PROJECT 2: BREAST CANCER PREDICTION

Data Set: Breast Cancer Wisconsin (Diagnostic) dataset

1. Data Preprocessing:

- **Handling Missing Values:** The dataset contained a column named 'Unnamed: 32' with 569 null values. This column was dropped from the dataset.

- **Outlier Detection and Handling**: Outliers were detected using the Z-score method, and rows containing outliers were removed from the dataset.

- **Normalisation:** The remaining numeric features were normalised using StandardScaler to ensure that each feature contributes equally to the analysis.

- **Encoding Labels:** The target variable 'diagnosis' was encoded into numerical labels using LabelEncoder.

2. Feature Selection:

- **Feature Selection Method:** The SelectKBest method with ANOVA F-test was employed to select the top 10 most relevant features for predicting breast cancer.

- **Selected Features:** The following features were selected: 'radius_mean', 'perimeter_mean', 'area_mean', 'concavity_mean', 'concave points_mean', 'area_se', 'radius_worst', 'perimeter_worst', 'area_worst', and 'concave points_worst'.

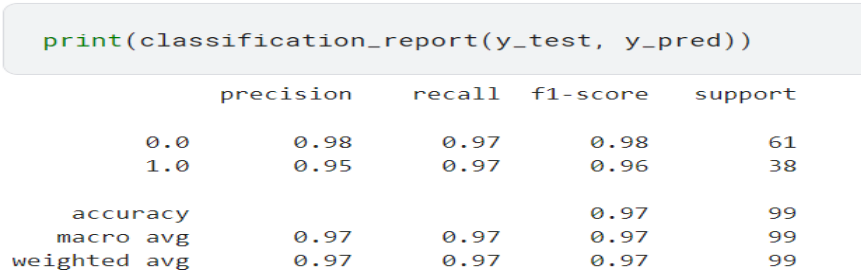
3. Machine Learning Model Implementation:

- **Model:** Support Vector Machine (SVM) with a linear kernel was chosen for breast cancer prediction.

- **Training and Evaluation:** The SVM model was trained on the training set and evaluated on the testing set using metrics such as accuracy, precision, recall, and F1-score.

- **Performance Metrics:** The performance metrics of the model are as follows:
  - Precision: Precision measures the proportion of true positive predictions among all positive predictions. For class 0 (benign), precision is 0.98, and for class 1 (malignant), precision is 0.95.
  - Recall: Recall measures the proportion of true positive predictions among all actual positive instances. For class 0, recall is 0.97, and for class 1, recall is 0.97.

- F1-score: F1-score is the harmonic mean of precision and recall, giving a balance between them. For class 0, F1-score is 0.98, and for class 1, F1-score is 0.96.
- Accuracy: Overall accuracy of the model is 0.97.

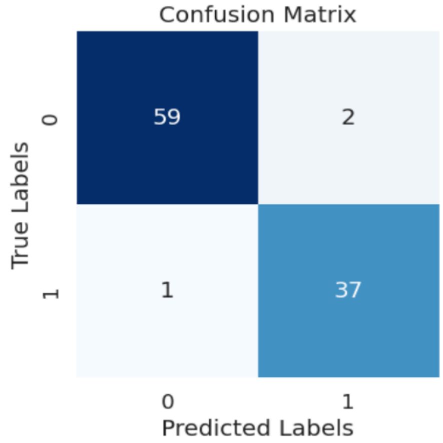


```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

         0.0       0.98      0.97      0.98        61
         1.0       0.95      0.97      0.96        38

    accuracy                           0.97        99
   macro avg       0.97      0.97      0.97        99
weighted avg       0.97      0.97      0.97        99
```



4. Challenges Faced:

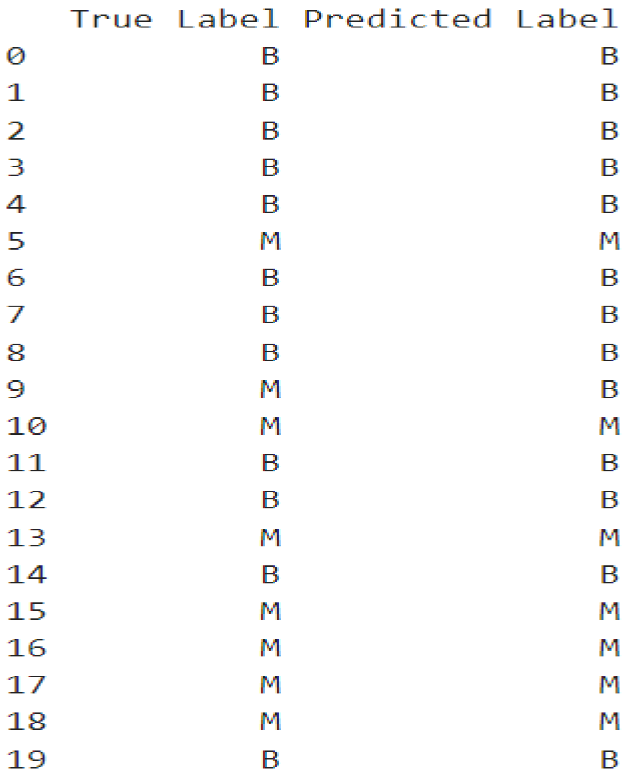
- **Handling Missing Values:** The presence of missing values in the 'Unnamed: 32' column required careful handling to avoid biases in the analysis.

- **Outlier Detection:** Identifying and handling outliers effectively without losing important information was challenging. The Z-score method was used, but other methods such as IQR (Interquartile Range) could also be considered.

- **Feature Selection:** Selecting the most relevant features from a large set of features can be challenging. The ANOVA F-test method was chosen, but other methods such as Recursive Feature Elimination (RFE) could also be explored.

5. Conclusion

Overall, the SVM model achieved high performance in predicting breast cancer, with an **accuracy of 97%**. The model demonstrated good precision, recall, and F1-score for both benign and malignant classes, indicating its effectiveness in classification tasks.



```
   True Label Predicted Label
0           B               B
1           B               B
2           B               B
3           B               B
4           B               B
5           M               M
6           B               B
7           B               B
8           B               B
9           M               B
10          M               M
11          B               B
12          B               B
13          M               M
14          B               B
15          M               M
16          M               M
17          M               M
18          M               M
19          B               B
```