

# Brain Tumor Prediction

Name:	<b>Tej Tarun Sharma</b>
Registration No./Roll No.:	2310225
Institute/University Name:	IISER Bhopal
Program/Stream:	Ph.D Chemistry
Problem Release date:	August 17, 2023
Date of Submission:	November 19, 2023

## Introduction

Dataset of cancer patients is given in the problem statement. This dataset (TCGA dataset)[1] consists total no. of 23 features. 20 features are molecular and 3 features are clinical.

Total Number of training instances = 775

Number of test instances = 87

Number of classes = 2 (GBM and LGG)

Training instances belong to GBM class = 326

Training instances belong to LGG class = 449

A supervised machine learning model has to be trained using the given training data set.

Paper: Tasci, E., Zhuge, et al. Hierarchical Voting-Based Feature Selection and Ensemble Learning Model Scheme for Glioma Grading with Clinical and Molecular Characteristics. International Journal of Molecular Sciences, 23(22), 14155, 2022.[1]

## Methods

This is a classification problem and classifier should be trained to solve the problem. Before training the classifier we have to prepare the dataset for the classifier (i.e. data pre-processing).

### Data Pre-Processing

1. 'Age at diagnosis' feature had continuous values in terms of years and days so I converted all of the values to total age value in days. (for training data as well as test data)
2. For imputing the missing values, I took the mean for the continuous feature and impute the missing values with the mean. And for other features which had categorical values, I assigned the category which was occurring maximum times in that particular feature.
3. To handle the categorical data, I did the one hot encoding using sci-kit learn module.
4. Training data didn't had target values. There was a different csv data file for target values. So, I added training data targets to the training data.

### Feature Selection

After pre-processing, data is ready to train. But there are a lot of features and not every feature is important, so I performed feature selection and explored chi2, mutual information and anova-f techniques to select the best feature values using sk-learn modules.

## **Training**

I used stratified k-fold cross validation to train the data k-times on different disjoint subsets of the original data using sk-learn module. I also got the classification report for different classifiers during the training phase. During the training phase of every classifier, it used grid search to look for the best classification parameters.

## **Testing**

Now when the data is trained, I classified the tested data using that classifier training and got the target labels (LGG and GBM) for every test instance.

## **Experimental Analysis**

Classification report during the training phase of different classifiers:

### **AdaBoost Classifier**

The Probability of Confidence of the Classifier: 0.605  
Precision: 0.9984709480122325  
Recall: 0.9988864142538976  
F1-Score: 0.9986768896150346

### **Decision Tree Classifier**

The Probability of Confidence of the Classifier: 0.995  
Precision: 0.9962393162393162  
Recall: 0.9958189295913209  
F1-Score: 0.9960273319561417

### **Logistic Regression Classifier**

The Probability of Confidence of the Classifier: 0.921  
Precision: 0.9804250863499143  
Recall: 0.9827291732138221  
F1-Score: 0.9815141684953725

### **Multinomial Naive Bayes Classifier**

The Probability of Confidence of the Classifier: 0.985  
Precision: 0.9767917511832319  
Recall: 0.9817624714771749  
F1-Score: 0.978938214232332

### **SVM Classifier**

The Probability of Confidence of the Classifier: 0.842  
Precision: 0.8240688751637657  
Recall: 0.8312200254143496  
F1-Score: 0.817626059571315

## K-Nearest Neighbors Classifier

The Probability of Confidence of the Classifier: 0.752

Precision: 0.7052100840336134

Recall: 0.7085411343544619

F1-Score: 0.7061730735379741

## Random Forest Classifier

The Probability of Confidence of the Classifier: 0.998

Precision: 0.9984709480122325

Recall: 0.9988864142538976

F1-Score: 0.9986768896150346

Observing the results of the experimental analysis, Random forest must be the best classifier for this problem. Classification Report on test data by Random Forest Classifier:

no. of instances for 'LGG': 53

no. of instances for 'GBM': 34

## Discussions

The purpose of developing the optimal machine learning technique for this brain tumor data-set is to help patients as well as doctors so the model should not overfit the data and specificity of the classifier should also be taken care of. Macro averaging is helping us for doing it. In the cited paper.[1], authors used some feature selection techniques and on the bases of the voting they selected novel features for the classifier. This voting from different techniques can have weights according to the limitations of that particular feature selection techniques. We can try weighted-voting based feature selection for selecting the novel features.

## GitHub

This is the link for GitHub Repository where the code and report is uploaded.

<https://github.com/tejtaruns/ML-Course-Project>

## References

- [1] Harpreet Kaur Kevin Camphausen Andra Valentina Krauze Erdal Tasci, Ying Zhuge. Hierarchical voting-based feature selection and ensemble learning model scheme for glioma grading with clinical and molecular characteristics. *Int. J. Mol. Sci.* 2022, 2022.