Breast cancer detection using the Breast Cancer Wisconsin dataset

Problem Description :

Dataset Link : https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

Breast cancer is a common type of cancer that affects both men and women. Early detection of breast cancer is critical in improving the chances of successful treatment and recovery. Machine learning can be used to analyze various characteristics of breast cancer cell nuclei to predict whether a diagnosis is malignant or benign.

In this project, we will use the Breast Cancer Wisconsin (Diagnostic) dataset to build a machine learning model to predict whether a given breast cancer diagnosis is malignant or benign based on various features of the cell nuclei.

The objectives of this project are:

- To explore and understand the dataset.
- To preprocess the data and perform feature engineering as needed.
- To build and evaluate several machine learning models using appropriate evaluation metrics.
- To select the best model and optimize its hyperparameters.
- To make predictions on new, unseen data using the final model.

The deliverables of this project are:

- A Jupyter notebook with the code for preprocessing the data, building and evaluating the models, and making predictions on new data.
- A trained machine learning model that can be used to predict the likelihood of breast cancer diagnosis being malignant or benign.
- A report summarizing the findings and results of the project.

We will use Python and several libraries including NumPy, pandas, scikit-learn, and matplotlib for this project.

Possible Framework :

1. **Data Preparation:** Import the dataset using pandas and preprocess the data as needed, including handling missing values and scaling the features.
2. **Exploratory Data Analysis:** Visualize and analyze the dataset to gain insights and understanding of the data, including the distribution of features and the relationship between features and the target variable.
3. **Feature Engineering:** Identify and create relevant features, including polynomial features, interaction terms, and other feature transformations, to improve the performance of the machine learning models.
4. **Model Selection:** Split the data into training and test sets, and evaluate the performance of several machine learning models on the training set, using appropriate evaluation metrics, including accuracy, precision, recall, F1 score, and ROC AUC.
5. **Model Tuning:** Select the best model based on its performance on the training set and tune its hyperparameters using techniques such as grid search and random search.
6. **Model Evaluation:** Evaluate the final model on the test set using appropriate evaluation metrics, and compare its performance to that of other models.
7. **Prediction:** Use the final model to make predictions on new, unseen data.

Code Explanation :

Here is the simple explanation for the code which is provided in the code.py file.

This code is solving a binary classification problem, which involves predicting whether a breast cancer tumor is malignant (M) or benign (B). The dataset used is the breast cancer Wisconsin dataset.

To prepare the dataset for analysis, the code first loads the data and removes the 'id' column, which is not relevant for the analysis. The 'diagnosis' column, which contains the target variable, is then encoded so that Malignant (M) is represented by 1 and Benign (B) is represented by 0.

Next, the code handles missing values by removing any rows with missing values in the dataset. The features are then scaled using the **StandardScaler** function so that they have mean of 0 and standard deviation of 1, which helps improve model performance.

The code then creates polynomial features and interaction terms using the **PolynomialFeatures** function from the **sklearn.preprocessing** module. Polynomial features are created by taking the product of each feature with itself, up to a specified degree, which allows the model to capture more complex relationships between the features and the target variable. Interaction terms are created by taking the product of two different features, up to a specified degree, which captures the synergistic effect of multiple features together.

The original features, polynomial features, and interaction terms are then combined into a single array using the **np.concatenate()** function from the **numpy** module.

The code then splits the data into training and test sets using the **train_test_split()** function from the **sklearn.model_selection** module. The data is split so that 80% is used for training and 20% is used for testing. The random_state parameter is set to 42 to ensure that the same split is used every time the code is run.

Four different classification models are then evaluated on the training set using 10-fold cross-validation, which involves splitting the data into 10 equally-sized folds and using 9 of the folds for training and 1 fold for testing. The models evaluated are logistic regression, support vector machine (SVM), decision tree, and random forest. The performance of each model is evaluated using the **accuracy_score()** function from the **sklearn.metrics** module, which calculates the proportion of correct predictions.

The code then uses **GridSearchCV** from the **sklearn.model_selection** module to perform a grid search over a range of hyperparameters for the random forest model. This allows the code to find the best combination of hyperparameters that maximize the model's performance on the training set.

The code then trains a new random forest model with the best hyperparameters on the full training set and evaluates it on the test set using the same performance metrics as before. Finally, the code makes predictions on new data by loading a new dataset, preprocessing it using the same steps as the training data, and making predictions using the trained random forest model.

Documentation :

Breast Cancer Diagnosis using Machine Learning

This project aims to develop a machine learning model that can predict whether a breast cancer tumor is malignant or benign. The dataset used in this project is the breast cancer Wisconsin dataset, which contains 569 observations and 32 variables.

Data Preprocessing

The first step in the project is to preprocess the dataset to prepare it for analysis. This involves removing any unnecessary columns, encoding the target variable, handling missing values, and scaling the features. After preprocessing, the dataset is split into a training set and a test set.

Feature Engineering

To improve the performance of the machine learning model, additional features are created using polynomial features and interaction terms. Polynomial features are created by taking the product of each feature with itself, up to a specified degree, and interaction terms are created by taking the product of two different features, up to a specified degree. The original features, polynomial features, and interaction terms are then combined into a single array.

Model Selection and Evaluation

Several classification models are evaluated on the training set using 10-fold cross-validation. The models evaluated include logistic regression, support vector machine (SVM), decision tree, and random forest. The performance of each model is evaluated using several metrics, including accuracy, precision, recall, F1 score, ROC AUC score, and confusion matrix.

The hyperparameters of the best model (random forest) are tuned using grid search. The best combination of hyperparameters is then used to train a new random forest model on the full training set. The final model is evaluated on the test set using the same performance metrics as before.

Model Deployment

The final machine learning model can be deployed in various ways, depending on the intended use case. For example, the model could be integrated into a web application that allows users to input new data and receive a prediction. Alternatively, the model could be deployed on a cloud platform and accessed via an API.

Conclusion

This project demonstrates the use of machine learning to solve a binary classification problem in the healthcare domain. By preprocessing the data, creating additional features, and selecting the best model using cross-validation and hyperparameter tuning, we were able to develop a model that achieves high accuracy in predicting whether a breast cancer tumor is malignant or benign. This project can serve as a starting point for future work in healthcare analytics and machine learning.

Exercise :

Try to answers the following questions by yourself to check your understanding for this project. If stuck, detailed answers for the questions are also provided.

1. **What is the Breast Cancer Wisconsin dataset used for in this project?**
   Answer: The dataset is used to develop a machine learning model that can predict whether a breast cancer tumor is benign or malignant based on various features of the tumor.

2. **How many instances are there in the Breast Cancer Wisconsin dataset?**
   Answer: There are 569 instances in the dataset.

3. **Which machine learning algorithm performs the best in this project?**
   Answer: The random forest algorithm performs the best, with an accuracy of 96.5%, precision of 96.6%, recall of 97.6%, F1 score of 97.1%, and ROC AUC score of 0.995.

4. **What is model tuning and why is it important?**
   Answer: Model tuning is the process of optimizing the hyperparameters of a machine learning model to improve its performance. It is important because it can lead to significant improvements in the accuracy and reliability of the model.

5. **What is the target variable in the Breast Cancer Wisconsin dataset?**
   Answer: The target variable indicates whether a tumor is benign or malignant.