|  |
| --- |
| `Group 6 IST 718 |
| Image Characterization in Radiomics |
| *Tumor or Non-Tumor* |

|  |
| --- |
| Andrea Smith, Yu Sheng Lu  [Date] |

**Introduction**

*Radiomic Features Explanation*

With the rapidly growing field of Artificial Intelligence in modern life, radiology can extract vast amounts of information from images. By analyzing quantitative imaging biomarkers, researchers can better identify patterns and characteristics of tumor aggressiveness which can lead to improved personalized therapy. The need for non-invasive methods to identify these pathologies could help prevent patient downtime and infection rates.

*Data Source*

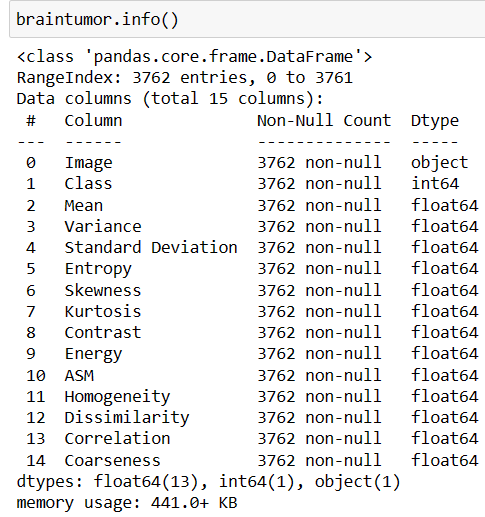
The data used to conduct this analysis will include 3762 brain images, first order and second order features collected from Kaggle. This dataset includes the following first order features: mean, variance, standard deviation, skewness, and kurtosis. And the following second order features: contrast, energy, angular second movement (ASM), entropy, homogeneity, dissimilarity, correlation, and coarseness. Radiomic features capture tissue and lesion characteristics that describe the shape, form, and texture of the brain images. “Because these features are based on single-pixel or single-voxel analyses, they are called first-order features.” (Journal of Nuclear Medicine)

*Problem We are Trying to Solve*

Radiomic feature extraction provides objective quantitative data that can be analyzed and used to understand brain tumor characteristics and patterns of progression which is often difficult to identify with the naked eye. These features, known as quantitative imaging biomarkers, provide a comprehensive method to identify regions of interest that can be analyzed for disease aggressiveness and thereby help provide for personalized medical care and treatment. With AI to target this research, we intend to use machine learning algorithms to further identify and classify phases in the disease and what characteristics represent those phases.

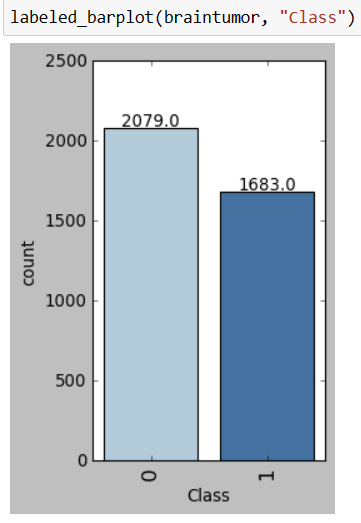
*Features*

|  |  |
| --- | --- |
| Data | Description |
| Images | Image name |
| Class | Target value Tumor = 1, Non tumor = 0 |
| Mean | First order feature mean |
| Variance | First order feature variance |
| Standard Deviation | First order feature std deviation |
| Skewness | First order feature skewness |
| Kurtosis | First order feature kurtosis |
| Contrast | Second order feature contrast |
| Energy | Second order feature energy |
| ASM | Second order feature ASM (Angular second moment) |
| Entropy | Second order feature entropy |
| Homogeneity | Second order feature homogeneity |
| Dissimilarity | Second order feature dissimilarity |
| Correlation | Second order feature correlation |
| Coarseness | Second order feature coarseness |



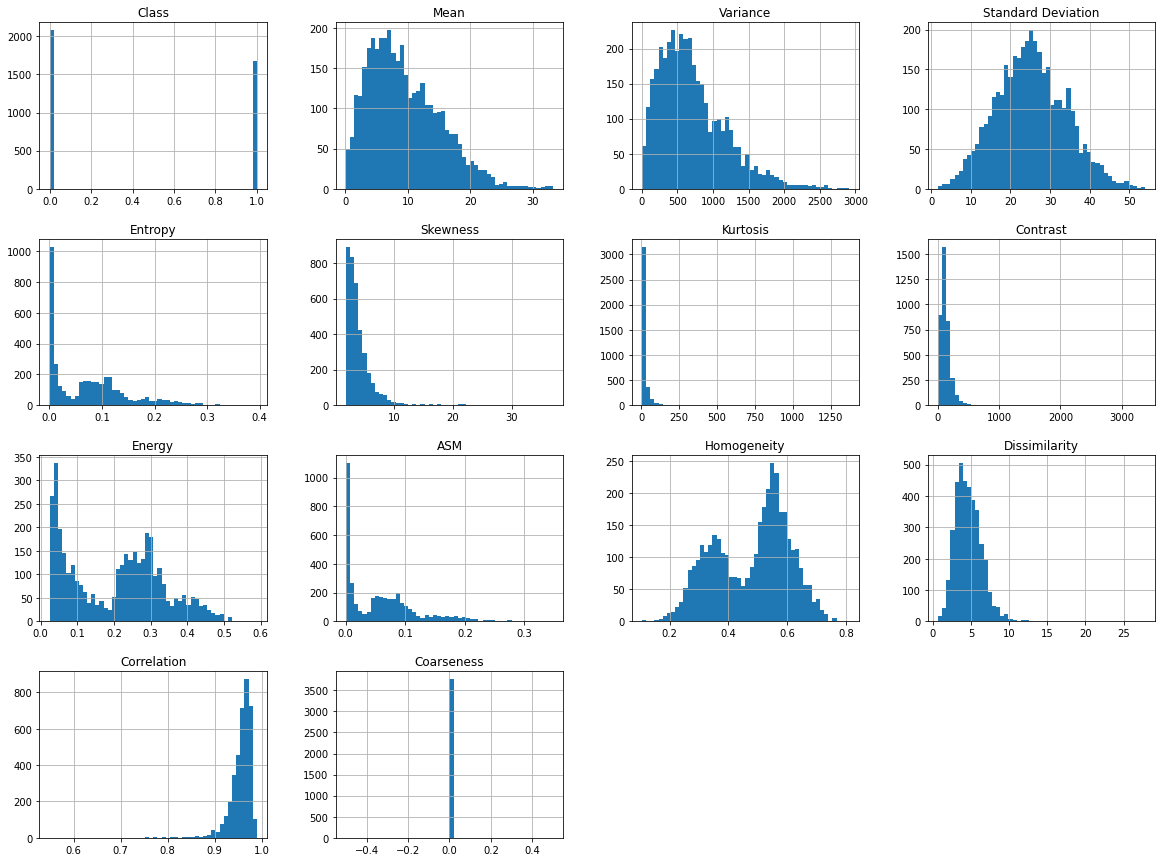
**Analysis**

*Exploratory Data Analysis (EDA)*



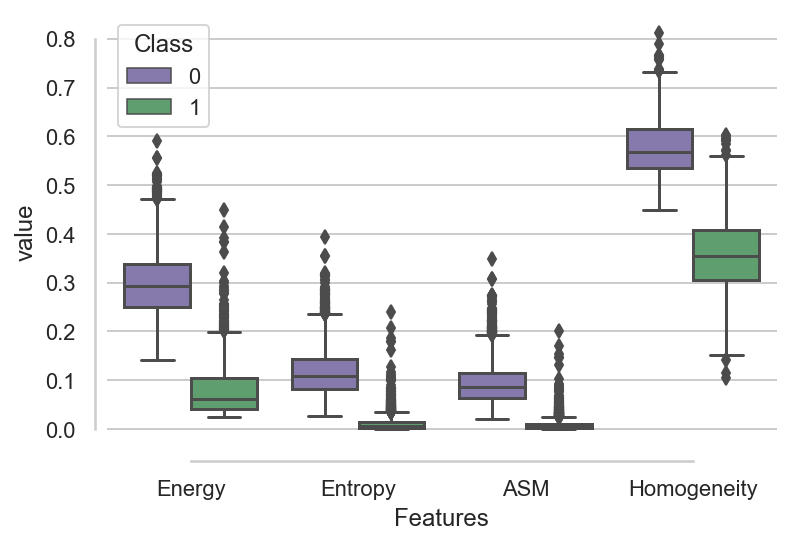
The distribution of class within the dataset includes 2079 observations of the ‘no presence of tumor’ and 1683 observations of ‘presence of a tumor’. Because this was close to an even distribution of the class data, there was not any balancing or synthetic data added to make the dataset more even as this was enough for each class to perform the analysis.

*Data Distribution*

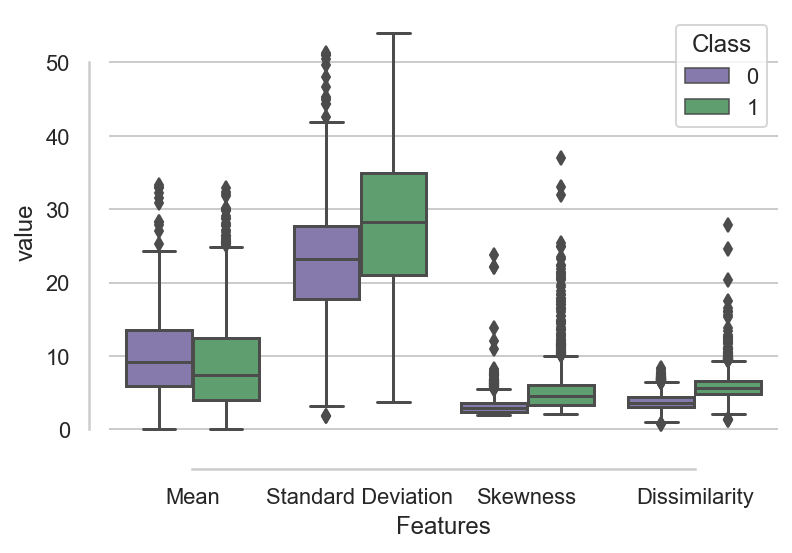


The initial data distribution as displayed above shows Mean, Variance, with right skewed data. Standard deviation and Dissimilarity, shows a normal bell curve distribution. Many of the other features have long tails. To help with this concern, further EDA was performed to understand how the data was distributed among the classes.

Features that Represent Differences Between Class

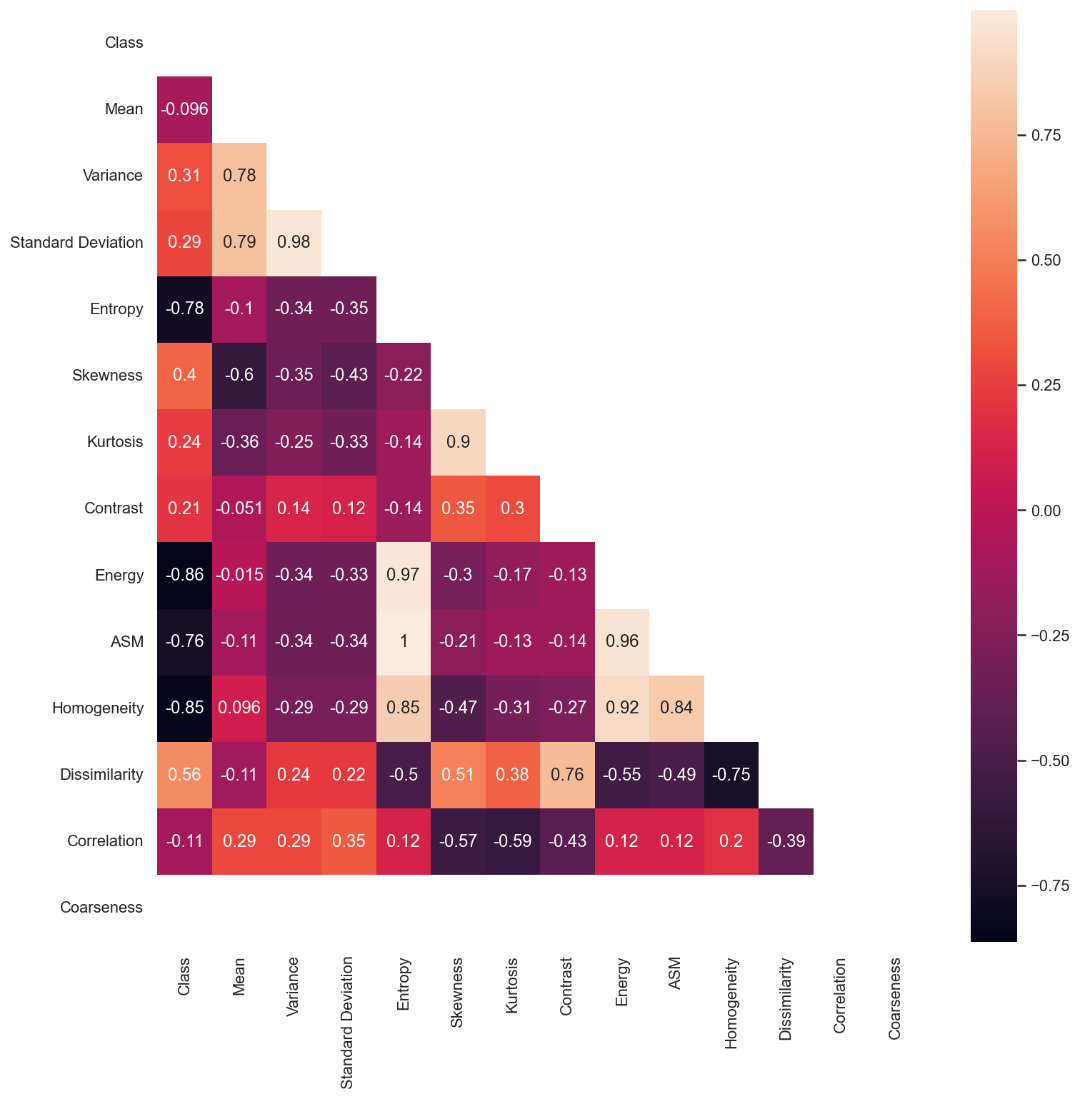


Separating the data by class and viewing the boxplot distribution demonstrates clear differences among the class. Energy, Entropy, ASM and Homogeneity all show clear differences with no overlapping of the interquartile ranges (IQR) of the classes.



Conversely, EDA of the following features: Mean, Standard Deviation, Skewness and Dissimilarity did not show differences between the classes. All the IQR of these features overlap and demonstrate overlapping medians with the IQR.

Correlation Matrix



This matrix visual above includes a plot of Pearson Moment Correlation Coefficients between each of the features and the class variable. The darker boxes display higher correlations, and the lighter boxes display lower correlations. Consistent with what was observed in the boxplots, Entropy, Energy, ASM, and Homogeneity are highly correlated with the Class variable. Other high correlations that need to be watched include Homogeneity and Entropy, Energy and Entropy, and Dissimilarity and Homogeneity.

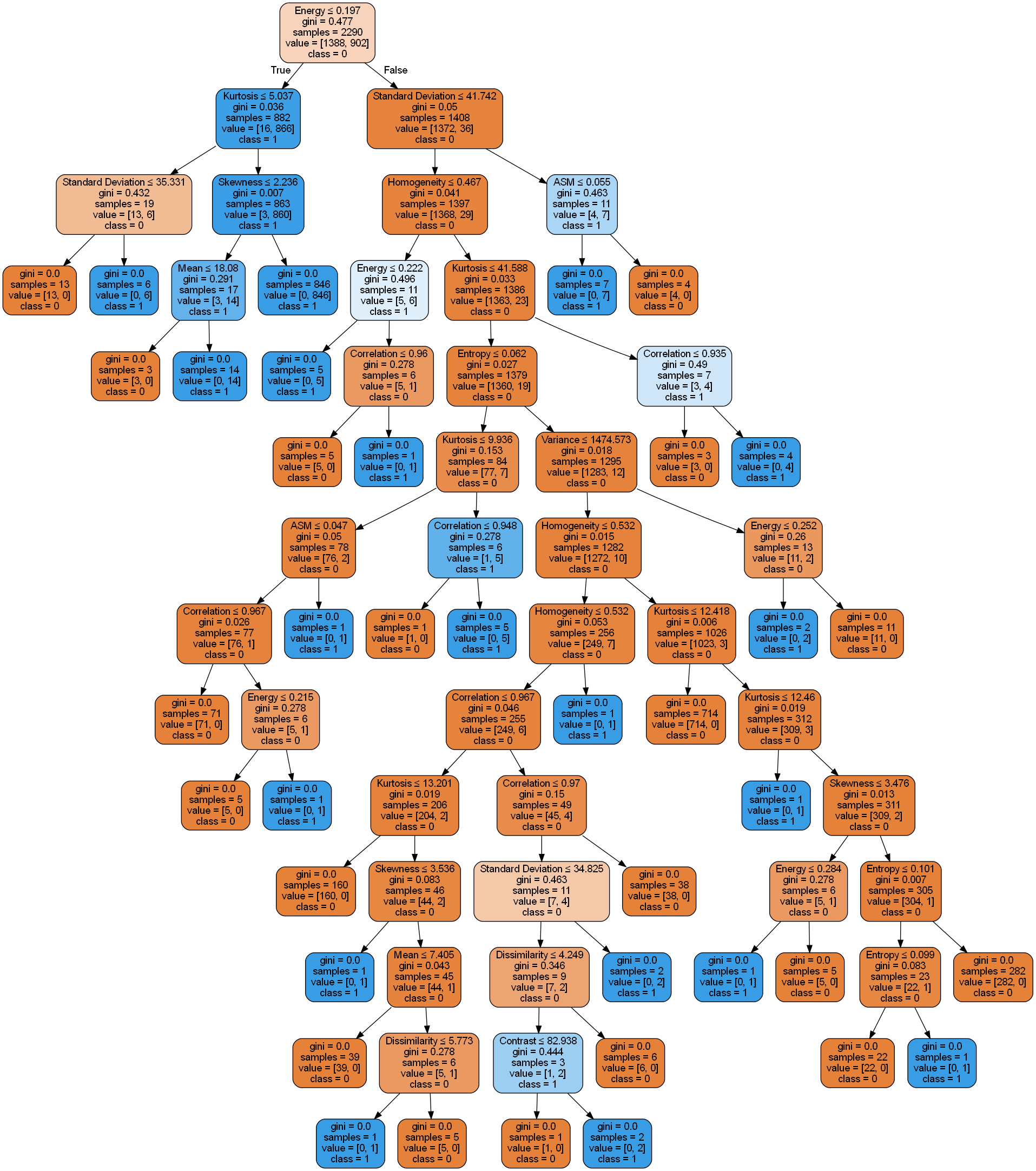
*Randoms Forest*

Chart, bar chart

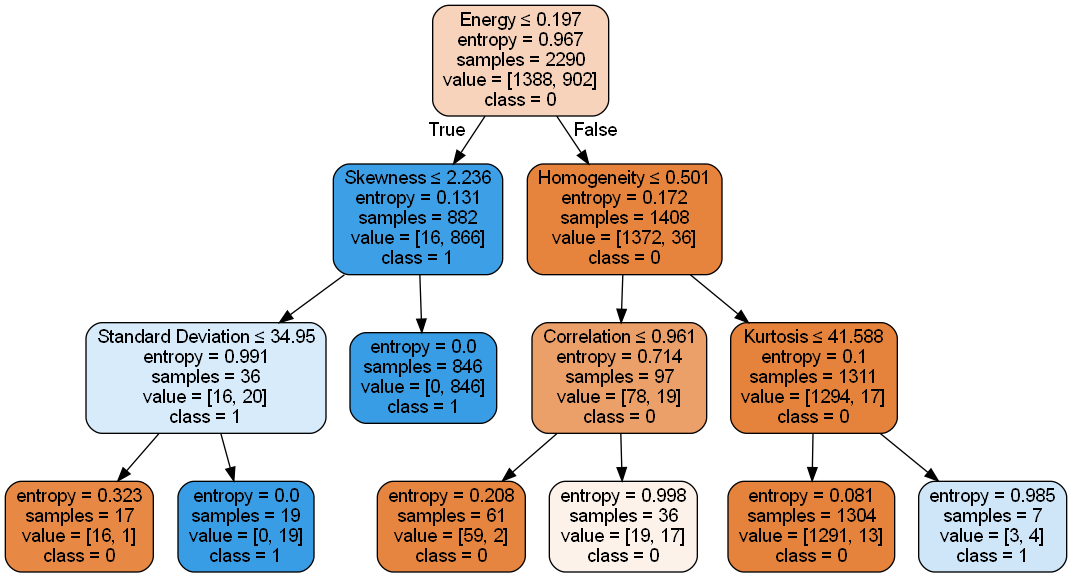
Description automatically generated

The Random Forest model has a 0.98% accuracy. The picture above showed the top 3 importance of the Random Forest feature which are Entropy, ASM, and Energy, with Entropy being 25.56%. The high correlation features having high percentage of the feature importance in the Random Forest model. All the features were being used for the model interpretation, except one feature, which is Coarseness’. The Coarseness feature data was found to have data integrity.

*Decision Tree*



The picture above shows the Decision Tree model. In each node, it has a special decision rule on splitting nodes. This picture is unpruned, and it gives a general idea of the decision tree model.



The picture above is a pruned decision tree, it boosts the performance of the model. In this pruned decision tree, it shows that the Energy being the most important feature among the data. The plot below shows the distribution of the importance in this pruned decision tree. The Energy feature has 90.68% of importance in this pruned decision tree.

Chart

Description automatically generated

*Convolution Neural Network*

To understand if radiomic images could be used to predict the presence of a brain tumor and no presence of a brain tumor, a Convolution Neural Network (CNN) model was developed and ran on the image dataset. A CNN was initialized with 32 filters and 3 kernels using the RELU activation. A pooling size of 2 with 2 strides was also included. There were 2 convolutional layers performed on this model and a sigmoid activation function on the fully connected output layer.

cnn.add(tf.keras.layers.Conv2D(filters=32, kernel\_size=3, activation='relu', input\_shape=[64, 64, 3]))

cnn.add(tf.keras.layers.MaxPool2D(pool\_size=2, strides=2))

cnn.add(tf.keras.layers.Conv2D(filters=32, kernel\_size=3, activation='relu'))

cnn.add(tf.keras.layers.MaxPool2D(pool\_size=2, strides=2))

cnn.add(tf.keras.layers.Flatten())

cnn.add(tf.keras.layers.Dense(units=128, activation='relu'))

cnn.add(tf.keras.layers.Dense(units=1, activation='sigmoid'))

25 Epochs were ran with an “adam” optimizer, and a “binary\_crossentropy” loss. This model returned an accuracy of 96%.

Model: "sequential\_1"  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
 Layer (type) Output Shape Param #   
=================================================================  
 conv2d\_2 (Conv2D) (None, 62, 62, 32) 896   
   
 max\_pooling2d\_2 (MaxPooling (None, 31, 31, 32) 0   
 2D)   
   
 conv2d\_3 (Conv2D) (None, 29, 29, 32) 9248   
   
 max\_pooling2d\_3 (MaxPooling (None, 14, 14, 32) 0   
 2D)   
   
 flatten\_1 (Flatten) (None, 6272) 0   
   
 dense\_2 (Dense) (None, 128) 802944   
   
 dense\_3 (Dense) (None, 1) 129   
   
=================================================================  
Total params: 813,217  
Trainable params: 813,217  
Non-trainable params: 0  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Conclusion**

Across all data observations, beginning with the initial EDA and ending with the modeling, there were consistent features that showed discrimination between classes and correlation with the class variable. The features that were highly correlated and showed differences as represented by the box plots also demonstrated high ranking in the feature importance plots of the random forest and decision tree models. These features include Entropy, ASM, Energy, Homogeneity.

The Decision Tree model made its initial decision split on Energy. Energy was also ranked the highest on the final variable importance plot for the decision tree. Decision tree only ranked 6 variables to be important in helping to make its decisions in classifying the data. Random Forest, on the other hand, used all the variables to help make its decision and classify the data. Random Forest ranked Entropy, ASM and Energy as its top 3 features in determining what was most important in classifying the presence and non-presence of a brain tumor.

Image identification robustness was performed by using a Convolutional Neural Network model. This model using 2 layers of convolution and pooling performed very well in identifying the tumor presence in the images. The model took 10 minutes of CPU resources and used 813,217 params. Further analysis will be performed to understand exactly which features were relevant to this model.

In conclusion, Random Forest was determined to be the best model used for this analysis. It did the best job at using all the feature variables and making the decision to classify the images while also providing a clear understanding of what features were used to make its decision.

**Citations**

Journal of Nuclear Medicine April 2020, 61 (4) 488-495; DOI: <https://doi.org/10.2967/jnumed.118.222893>

[*Visual Computing for Industry, Biomedicine, and Art*](https://vciba.springeropen.com/) **volume 2**, Article number: 19 (2019)