NAME: ANAMITRA SENGUPTA
ROLL: 20IE10004
DATE: 16.09.2023
TITLE: *SURVIVAL PREDICTION ON BRATS 2020 DATASET USING EXTENSIVE FEATURE EXTRACTION, FEATURE PRUNING AND RANDOM FOREST CLASSIFIER*

Introduction: Survival Prediction was performed on patients with brain tumors, such as glioma. A radiomics model was used, wherein standard features were extracted from the MRIs and their transformed images. Due to the vast quantity of extracted features, pruning was essential to save time complexity and eliminate irrelevant features. Post-pruning, a random forest classifier was used to classify patients, based on the number of days: *d*, into three categories: (a) Short survival (*d<250*), (b) Medium Survival (*250<=d<=450*), (c) Long Survival (*d>450*).

Procedure & Results:

(A) Feature Extraction

(1) The dataset consisted of 4 image modalities per patient: *native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-flair).*

(2) The following image transformations were applied to each of the modalities, using the Pyradiomics 'imageoperations' module:
   (a) Laplacian of Gaussian filter (for σ = [5:0:-0.5]) [*10*]
   (b) Wavelet Filter (LLL,LLH,LHL,HLL,LHH,HLH,HHL,HHH) [*8*]
   (c) Square of image intensities [*1*]
   (d) Square-root of image intensities [*1*]
   (e) Logarithm of (the absolute value of original image + 1) [*1*]
   (f) Exponential of the original image [*1*]
   (g) Gradient Magnitude in the image [*1*]
   (h) Local Binary Pattern (LBP) in 2D (image processing in a by-slice operation) [*1*]
   (i) Local Binary Pattern (LBP) in 3D (using spherical harmonics) [*1*]

   The number beside each transformation indicates the number of images obtained. We have obtained a total of 26 images (including the original), per modality.

(3) The standard Pyradiomics features were extracted from each of the transformed and original images. These include:
   (a) First Order Statistics (19 features)
   (b) Shape-based (3D) (16 features)
   (c) Shape-based (2D) (10 features)
   (d) Gray Level Co-occurrence Matrix (24 features)
   (e) Gray Level Run Length Matrix (16 features)
   (f) Gray Level Size Zone Matrix (16 features)
   (g) Neighbouring Gray Tone Difference Matrix (5 features)
   (h) Gray Level Dependence Matrix (14 features)

   A total of 120 features, were extracted for each image.

Thus, in total, we have obtained 4*26*120 = 12,480 features per patient.

(B) Feature Elimination

(1) For the given problem of survival prediction, age of the patient was added as a feature.
(2) All the features were scaled using MinMaxScaler.
(3) The dataset consisted of 117 entries. The train-test ratio was 80:20.
(4) Fisher scores were calculated from the training dataset. The top 2000 features were extracted from the sorted list of feature-score pairs.
(5) Recursive Feature Elimination was performed with the Random Forest Classifier model (n_estimators = 100). The number of features were pruned to 400.

(C) Class Prediction

(1) Random Forest Classifier was used as the prediction model.
(2) For hyperparameter tuning, RandomizedSearchCV was used.The following hyperparameter values were obtained as a result:

```
{'n_estimators': 200,
 'min_samples_split': 2,
 'min_samples_leaf': 1,
 'max_depth': 80,
 'bootstrap': True}
```

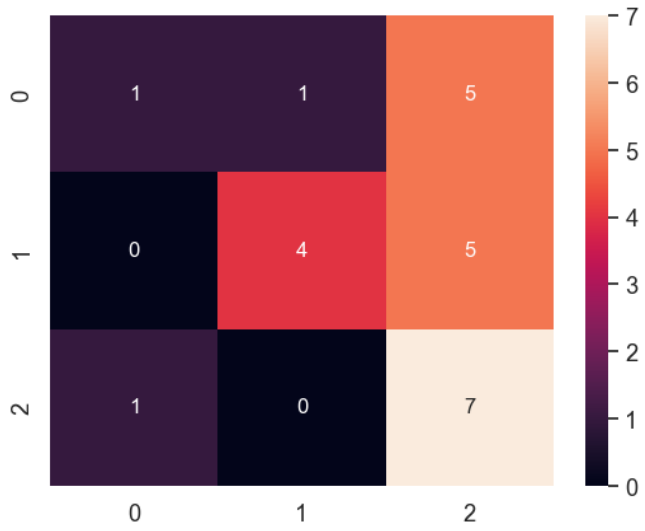
(3) For finer tuning, GridSearchCV was used. The parameter list was taken to be around the values obtained using RandomizedSearchCV.

```
{'bootstrap': True,
 'max_depth': 100,
 'min_samples_leaf': 3,
 'min_samples_split': 3,
 'n_estimators': 200}
```

(4) Random Forest Classifier with these parameters was then used for the class-prediction. The performance report is as follows:

	precision	recall	f1-score	support
0	0.50	0.14	0.22	7
1	0.80	0.44	0.57	9
2	0.41	0.88	0.56	8
accuracy			0.50	24
macro avg	0.57	0.49	0.45	24
weighted avg	0.58	0.50	0.47	24

(5) The confusion matrix is as follows:



Reference: *"Glioma Segmentation Using Ensemble of 2D/3D U-Nets and Survival Prediction Using Multiple Features Fusion" - Muhammad Junaid Ali, Muhammad Tahir Akram, Hira Saleem, Basit Raza, and Ahmad Raza Shahid*
The solution was a modified implementation of the above mentioned research paper.
The percentage accuracy in the original paper was 48.3% for regression. The modified implementation has an accuracy of 50%. However, it is a classification (by dividing the predicted survival days into three classes).