Brain tumor prediction model - HarvardX Data science Capstone assignment #2

edX learner DBel_17

April 22nd, 2023

INTRODUCTION

Glioma is the most frequent brain tumor and consists of cells that most closely resemble those brain cells which normally fulfill supportive and structural tasks, i.e. glial and astrocyte cells, as opposed to the signal transmitting neurons [1]. Clinical diagnosis and prognosis of glioma depend on magnetic resonance tomography (MRI) imaging and evaluation of the scans by experienced radiologists. In the past decade, machine-learning methods have been devised to identify and demarcate different regions of brain tumors, which is referred to as brain tumor segmentation. The goal of these endeavors is to facilitate and objectify tumor diagnosis. Precise diagnosis is at the core of appropriate patient care and prognostic projections [2].

In the present project, a pre-processed dataset was retrieved from the Kaggle repository [3], containing a series of MRI scan images from patients diagnosed with glioma. The scans show either completely normal or tumor-infiltrated areas of tissue. Each image was assigned a ground truth allowing for evaluation of prediction algorithms. Although the purpose of the original dataset was to classify and segment the MRI sections into areas of either normal tissue or four different types of tumorous tissue [4], the creator of the Kaggle dataset had simplified the task. The outcome measure now only differentiates between images strictly positive or negative for the appearance of a tumor. Thus, it becomes a binary classification challenge.

Furthermore, the texture features of the images were extracted using mathematical methods using the gray-level co-occurrence matrix (GLCM) [6] resulting in first-order and second-order features. First-order features, namely variance (or standard deviation), kurtosis, skewness and mean, provide a measure of how grey pixels are overall distributed across the image, while second-order features such as angular second moment (ASM), entropy, contrast or dissimilarity quantify the relationship between pixels within the image and can be used to extrapolate to the coarseness or smoothness of an image.

METHODS

Data partitioning

To independently validate our model, we partitioned the initial dataset containing 3762 images into a development and a final holdout dataset, using a 90/10 partition, leaving enough data for training. The development data were further split into a training and test set, again using a 90/10 partition. Models were then trained and performance assessed. Finally, the models were trained on the complete development set to feed as much data as possible and the prediction algorithm evaluated on the final holdout set.

Assessing model performance

Since we had a discrete binary outcome, we used Accuracy **A** as a measure of model performance which is defined as

$\mathbf{A} = \frac{TP+TN}{TP+TN+FP+FN}$, with T being "true", F being "false", P being "positive" and N "being negative".

Any model performance has to be measured against a random prediction which in the case of a binary outcome would have an accuracy of 50%.

Feature selection

There were 13 predictors present in the Kaggle dataset: **Mean, Variance, Standard.Deviation, Entropy, Skewness, Kurtosis, Contrast, Energy, ASM, Homogeneity, Dissimilarity, Correlation, Coarseness**. Predictors with low variance were identified with nearZeroVar() from the *caret* package and thus **Coarseness** was removed.

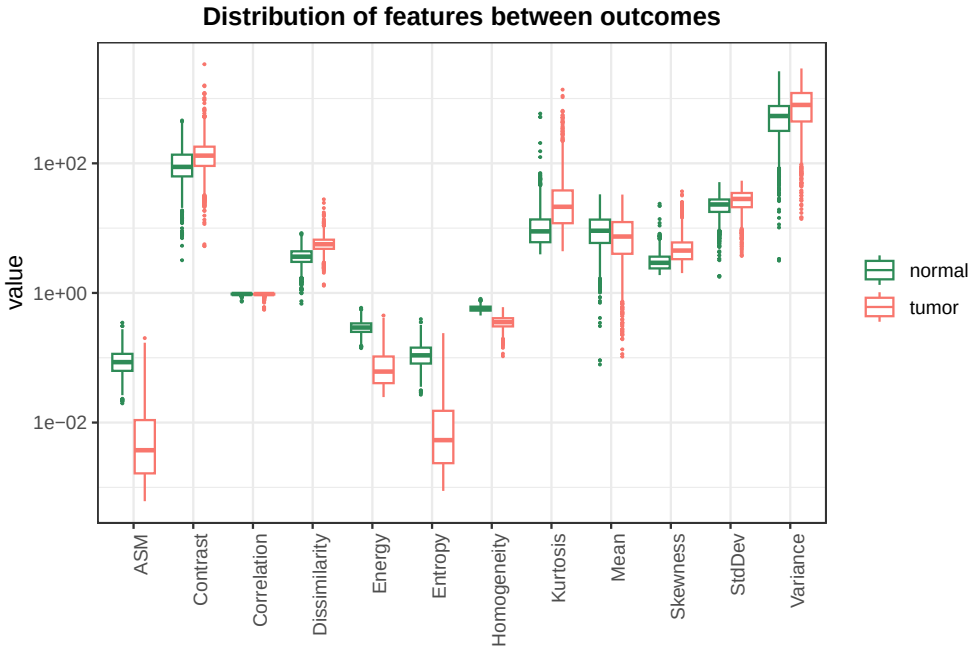

Figure 1: Outcome discrimination by features

As can be seen from Fig. X, some features have more discriminative power than others and some like ASM and Entropy display almost the same distribution. To exclude redundant features and therefore to decrease the possibility of overfitting as well as to save computational resources, a filter method was applied examining the correlation and relationships between predictors among themselves and between predictors and the outcome. To presume that correlation is given, the Pearson correlation coefficient had to be greater than 0.7. Deterministic relationships between features spoke in favor of retaining only the one with the highest correlation to the outcome, while stochastic relationships were weighed in favor of retaining a feature in a correlation cluster. To not lose important information, the decision to exclude features was therefore conservative. A pairwise scatter plot grid (Fig. X) and a pairwise correlation matrix (Fig. X) guided the selection.

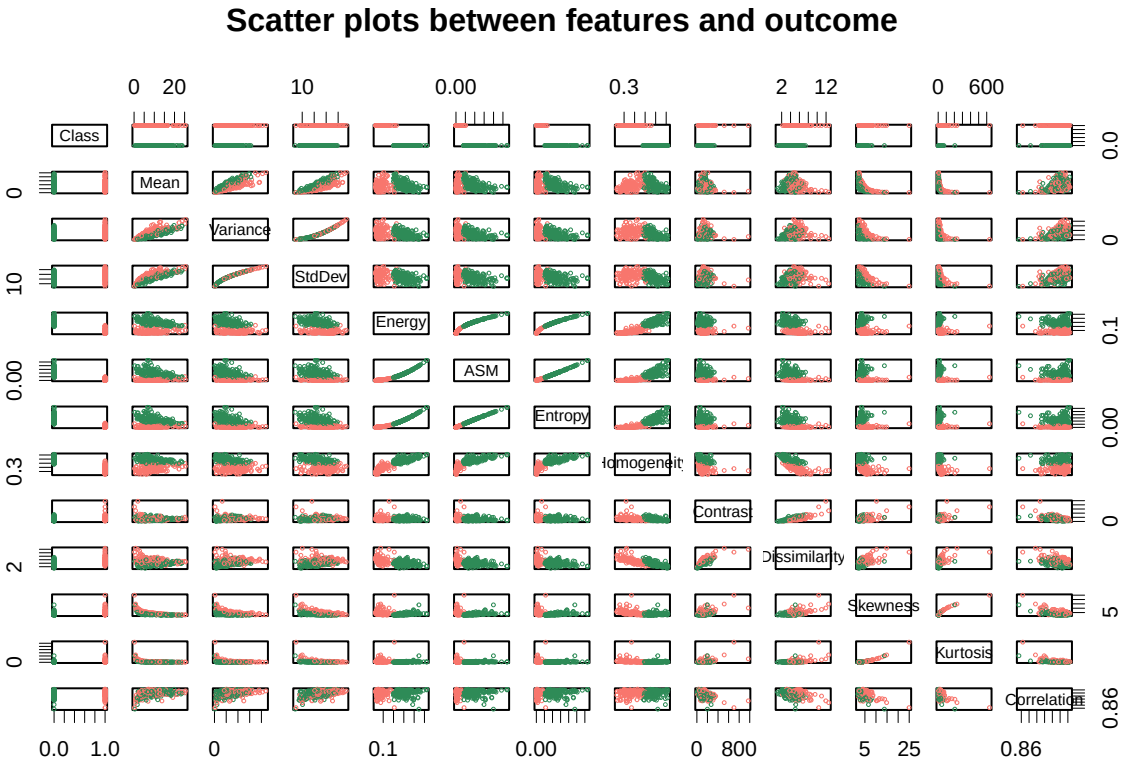

Figure 2: Scatter plots between features and outcome

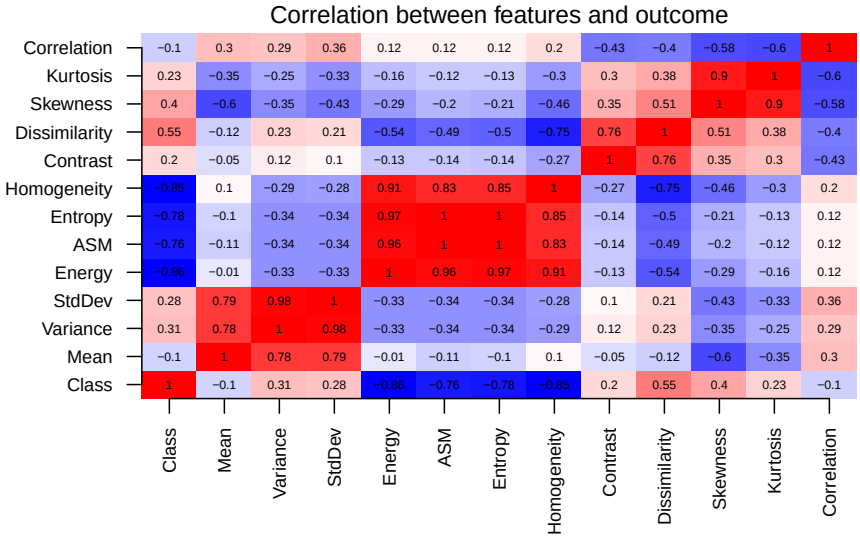

Figure 3: Correlation between features and outcome

Mean, Variance and Std.Dev. are highly correlated between each other. Variance has the highest positive correlation to the outcome ("Class"), but Mean is anti-correlated and shows a stochastic relationship to Variance. Therefore, Mean and Variance were kept.

Entropy, Energy, ASM and Homogeneity are highly correlated. Energy has the highest anti-correlation to class. Homogeneity is the only one to show a stochastic relationship to the other three features in the cluster. We therefore kept Energy and Homogeneity.

Dissimilarity and Contrast are highly correlated. Dissimilarity has higher correlation to Class and was retained.

Kurtosis and Skewness are highly correlated and have a deterministic relationship. Skewness has the stronger correlation to Class and was retained.

In summary, we kept the following 7 of the 12 non-zero-variance features for training: **Variance, Mean, Energy, Homogeneity, Skewness, Dissimilarity and Correlation**.

Scan visualization

MRI scans were read with the readJPEG() function of the *jpeg* package and visualized using the base *R* plot(as.raster()) function. An example of randomly selected images of normal and tumorous tissues is provided in Fig. X. As can be seen, all sections are in the transverse plane.

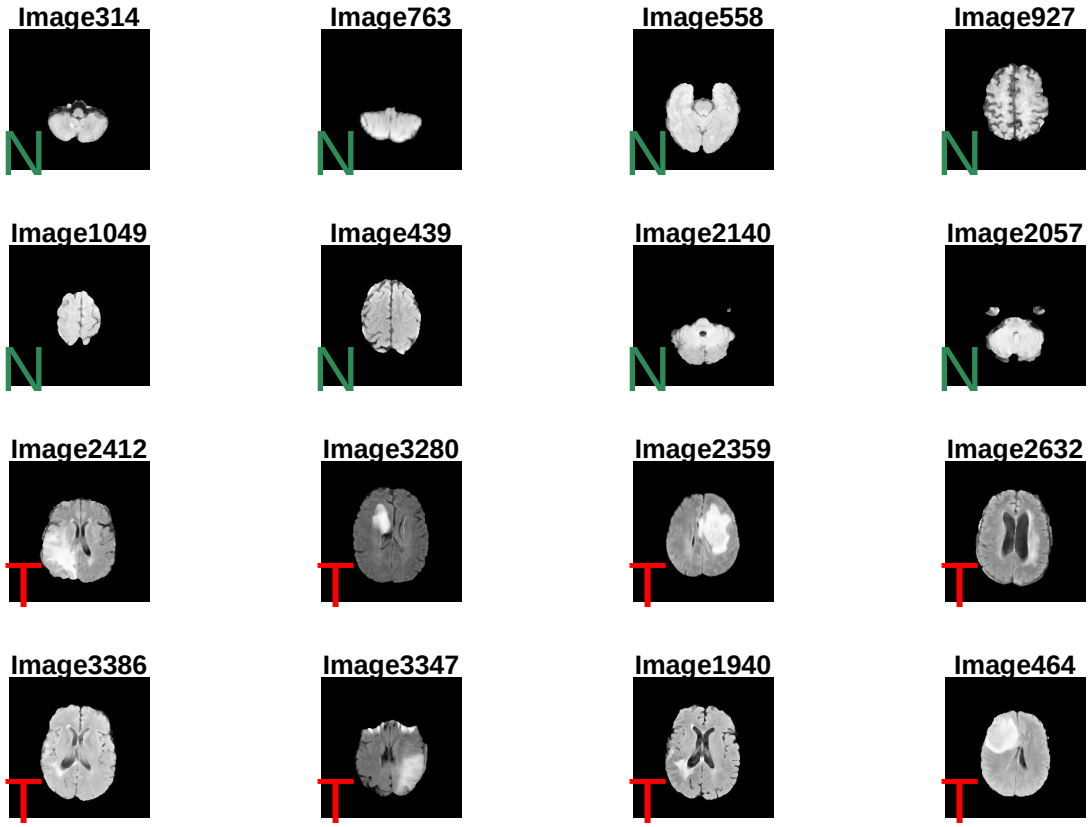

Figure 4: Examples of MRI images

1 Zong, H. *et al.* The cellular origin for malignant glioma and prospects for clinical advancements. *Expert Review of Molecular Diagnostics*, 2012, 12 (4), pp. 383–394. DOI: 10.1586/erm.12.30

2 Menze, B. *et al.* The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 2014, p. 33. DOI: 10.1109/TMI.2014.2377694

3 Bohaju, J. Brain tumor. 2020. DOI: 10.34740/KAGGLE/DSV/1370629

4 *BRATS 2015: Brain tumor image segmentation challenge.* 2015. Available from: https://www.smir.ch/BRATS/Start2015 [Accessed 22 April 2023]

5 Haralick, R.M. *et al.* Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 1973, SMC-3 (6), pp. 610–621. DOI: 10.1109/TSMC.1973.4309314

6 Aggarwal, N. *et al.* First and second order statistics features for classification of magnetic resonance brain images. *Journal of Signal and Information Processing*, 2012, 3 (2), pp. 146–153. DOI: 10.4236/jsip.2012.32019.