Deepiotics

Table of contents

1. About Problem Statement and Data Set

2. Exploratory Data Analysis

3. Data Preprocessing

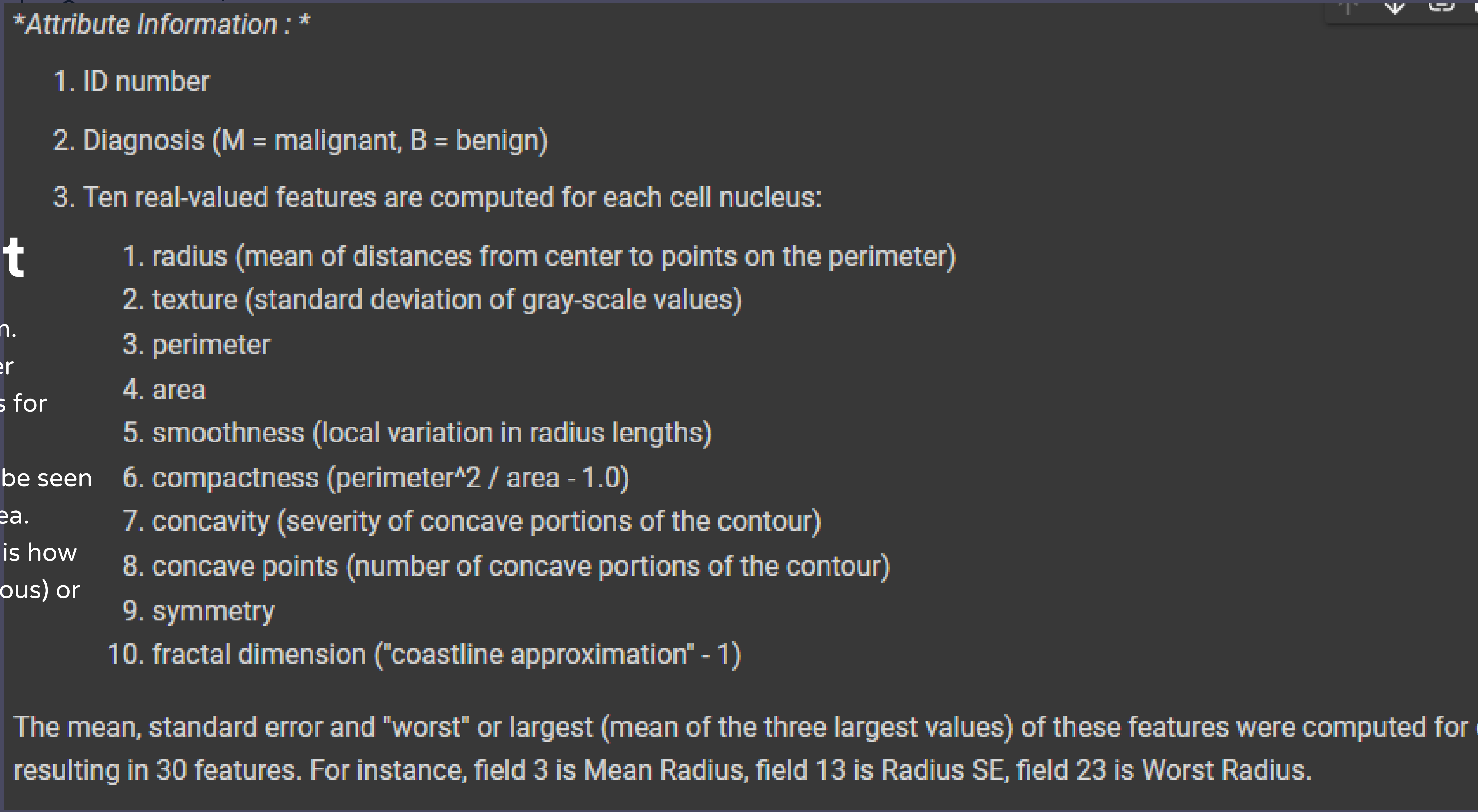
2. Predictions & ML models Accuracy

Problem Statement

- Its a Breast Cancer Classification Problem.
- Breast cancer is the most common cancer amongst women in the world. It accounts for 25% of all cancer cases
- These cells usually form tumors that can be seen via X-ray or felt as lumps in the breast area.
- The key challenges against it’s detection is how to classify tumors into malignant (cancerous) or benign(non cancerous).

Data Source : https://www.k

About Data Set

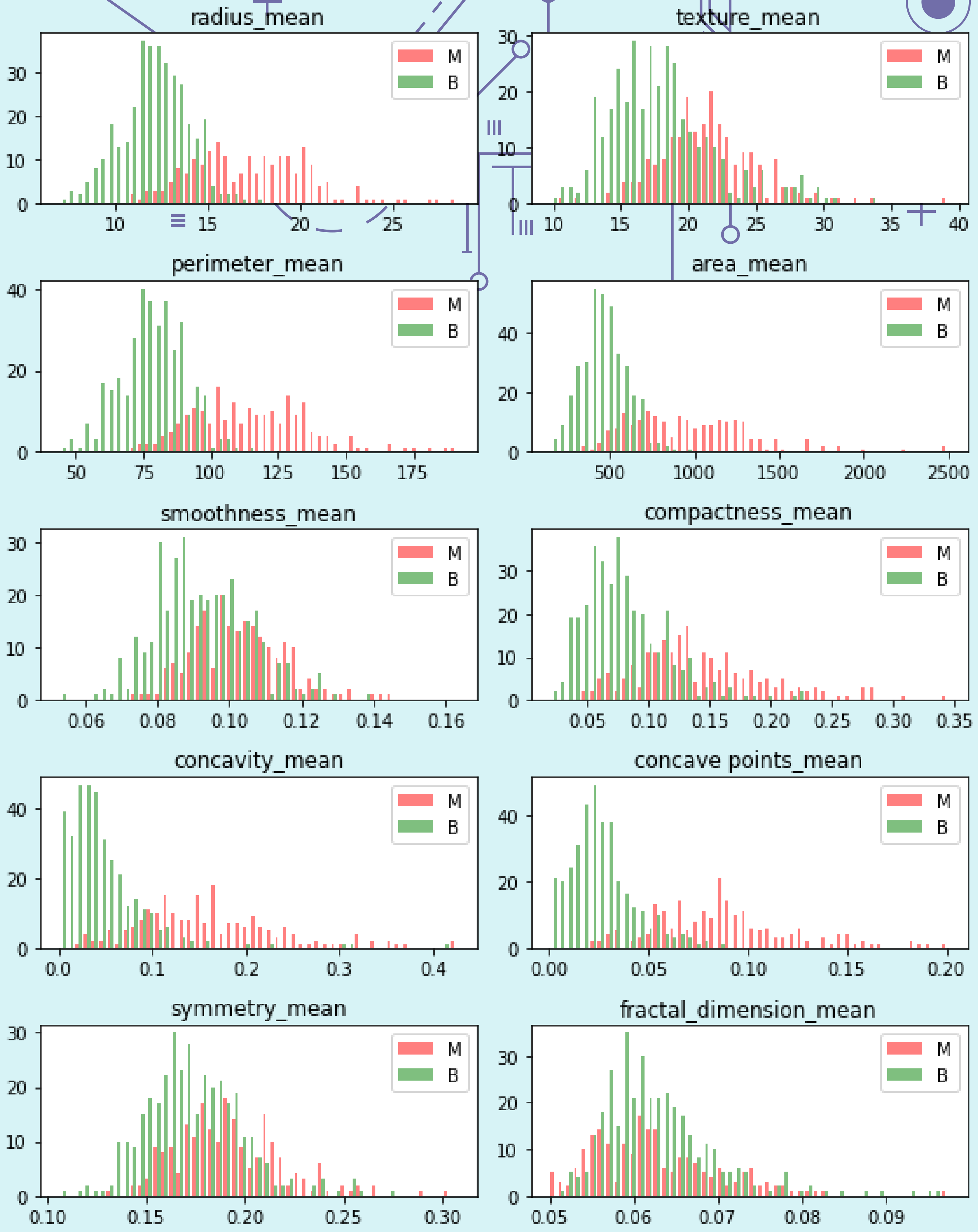


*Attribute Information : *

1. ID number
2. Diagnosis (M = malignant, B = benign)
3. Ten real-valued features are computed for each cell nucleus:
    1. radius (mean of distances from center to points on the perimeter)
    2. texture (standard deviation of gray-scale values)
    3. perimeter
    4. area
    5. smoothness (local variation in radius lengths)
    6. compactness (perimeter^2 / area - 1.0)
    7. concavity (severity of concave portions of the contour)
    8. concave points (number of concave portions of the contour)
    9. symmetry
    10. fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

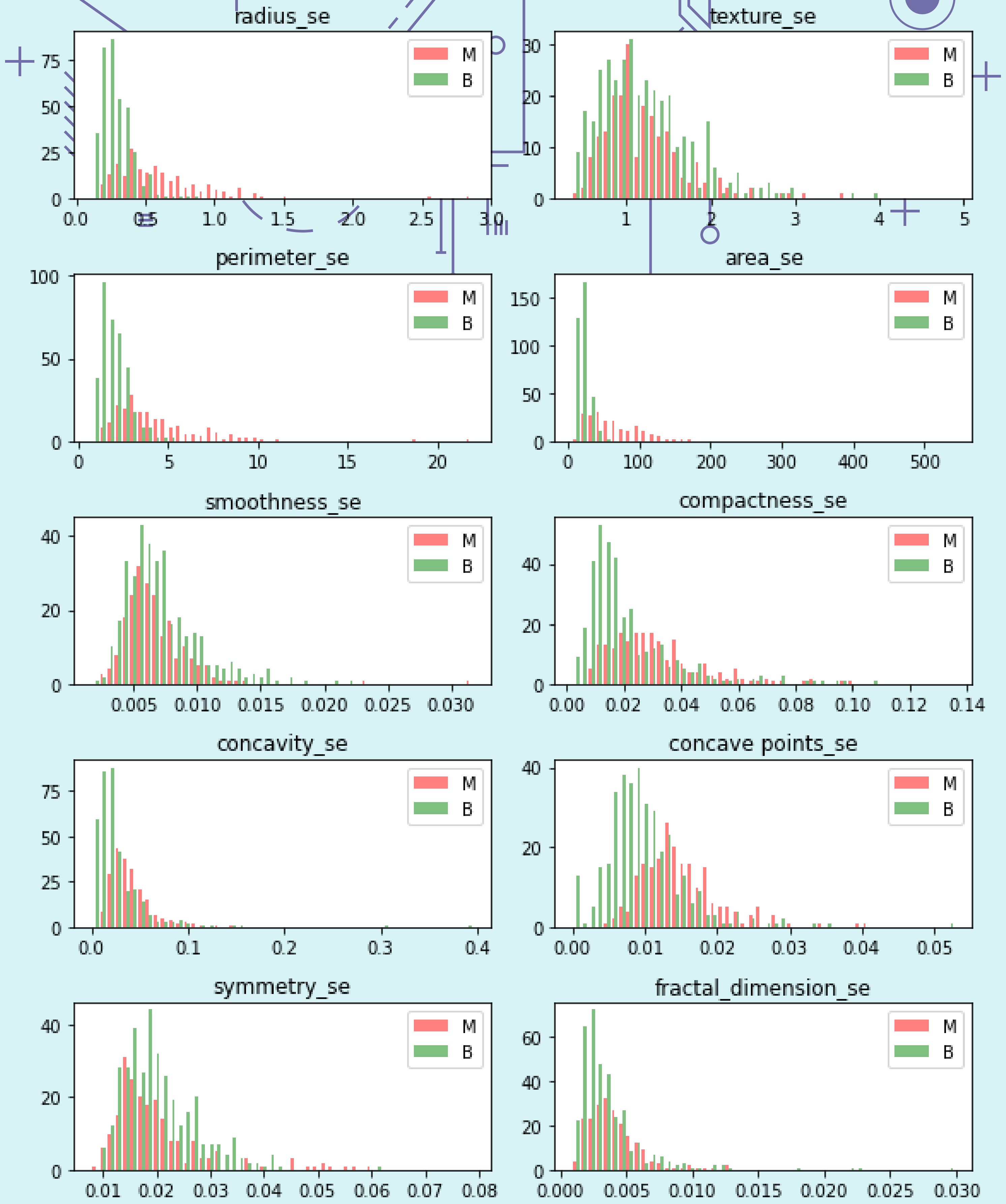
EXPLORATORY DATA ANALYSIS



MEAN OF FEATURES

*   It can be observed that Larger the mean values of radius" , concave points" , concavity" , "compactness, area and perimeter" shows correlation with Malignant tumors.
So these features can be used for classification.
*   On the other hand mean values of texture, smoothness, symmetry or fractual dimension does not show a particular preference of one diagnosis over the other

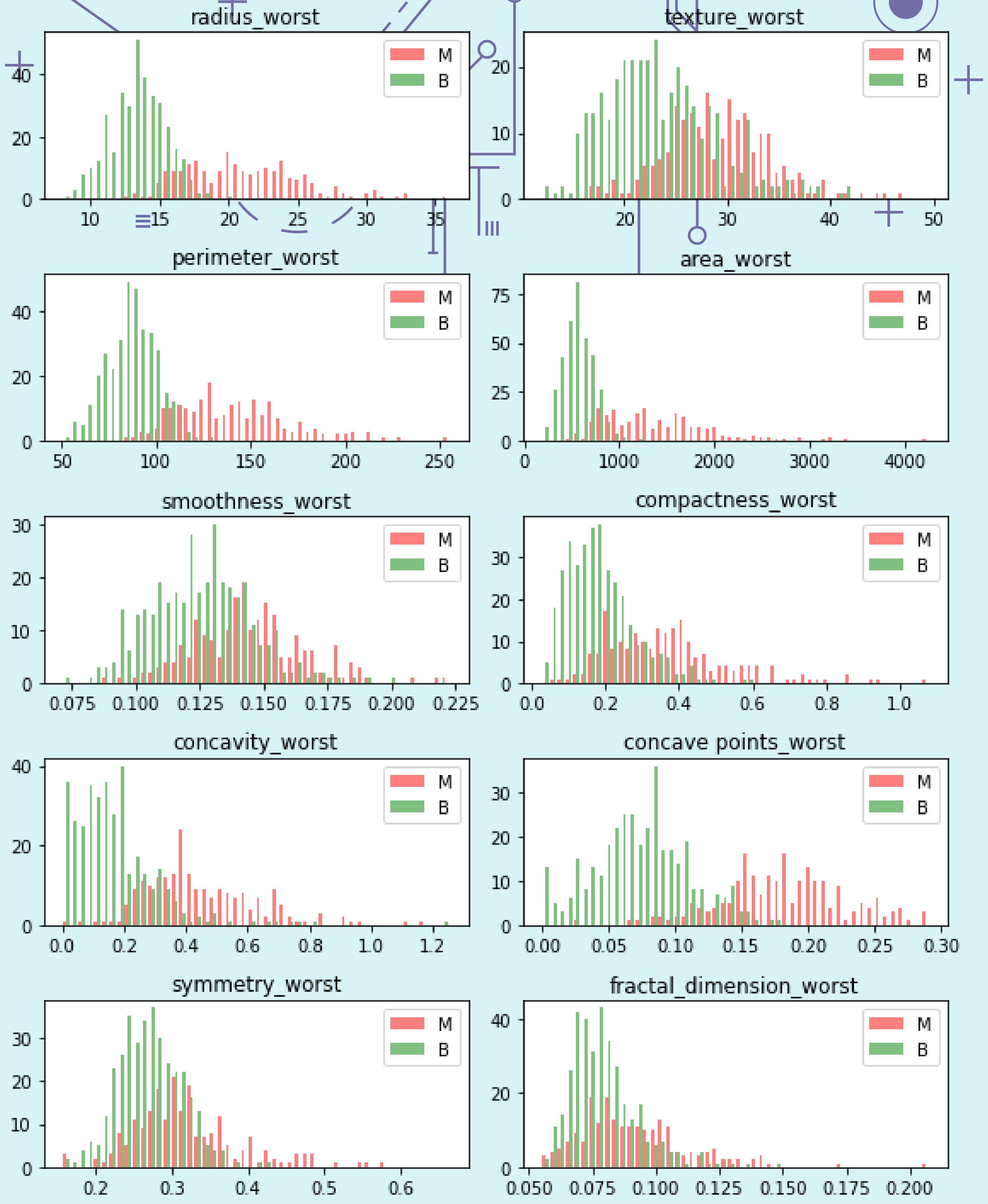
EXPLORATORY DATA ANALYSIS



STANDARD ERROR OF ATTRIBUTES

* It can be observed that Larger the standard error values of radius" , concave points", "compactness, area and perimeter" shows correlation with Malignant tumors.
So these features can be used for classification.

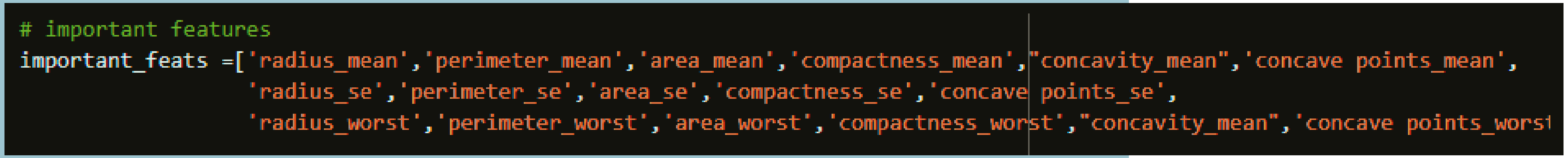
EXPLORATORY DATA ANALYSIS



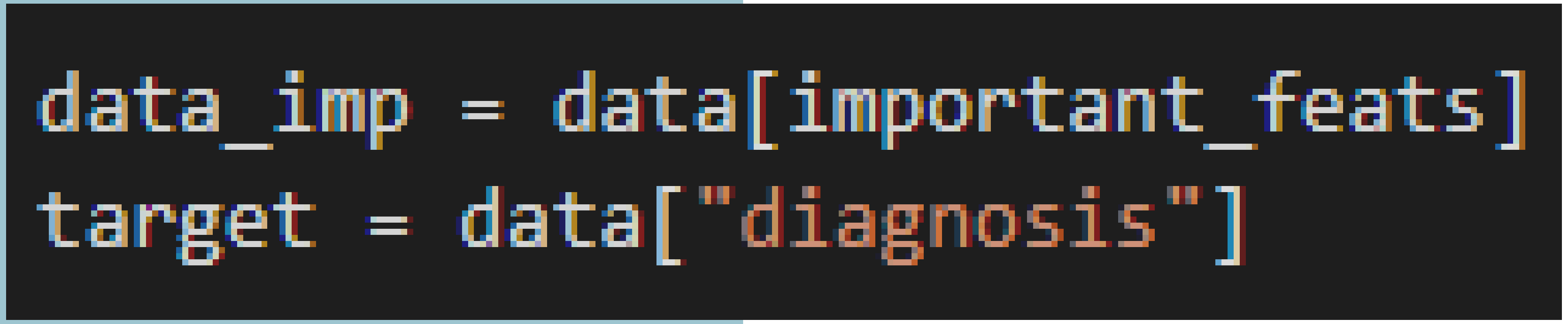
WORST/LARGEST VALUE OF ATTRIBUTES

* Similar to means values, It can be observed that Larger the worst values of radius" , concave points" , concavity" , "compactness, area and perimeter" shows correlation with Malignant tumors.So these features can be used for classification.

* On the other hand worst values of texture, smoothness, symmetry or fractual dimension does not show a particular preference of one diagnosis over the other



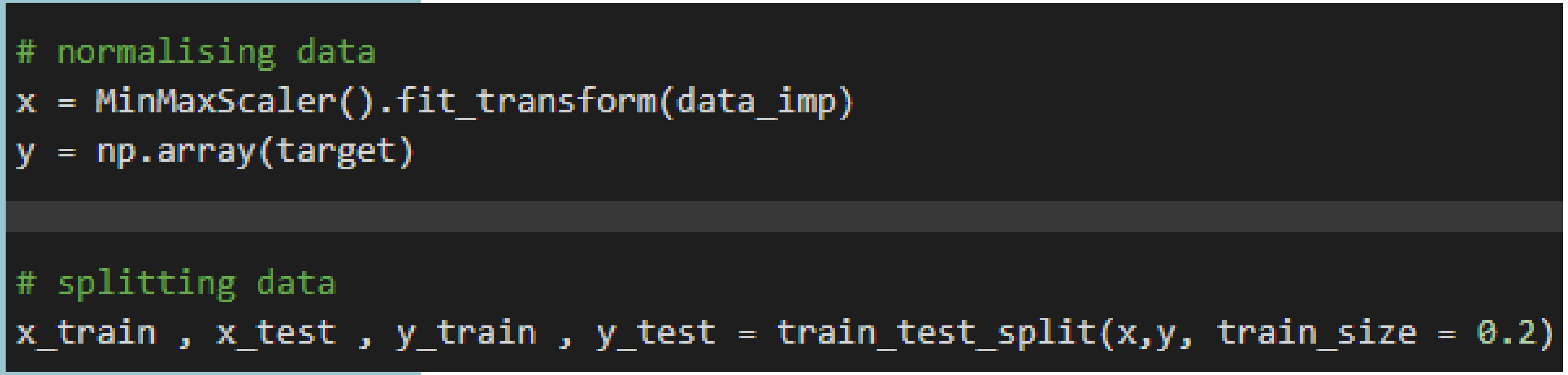
```python
# important features
important_feats =['radius_mean','perimeter_mean','area_mean','compactness_mean',"concavity_mean",'concave points_mean',
                 'radius_se','perimeter_se','area_se','compactness_se','concave points_se',
                 'radius_worst','perimeter_worst','area_worst','compactness_worst',"concavity_mean",'concave points_worst
```



```python
data_imp = data[important_feats]
target = data["diagnosis"]
```

Data Preprocessing

- New Dataset of Important features is created.
- Dependent and independent variables are seperated.
- Independent variables are **Normalized** using Min Max Scalar .
- Splitting the data into train and test data.



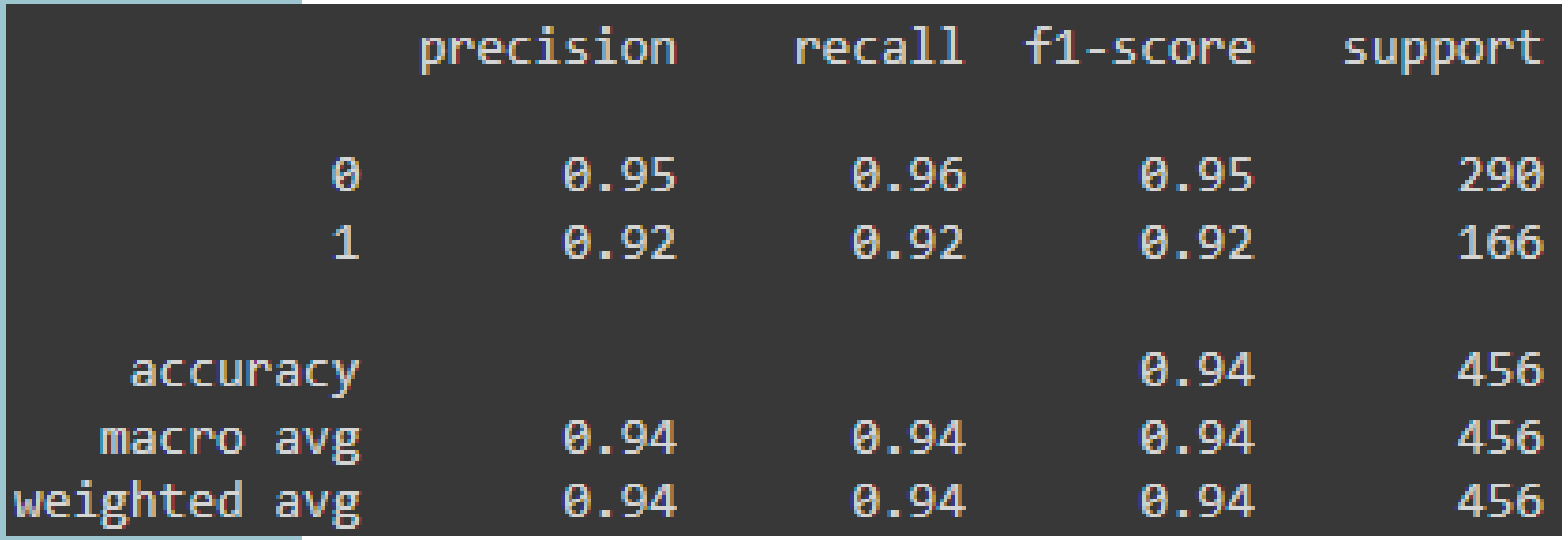
```python
# normalising data
x = MinMaxScaler().fit_transform(data_imp)
y = np.array(target)
```

```python
# splitting data
x_train , x_test , y_train , y_test = train_test_split(x,y, train_size = 0.2)
```

Predictions and ML Models

- I have trained data on 2 different machine learning models Random Forest classifier and SVM
- Out of 2 Random Forest Classifier outperformed with 94% Accuracy , 95 % precision and 96% recall.

Random Forest Classification Report

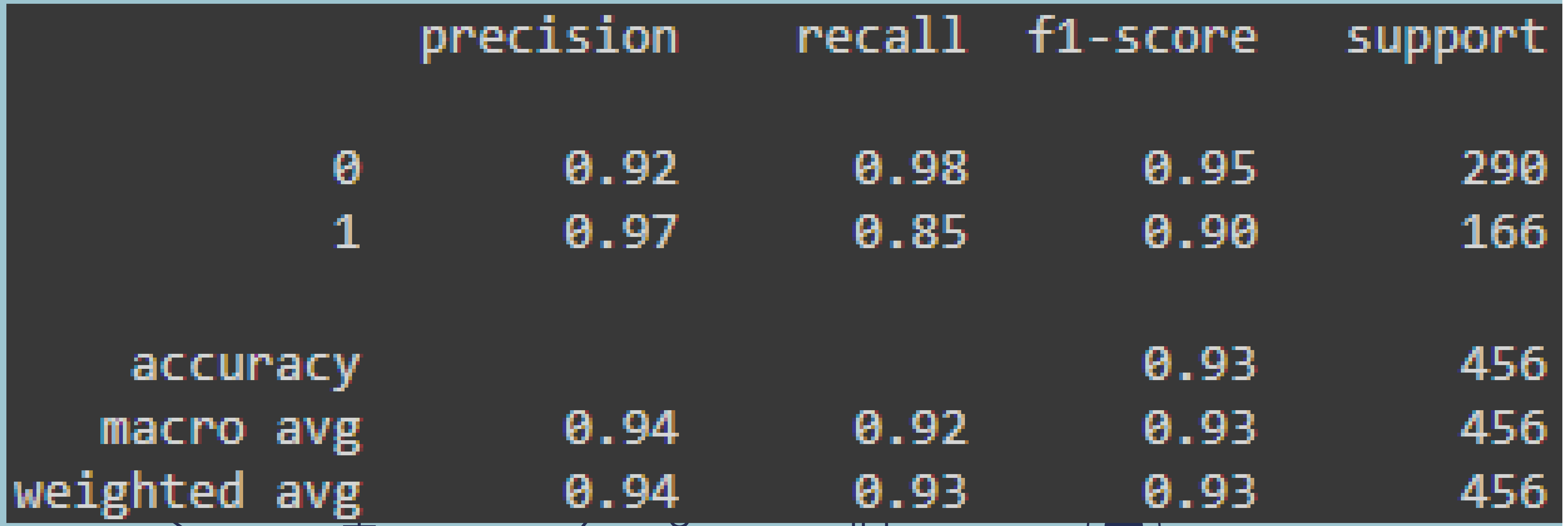


```
              precision    recall  f1-score   support

           0       0.95      0.96      0.95       290
           1       0.92      0.92      0.92       166

    accuracy                           0.94       456
   macro avg       0.94      0.94      0.94       456
weighted avg       0.94      0.94      0.94       456
```

SVM Classification Report



```
              precision    recall  f1-score   support

           0       0.92      0.98      0.95       290
           1       0.97      0.85      0.90       166

    accuracy                           0.93       456
   macro avg       0.94      0.92      0.93       456
weighted avg       0.94      0.93      0.93       456
```