**INTERN PROJECT PHASE – 1**

**Data Science Projects: Predictive Modeling**

**Project 2: Breast Cancer Prediction**

<https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data>

**Introduction:**

The objective of this project is to predict whether a tumor is malignant or benign based on the Breast Cancer Wisconsin (Diagnostic) dataset. This involves several steps including data preprocessing, feature selection, and implementing a Support Vector Machine (SVM) model.

**Data Preprocessing:**

* The dataset was loaded using pandas and checked for missing values.
* Missing values were handled using the mean strategy to ensure no data was lost.
* Columns that were not useful for model training, such as 'id' and 'Unnamed: 32', were removed.
* The target variable 'diagnosis' was encoded to numerical values, with 'M' mapped to 1 and 'B' mapped to 0. The data was split into features (X) and target (y).
* Outliers were detected using the z-score method and removed if they exceeded a threshold of 3. This was to ensure that the data used for training was clean and did not contain extreme values that could skew the model's performance.
* Features were then normalized using the Standard Scaler to standardize the data and make the model training more efficient. The dataset was split into training and testing sets with a ratio of 80:20.

**Feature Selection and Engineering:**

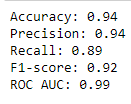
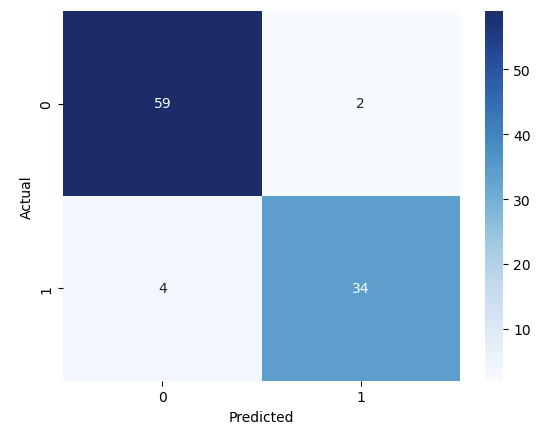
* A correlation matrix was used to identify important features. Features with a high correlation to the target variable were selected based on a threshold of 0.8.
* This step ensured that only the most relevant features were used for model training, improving the efficiency and performance of the model.
* The selected features were then normalized, and the data was split again into training and testing sets.

**Machine Learning Model (SVM):**

* A Support Vector Machine (SVM) model with a linear kernel was implemented. The model was trained on the selected and scaled features.
* Predictions were made on the test set, and various performance metrics were calculated to evaluate the model's performance. These metrics included accuracy, precision, recall, F1-score, and ROC AUC.
* The accuracy measures the proportion of correct predictions, while precision indicates the accuracy of the positive predictions.
* Recall measures the model's ability to detect positive samples, and the F1-score provides a balance between precision and recall.
* The ROC AUC indicates the model's ability to discriminate between positive and negative classes.

**Results and Evaluation:**

* The SVM model achieved a satisfactory performance based on the calculated metrics.
* The confusion matrix was plotted to visualize the number of true positives, true negatives, false positives, and false negatives.

****

**Challenges:**

Several challenges were faced during the project. Handling missing values and outliers effectively without losing significant data was crucial. Ensuring that the selected features contributed positively to the model's performance was also a key challenge.

In conclusion, this project successfully implemented a Support Vector Machine model to predict whether a tumor is malignant or benign. The steps taken in data preprocessing, feature selection, and model training were documented, and the model's performance was evaluated using various metrics.

Source Code: <https://github.com/msubhikshaT/Data-Science-Project-Series/blob/main/2%20-%20Breast%20Cancer%20Prediction.ipynb>