Glioma Grading: Clinical and Mutation Features

Gliomas are the most common primary tumors of the brain.

About Dataset

Gliomas are the most common primary tumors of the brain. They can be graded as LGG (Lower-Grade Glioma) or GBM (Glioblastoma Multiforme) depending on the histological/imaging criteria. Clinical and molecular/mutation factors are also very crucial for the grading process. Molecular tests are expensive to help accurately diagnose glioma patients.

In this dataset, the most frequently mutated 20 genes and 3 clinical features are considered from TCGA-LGG and TCGA-GBM brain glioma projects.

The prediction task is to determine whether a patient is LGG or GBM with a given clinical and molecular/mutation features. The main objective is to find the optimal subset of mutation genes and clinical features for the glioma grading process to improve performance and reduce costs.

Source Dataset: UC Irvine Machine Learning Repository | Glioma Grading Clinical and Mutation Features

The Cancer Genome Atlas (TCGA) Project – NCI funded the creation of this dataset.

Task

The task is to design a classifier to predict Glioma grade using the provided dataset, with input features such as Gender, Age_at_diagnosis, Race, and genetic markers like IDH1, TP53, etc.

Consider the following points in your design:

Data Exploration and Preprocessing:

- Explore the dataset to understand its structure, distributions, and missing values.

- Preprocess the data by handling missing values, encoding categorical variables, and scaling numerical features if necessary.

Data Splitting:

- Split the dataset into training and testing sets with an 80%-20% ratio.

Cross-validation:

- Implement 10-fold cross-validation on the training dataset to ensure robust model evaluation.

Model Selection and Hyperparameter Tuning:

- Evaluate the performance of various classification algorithms: logistic regression, decision tree, random forest, and support vector machine (SVM).

- Use GridSearchCV to tune the hyperparameters of these models for improved performance.

Pipeline Construction:

- Construct pipelines to integrate preprocessing steps (e.g., data scaling, encoding) with the classification algorithms.

Model Evaluation:

- Assess the performance of trained models using metrics such as Accuracy, Specificity, Sensitivity, and F1 score on the test set.

- Plot precision-recall curves and ROC curves for all trained classifiers to visualize their performance.

Feature Importance Analysis:

- Analyze feature importance from the trained decision tree and random forest models to understand which features contribute most to Glioma grade prediction.

Report Generation:

- reprot jupyter notebook file. Save it as HTML and submit it.

- report csv prediction for test.xlsx. This csv file should contains two columns namely the sample_id and your best model prediction