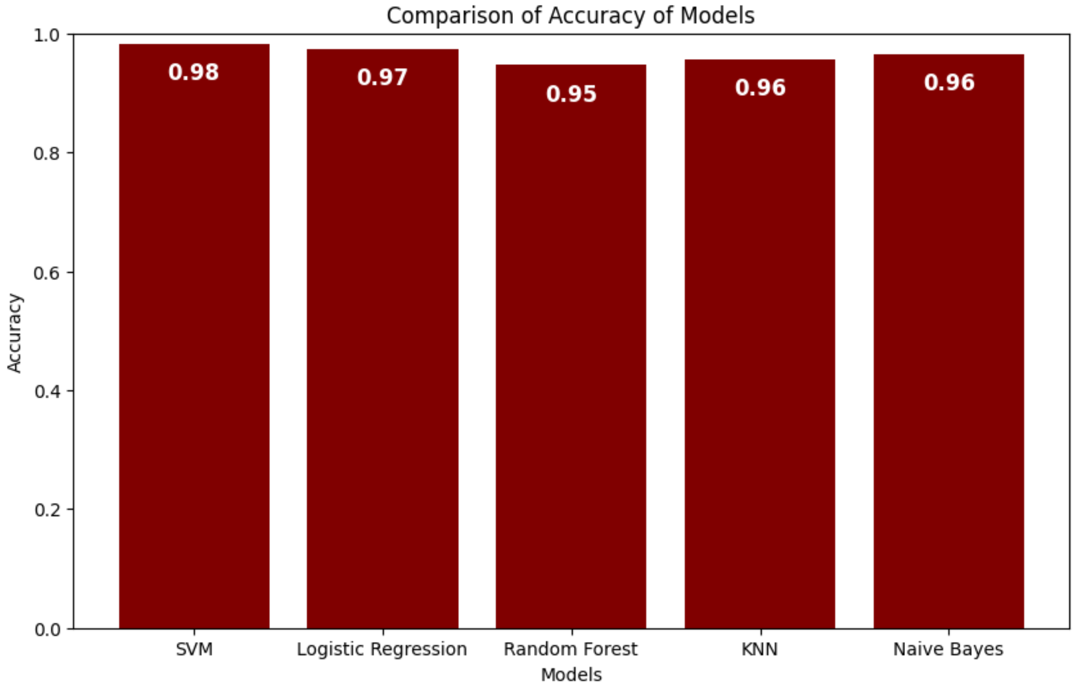
	precision	recall	f1-score	support
0	0.931818	0.953488	0.942529	43.00000
1	0.971429	0.957746	0.964539	71.00000
accuracy	0.956140	0.956140	0.956140	0.95614
macro avg	0.951623	0.955617	0.953534	114.00000
weighted avg	0.956488	0.956140	0.956237	114.00000



Classification Report of Naive Bayes:

	precision	recall	f1-score	support
0	0.975610	0.930233	0.952381	43.000000
1	0.958904	0.985915	0.972222	71.000000
accuracy	0.964912	0.964912	0.964912	0.964912
macro avg	0.967257	0.958074	0.962302	114.000000
weighted avg	0.965205	0.964912	0.964738	114.000000

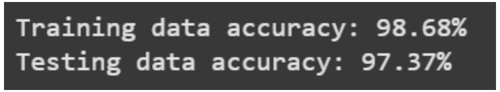
RESULT



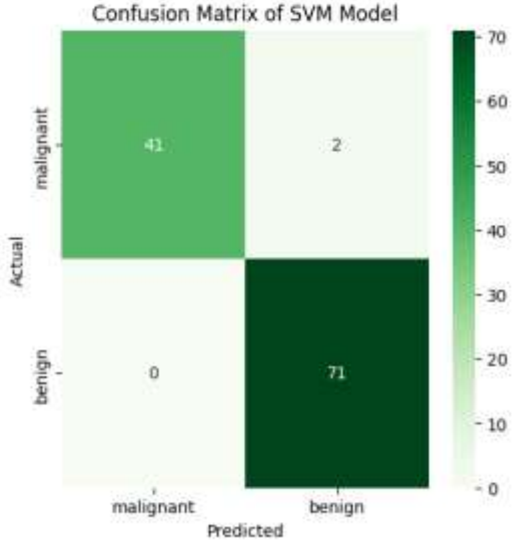
Based on the evaluation results of the five models, it was found that the Support Vector Machine has a higher accuracy compared to the other four models, namely Logistic Regression, Random Forest, KNN, and Naive Bayes. Therefore, the next step is to conduct a misclassification analysis and overfitting vs underfitting check.

BEST MODEL ANALYSIS

BEST MODEL ANALYSIS



```
Training data accuracy: 98.68%
Testing data accuracy: 97.37%
```



The confusion matrix shows that the SVM model performs well in classifying breast cancer cases. It correctly identifies 41 malignant and 71 benign cases, with only 2 false negatives and no false positives. The high training (98.68%) and testing (97.37%) accuracy indicate that the model generalizes well without signs of overfitting or underfitting.

CONCLUSION

Breast cancer classification was performed using five machine learning models: SVM, Logistic Regression, Random Forest, KNN, and Naïve Bayes. SVM achieved the highest accuracy (98.25%), making it the best model.

Misclassification analysis showed minimal errors (2 false negatives, 0 false positives). The training (98.68%) and testing (97.37%) accuracies were close, indicating no overfitting or underfitting.

DATE
26/02/2025

CONTACT

Email	: naftaliasphn94@gmail.com
GitHub	: https://github.com/naftaliasphna

BREAST CANCER CLASSIFICATION: MACHINE LEARNING APPROACH

BY : NAFTALIA SOPHIANA PURBA