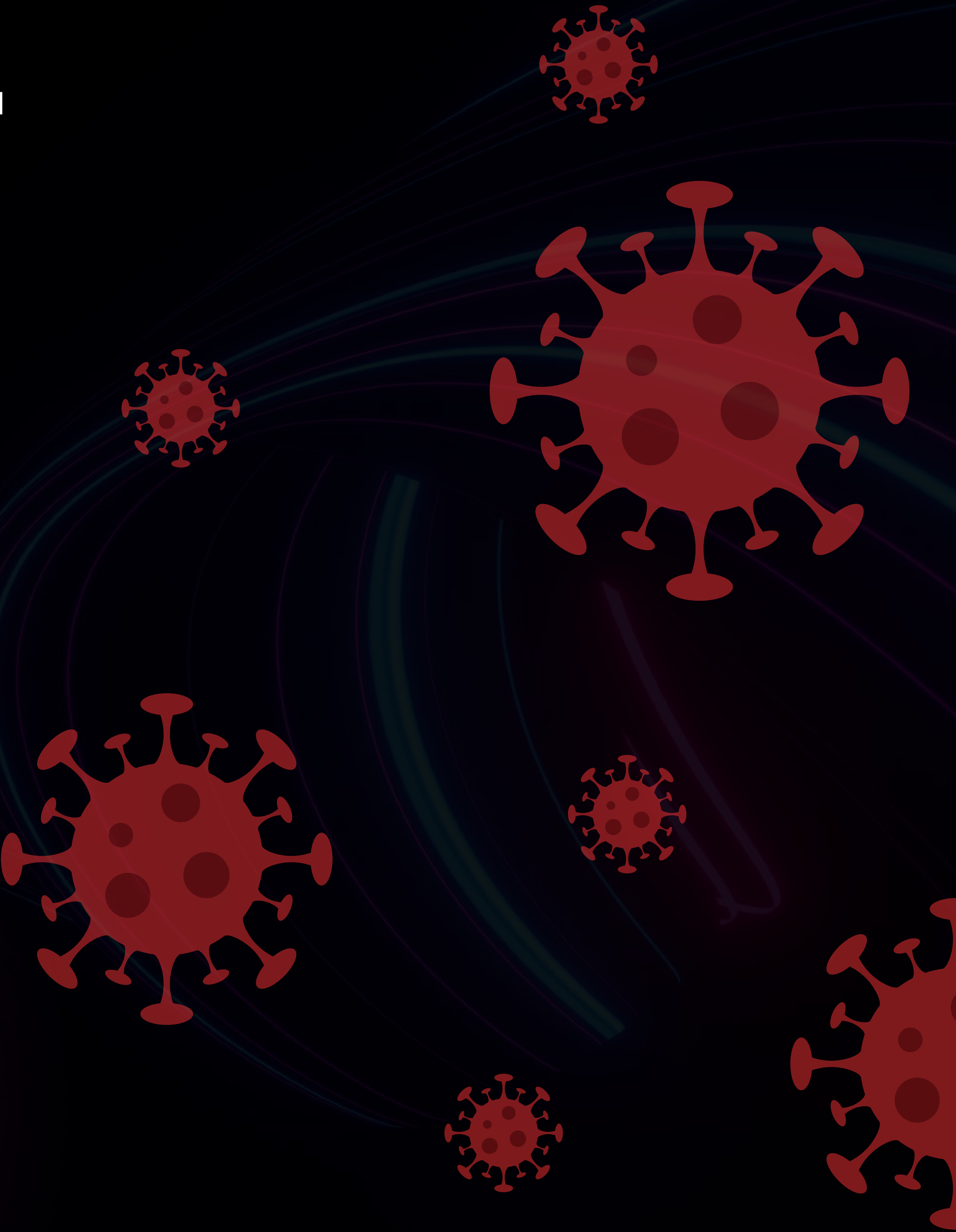
Uncovering Breast Cancer Patterns: A Study on Classification with Machine Learning

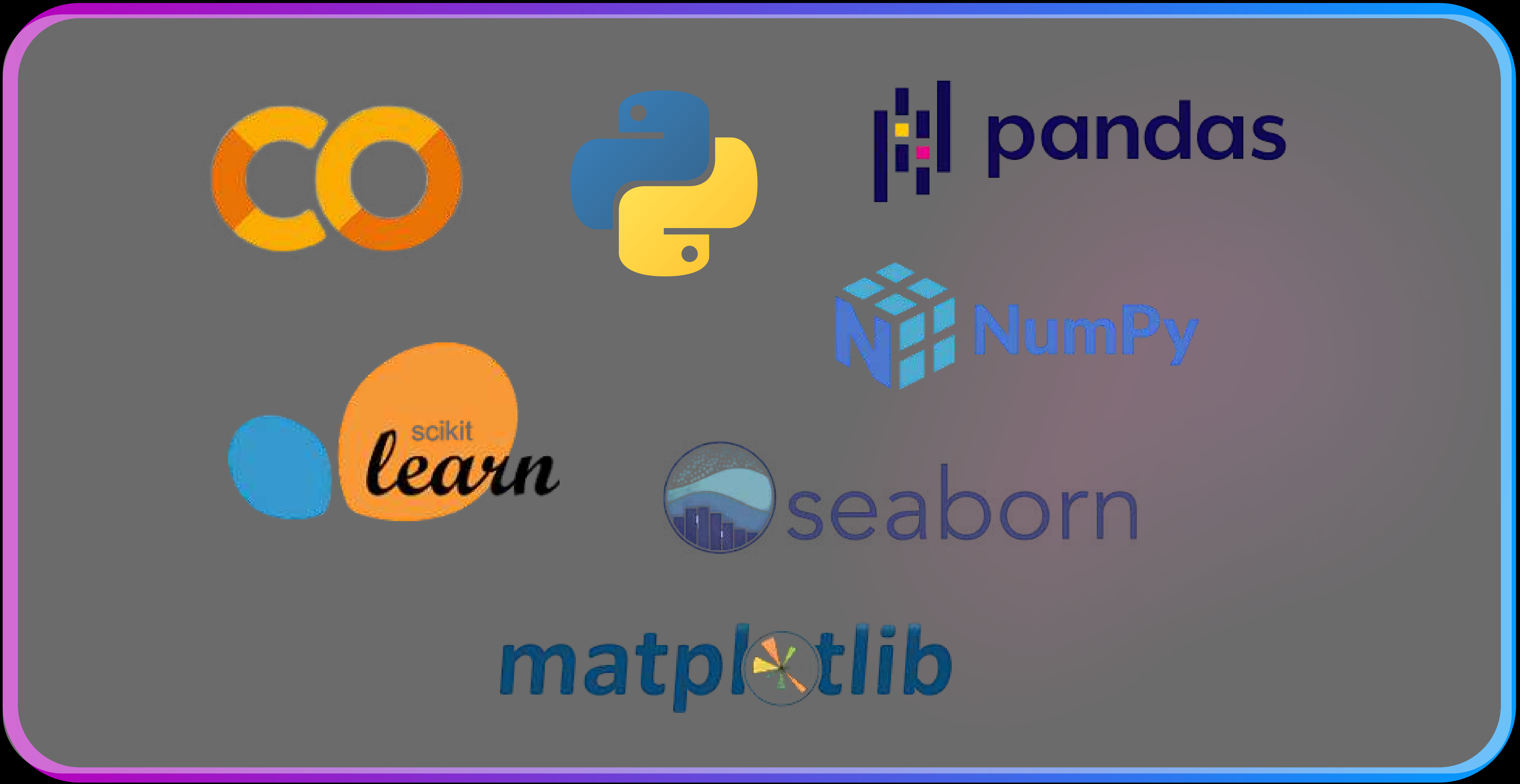
Breast cancer is a leading cause of mortality among women, highlighting the need for early detection. This study uses machine learning to classify breast cancer as benign or malignant, aiming to improve diagnostic accuracy.

Start Slide ▶

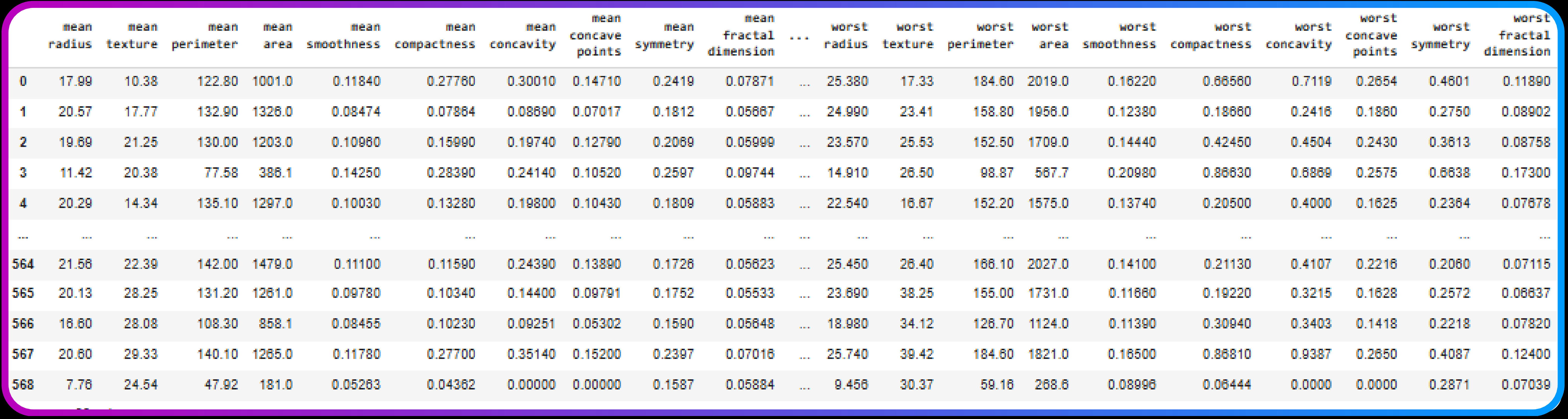


Presented by Ainun Khoirunni’mah

TOOLS



About Dataset

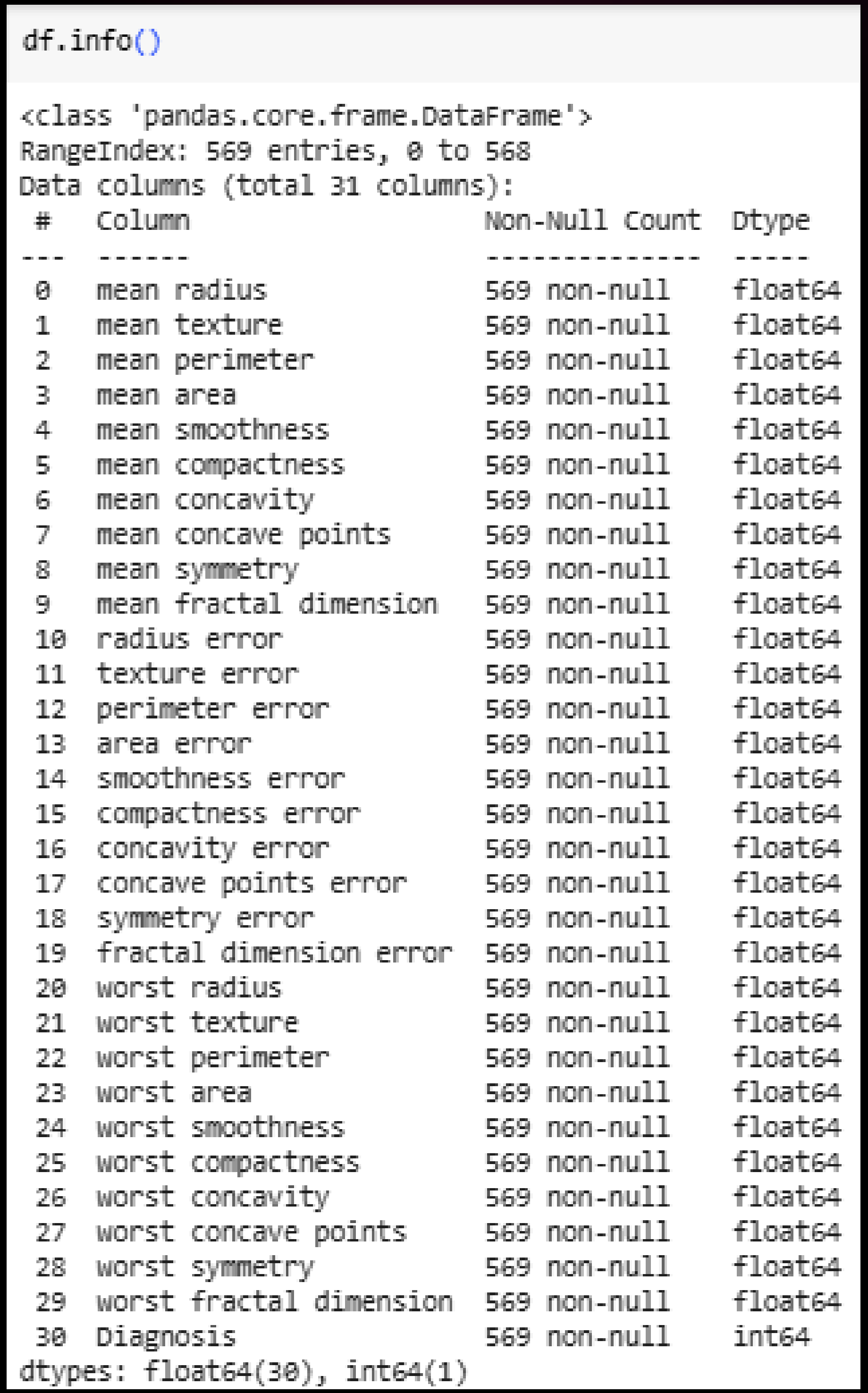


	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...	worst radius	worst texture	worst perimeter	worst area	worst smoothness	worst compactness	worst concavity	worst concave points	worst symmetry	worst fractal dimension
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	0.07871	...	25.380	17.33	184.60	2019.0	0.16220	0.66560	0.7119	0.2654	0.4601	0.11890
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	0.05667	...	24.990	23.41	158.80	1956.0	0.12380	0.18660	0.2416	0.1860	0.2750	0.08902
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	0.05999	...	23.570	25.53	152.50	1709.0	0.14440	0.42450	0.4504	0.2430	0.3613	0.08758
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	0.09744	...	14.910	26.50	98.87	567.7	0.20980	0.86630	0.6869	0.2575	0.6638	0.17300
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	0.05883	...	22.540	16.67	152.20	1575.0	0.13740	0.20500	0.4000	0.1625	0.2364	0.07678
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	0.05623	...	25.450	26.40	166.10	2027.0	0.14100	0.21130	0.4107	0.2216	0.2060	0.07115
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	0.05533	...	23.690	38.25	155.00	1731.0	0.11660	0.19220	0.3215	0.1628	0.2572	0.06637
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	0.05648	...	18.980	34.12	126.70	1124.0	0.11390	0.30940	0.3403	0.1418	0.2218	0.07820
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	0.07016	...	25.740	39.42	184.60	1821.0	0.16500	0.86810	0.9387	0.2650	0.4087	0.12400
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	0.05884	...	9.456	30.37	59.16	268.6	0.08996	0.06444	0.0000	0.0000	0.2871	0.07039

The Breast Cancer Wisconsin (Classification) dataset is a widely used binary classification dataset, often utilized for testing machine learning models. This dataset is readily available in **Scikit-learn**, a powerful and open-source machine learning library in Python.

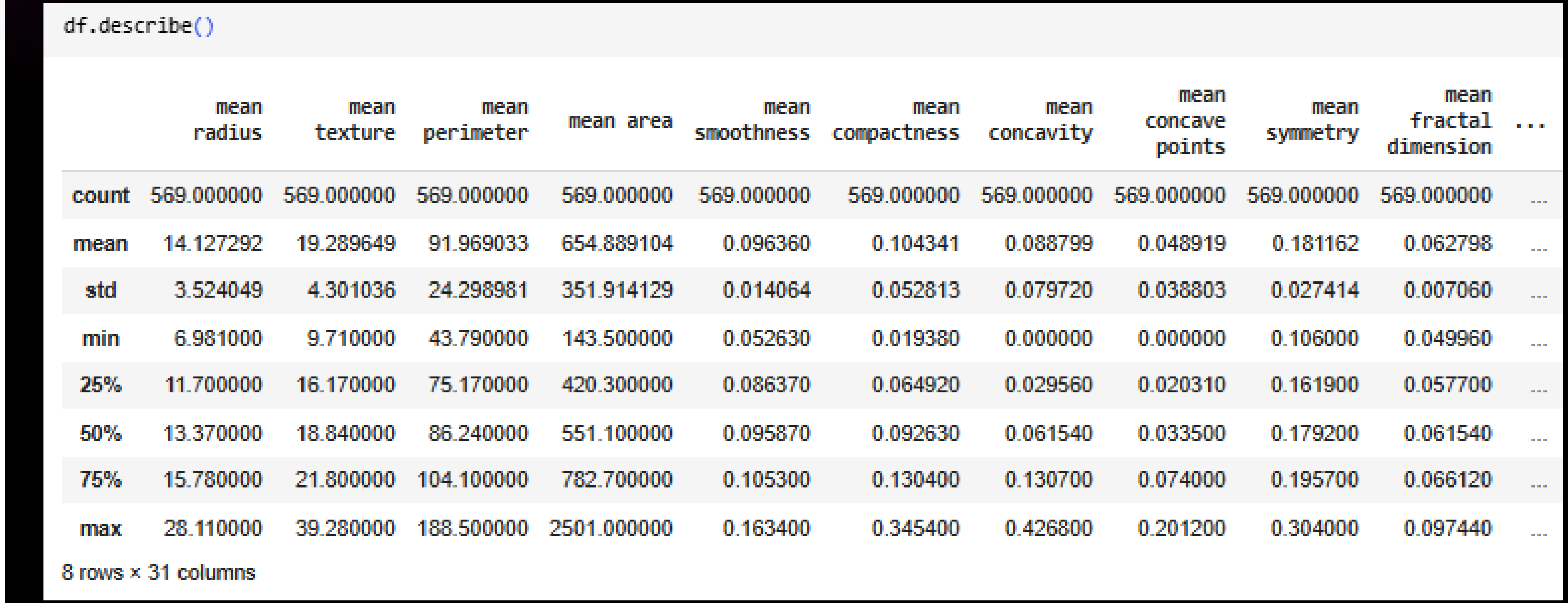
- **Type** : Binary Classification
- **Total Samples** : 569
- **Class Distribution** :
  - Malignant (M) : 212 samples
  - Benign (B) : 357 samples
- **Dimensionality** : 30 features

Exploratory Data & Analysis (EDA)



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
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  Diagnosis                569 non-null    int64  
dtypes: float64(30), int64(1)
```



```
df.describe()
```

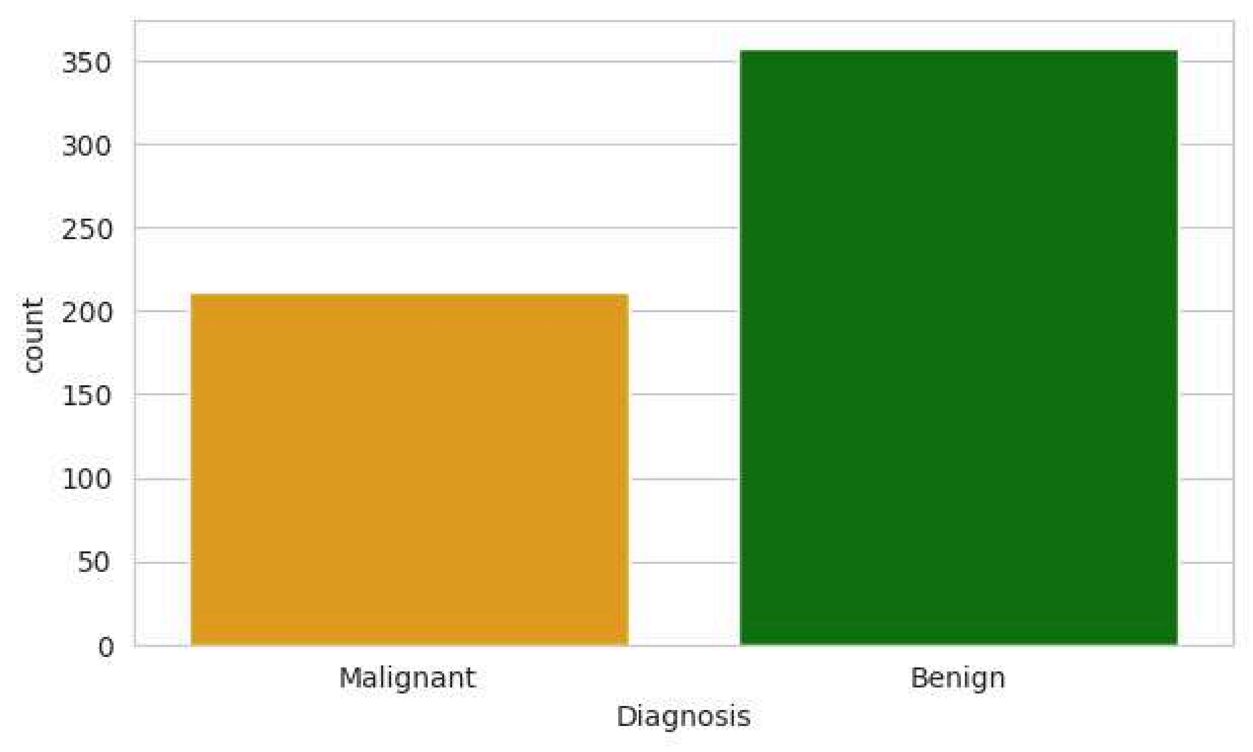
	mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	...
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.181162	0.062798	...
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.027414	0.007060	...
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000	0.049960	...
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900	0.057700	...
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.179200	0.061540	...
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700	0.066120	...
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.304000	0.097440	...

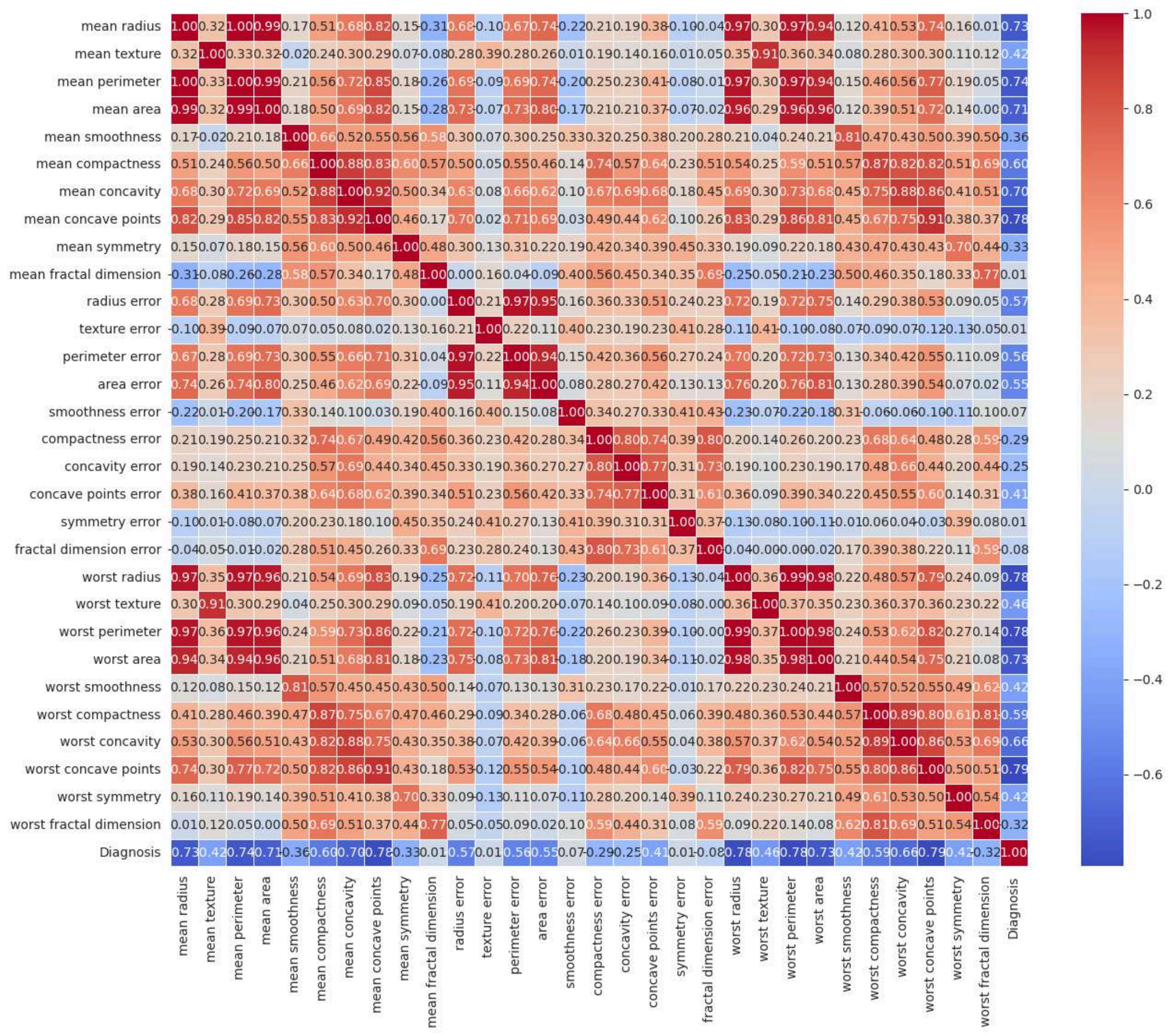
8 rows × 31 columns

Exploratory Data Analysis (EDA) is a crucial initial step in data analysis that involves exploring and understanding the dataset using statistical summaries and visualizations.

Exploratory Data & Analysis (EDA)

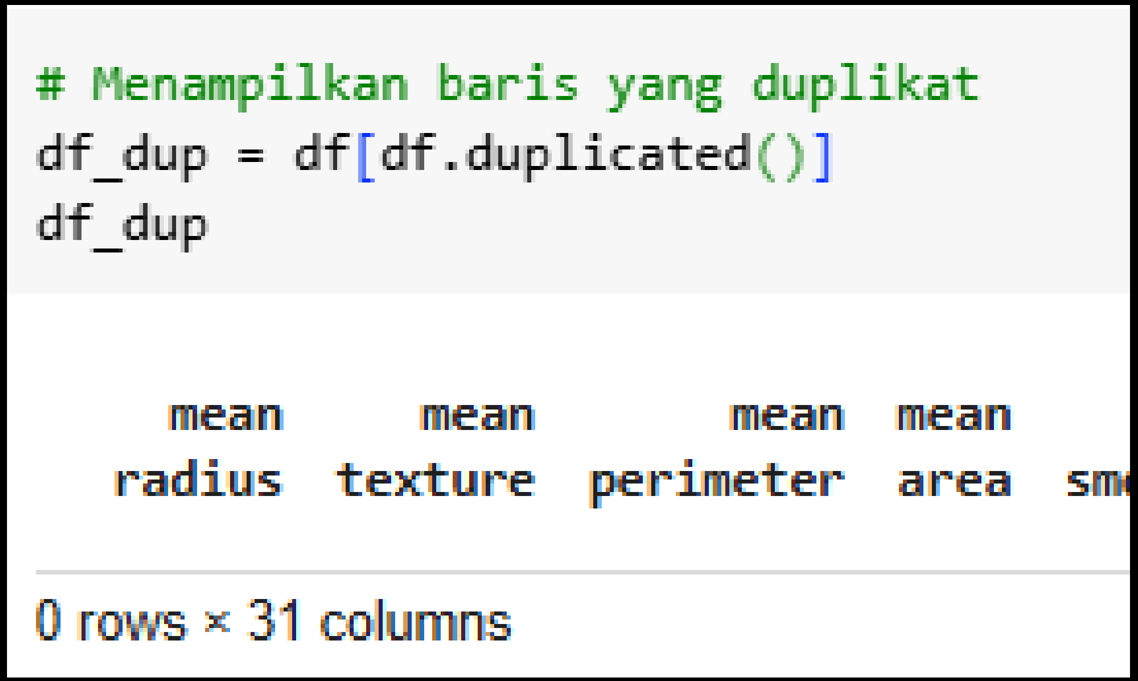
The **bar chart** shows the class distribution between Malignant and Benign categories, highlighting potential class imbalances. **The correlation heatmap** visualizes the relationships between features, helping identify highly correlated variables for better feature selection and multicollinearity management. Together, these visualizations provide insights to guide data preprocessing and modeling.





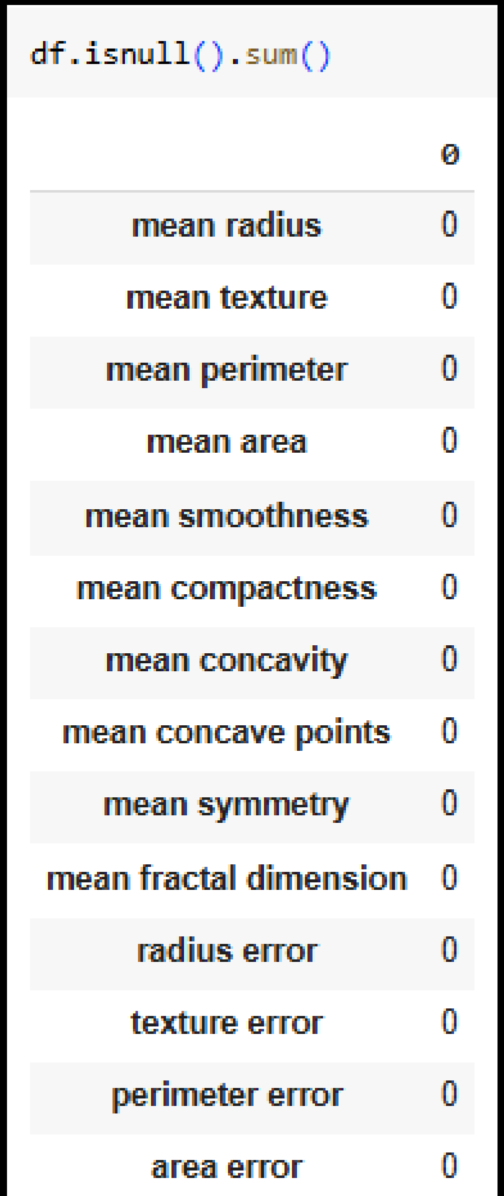
Data Prepocessing

Duplicate & Missing Values



```
# Menampilkan baris yang duplikat
df_dup = df[df.duplicated()]
df_dup
```

```
   mean     mean       mean    mean
 radius  texture  perimeter   area  sm
0 rows × 31 columns
```

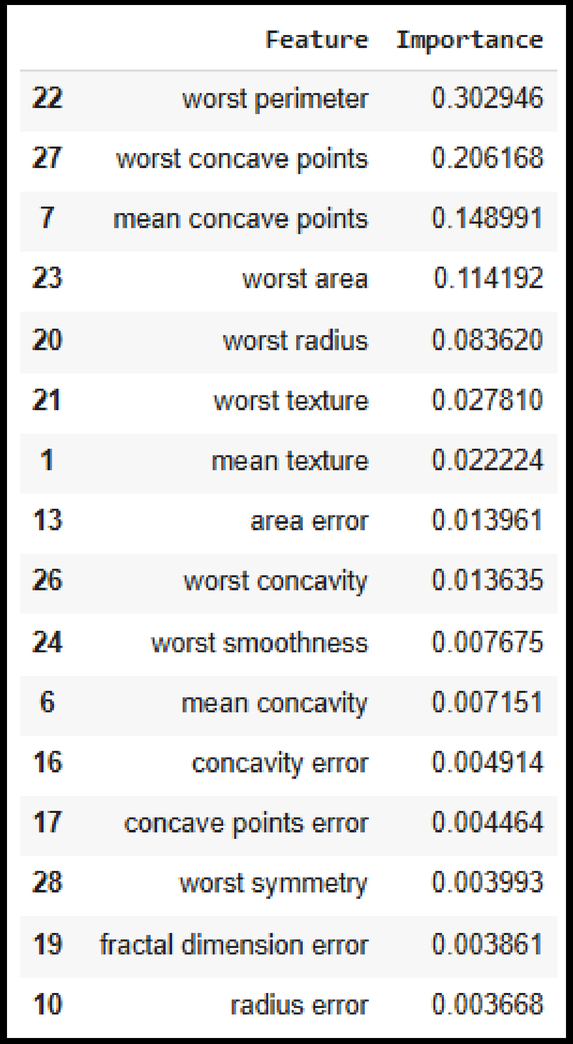


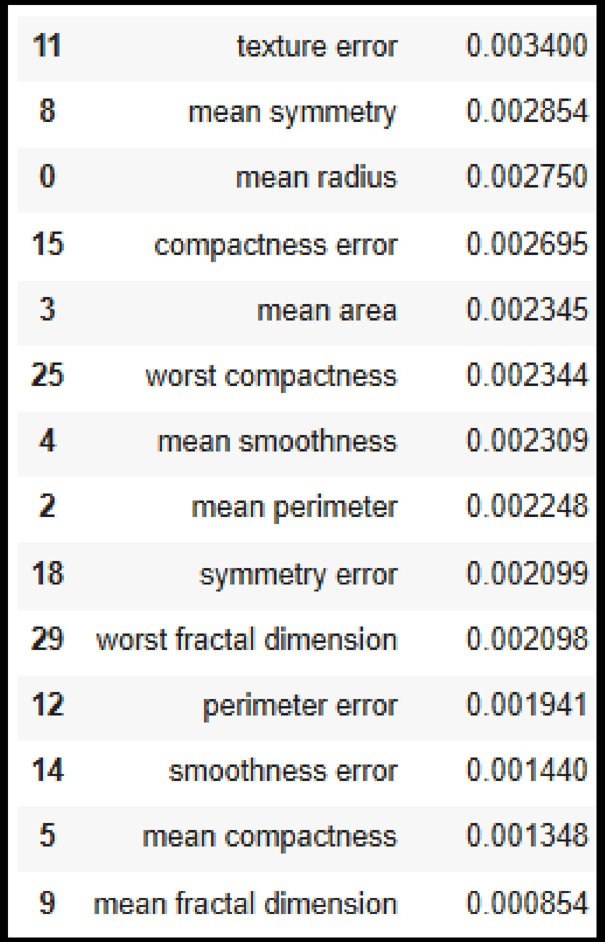
```
df.isnull().sum()
```

	0
mean radius	0
mean texture	0
mean perimeter	0
mean area	0
mean smoothness	0
mean compactness	0
mean concavity	0
mean concave points	0
mean symmetry	0
mean fractal dimension	0
radius error	0
texture error	0
perimeter error	0
area error	0

The data preprocessing involved checking for duplicate rows and missing values, with the results **showing no duplicates** and **no missing data** across all 31 features. This indicates that the dataset is clean and ready for the next steps in the machine learning workflow.

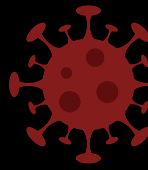
Feature Importance

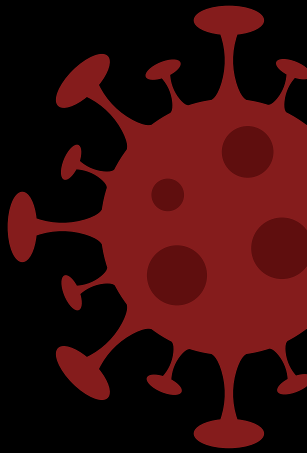




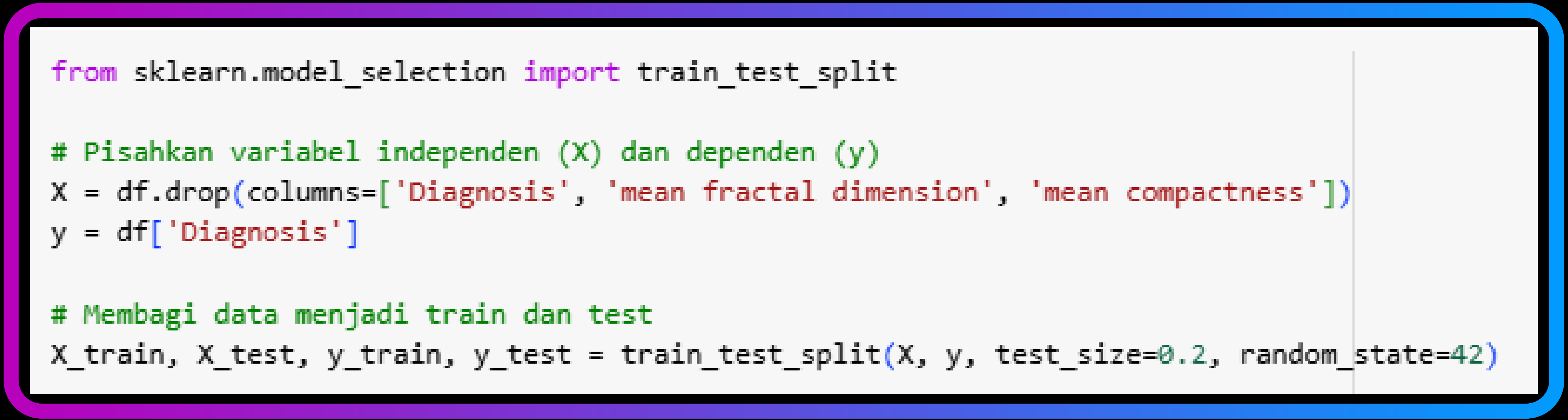
	Feature	Importance
22	worst perimeter	0.302946
27	worst concave points	0.206168
7	mean concave points	0.148991
23	worst area	0.114192
20	worst radius	0.083620
21	worst texture	0.027810
1	mean texture	0.022224
13	area error	0.013961
26	worst concavity	0.013635
24	worst smoothness	0.007675
6	mean concavity	0.007151
16	concavity error	0.004914
17	concave points error	0.004464
28	worst symmetry	0.003993
19	fractal dimension error	0.003861
10	radius error	0.003668
11	texture error	0.003400
8	mean symmetry	0.002854
0	mean radius	0.002750
15	compactness error	0.002695
3	mean area	0.002345
25	worst compactness	0.002344
4	mean smoothness	0.002309
2	mean perimeter	0.002248
18	symmetry error	0.002099
29	worst fractal dimension	0.002098
12	perimeter error	0.001941
14	smoothness error	0.001440
5	mean compactness	0.001348
9	mean fractal dimension	0.000854

Feature importance identifies which features most influence the model's predictions, helping improve model performance by focusing on the most relevant features.





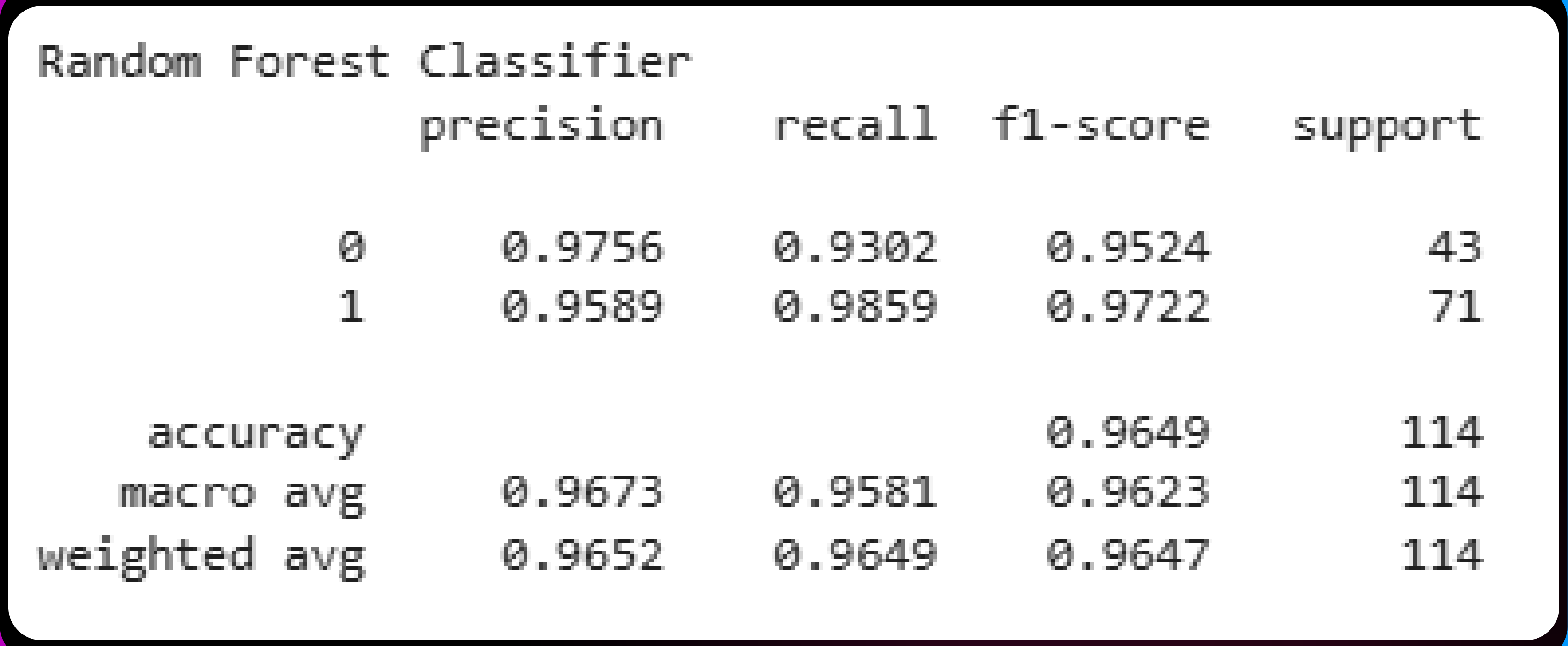
Splitting Data



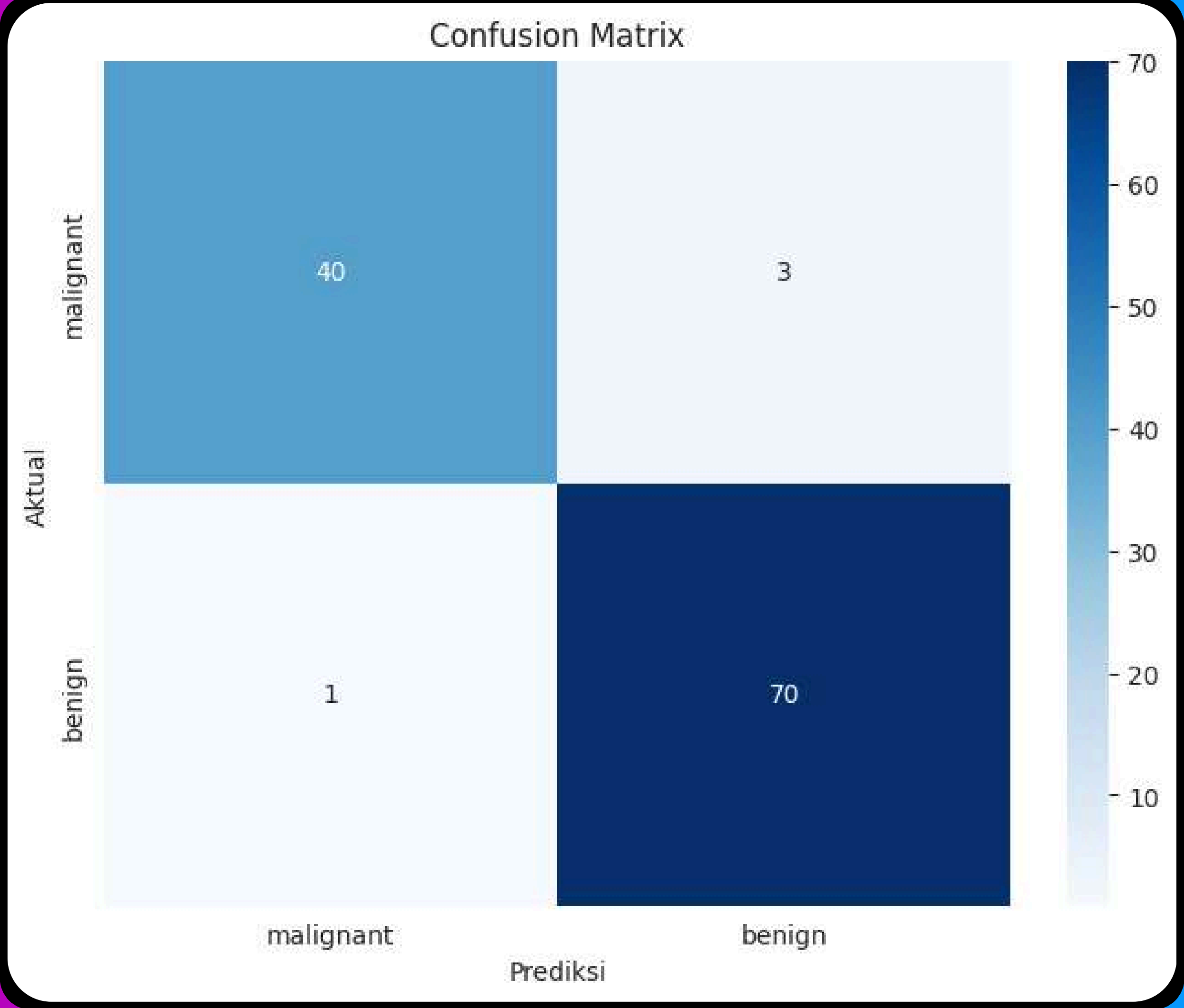
```
from sklearn.model_selection import train_test_split

# Pisahkan variabel independen (X) dan dependen (y)
X = df.drop(columns=['Diagnosis', 'mean fractal dimension', 'mean compactness'])
y = df['Diagnosis']

# Membagi data menjadi train dan test
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

The code splits the dataset into **independent variables (X) and the target variable (y), excluding irrelevant features to improve model performance.** It then divides the data into training and testing sets, with **80% of the data used for training and 20% for testing**, ensuring randomness and reproducibility with a random state of 42. This step is crucial for model evaluation, allowing the model to learn from the training data and validate its performance on unseen data.

Modelling - Random Forest

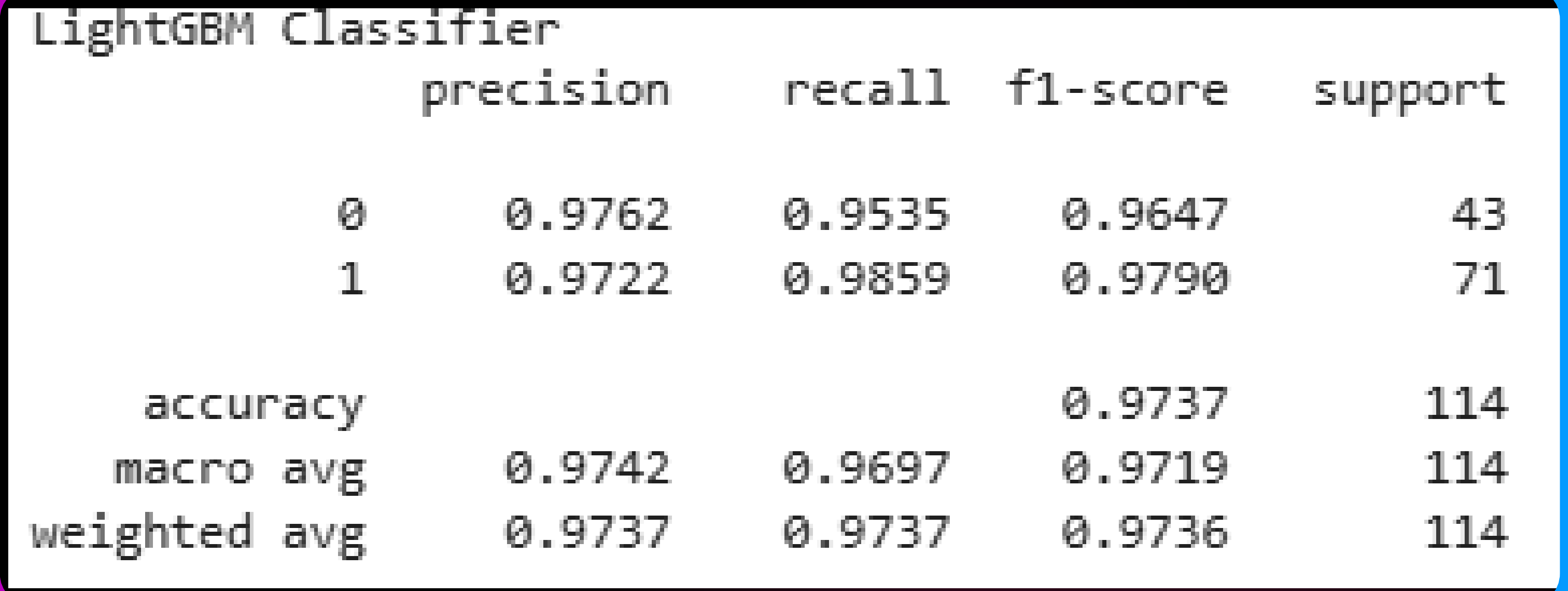


```
Random Forest Classifier
              precision    recall  f1-score   support

           0     0.9756    0.9302    0.9524        43
           1     0.9589    0.9859    0.9722        71

    accuracy                         0.9649       114
   macro avg     0.9673    0.9581    0.9623       114
weighted avg     0.9652    0.9649    0.9647       114
```

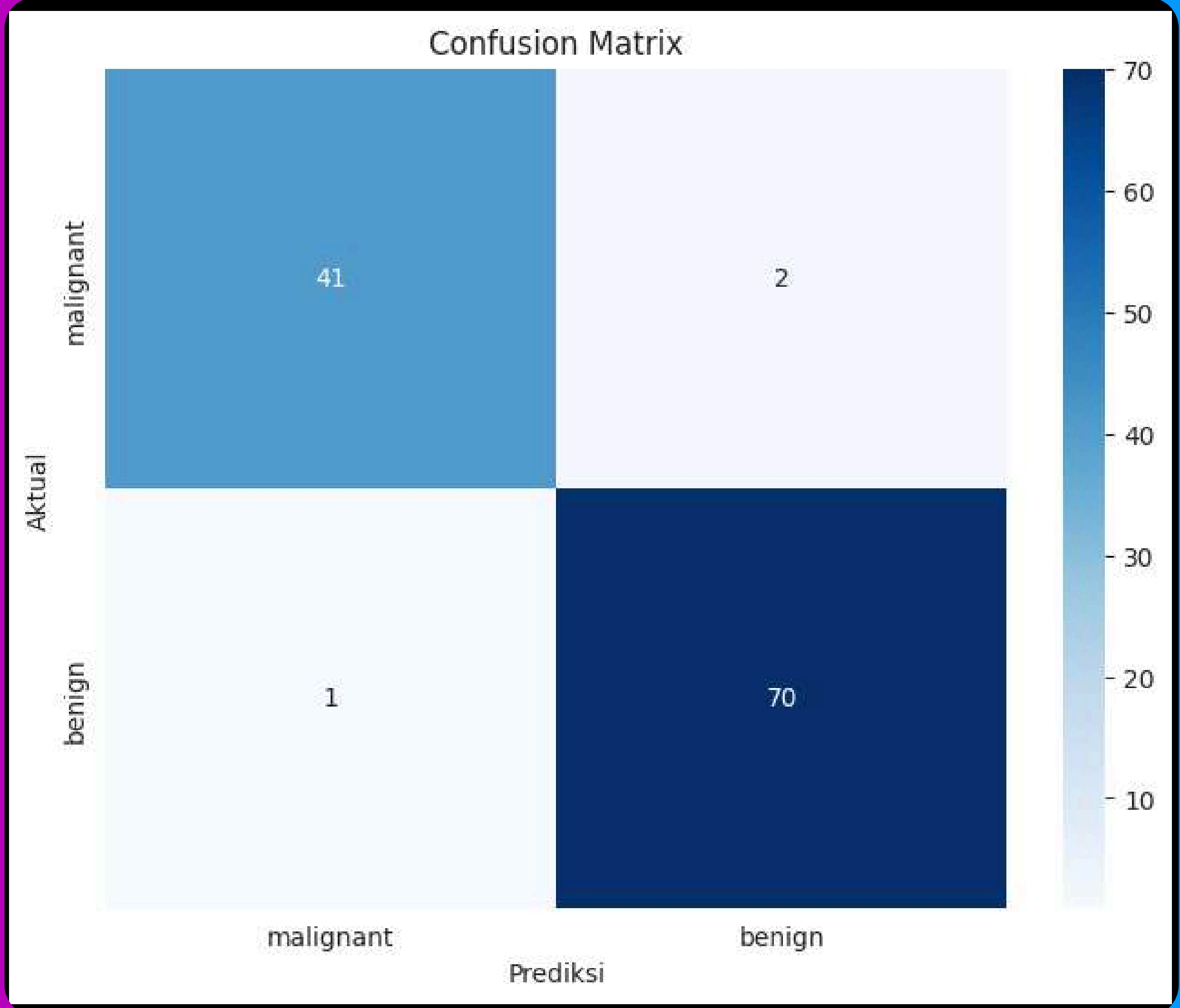


The **Random Forest Classifier** model achieved an accuracy of **96.49%** in classifying breast cancer cases as malignant or benign. This high accuracy indicates the model's strong ability to make correct predictions, with only **4 misclassifications out of 114 samples**, demonstrating its reliability for early breast cancer detection.

Modelling - LightGBM



```
LightGBM Classifier
              precision    recall  f1-score   support

           0     0.9762    0.9535    0.9647        43
           1     0.9722    0.9859    0.9790        71

    accuracy                         0.9737       114
   macro avg     0.9742    0.9697    0.9719       114
weighted avg     0.9737    0.9737    0.9736       114
```



The **LightGBM Classifie**r demonstrated a strong performance with an accuracy of **97.37%** in distinguishing between malignant and benign breast cancer cases. With only **3 misclassifications out of 114 samples**, the model proves to be highly effective in supporting accurate and timely breast cancer diagnosis.

Conclusion

The breast cancer classification analysis using machine learning demonstrated promising results. The data preprocessing stage confirmed **a clean dataset with no duplicates or missing values**, providing a solid foundation for modeling. Feature importance analysis helped identify key predictors, allowing the model to focus on relevant features. The dataset was split into training and testing sets for robust model evaluation.

Two models were tested: the **Random Forest Classifier achieved an accuracy of 96.49%**, while the **LightGBM Classifier outperformed with 97.37% accuracy,** demonstrating high precision and recall in distinguishing malignant and benign cases. The confusion matrices showed minimal misclassifications, highlighting the models' reliability in early breast cancer detection. Interestingly, the application of SMOTE (Synthetic Minority Over-sampling Technique) and standardization resulted in lower accuracy, indicating that the original data distribution and scaling provided better performance. Overall, the LightGBM model emerged as the best approach, offering accurate and consistent predictions to support early diagnosis and improved patient outcomes.

2025 Home Tools Main Contact

Thank You!

FOR YOUR ATTENTION

Email
ainunkhoirunnimah@mai.ugm.ac.id

LinkedIn
linkedin.com/in/ainun-khoirunnimah