Breast Cancer Prediction Project

Introduction

This project aims to build a machine-learning model to classify breast cancer tumours as either malignant or benign using the Breast Cancer Wisconsin (Diagnostic) dataset. We utilized a Support Vector Machine (SVM) for this task due to its effectiveness in classification problems.

Dataset

The dataset contains 569 entries with 33 columns. These columns include:

- Patient ID (which was dropped in preprocessing)
- Diagnosis of the tumour (either malignant or benign)
- 30 numerical features describing various characteristics of the cell nuclei in the breast tissue.

Data Preprocessing

1. **Loading Data**: The dataset was loaded into a Pandas DataFrame for ease of manipulation.
2. **Removing Irrelevant Columns**: Columns that did not contribute to the predictive modelling, such as the patient ID and an unnamed column with missing values, were removed.
3. **Encoding Categorical Variables**: The target variable, 'diagnosis', was converted into numerical values (1 for malignant and 0 for benign) to be used in the machine learning model.
4. **Handling Outliers**: Outliers were identified and removed using the Interquartile Range (IQR) method to ensure the robustness of the model.
5. **Normalizing Data**: The feature values were standardized to have a mean of zero and a standard deviation of one to facilitate effective training of the SVM.

Feature Selection

Correlation Matrix

A correlation matrix was computed to understand the relationships between different features. This analysis helps in identifying the most relevant features for the prediction task by showing how features correlate with each other and with the target variable.

Machine Learning Model

Model Implementation

An SVM with a linear kernel was chosen for this classification task. The data was split into training and testing sets, with 80% of the data used for training and 20% for testing. The model was trained on the training set and then used to make predictions on the test set.

Model Evaluation

The performance of the SVM model was evaluated using several metrics:

- **Accuracy**: The overall accuracy of the model was 93.75%, indicating that the model correctly classified 93.75% of the test instances.
- **Confusion Matrix**: The confusion matrix showed the number of true positives, true negatives, false positives, and false negatives, helping to understand the types of errors the model made.
- **Classification Report**: This report provided precision, recall, and F1 scores for both classes (malignant and benign). The high precision and recall values indicated that the model performed well in distinguishing between the two classes.

Challenges and Solutions

Challenges

- **Outliers**: Outliers in the data could significantly affect the performance of the model.
- **Feature Selection**: Identifying the most relevant features was crucial to improve the model's accuracy and interpretability.

Solutions

- **IQR Method**: This method was used to effectively identify and remove outliers, ensuring a more reliable model.
- **Correlation Analysis**: This analysis helped in selecting features that were most relevant to the prediction task, improving the model's performance.

Conclusion

The SVM model achieved high accuracy and demonstrated strong performance in classifying breast cancer tumours as malignant or benign. This project underscores the importance of thorough data preprocessing and careful feature selection in building effective machine-learning models.