Brain Tumor Prediction

Name:	**Ronit Raj**
Registration No./Roll No.:	21227
Institute/University Name:	IISER Bhopal
Program/Stream:	EECS
Problem Release date:	August 17, 2023
Date of Submission:	November 19 , 2023

1 Introduction

The project's objective is to create a predictive model for classifying glioma grades in brain tumors, specifically distinguishing between Lower-Grade Glioma (LGG) and Glioblastoma Multiforme (GBM). The training dataset comprises 775 instances with 24 features, and the test dataset has 87 instances with the same 24 features. The two classes in the dataset are LGG and GBM, with 449 instances in the LGG class and 326 in the GBM class in the training set. The features include Gender, Age at diagnosis, Primary Diagnosis, and Race, each having 4, 5, 4, and 4 unique values, respectively. There are 17 missing values across these features, which will be imputed with mode of the respective features. This is a classification problem, given that the classes are discrete (LGG or GBM). I started by label encoding categorical columns in our DataFrame. After that, we apply feature selection techniques to identify the most relevant features. Subsequently, we use SVM, Decision Tree, KNN, Adaptive Boosting, Random Forest, and Logistic Regression classifiers on the preprocessed data. I have fine-tune the model parameters to achieve the optimal precision, recall, and F1 score.

2 Methods

During the preprocessing phase, missing values are handled by replacing them with mean and mode values. One-hot encoding is then applied to the dataset, excluding a single integer column. To enhance dataset cleanliness, all '-' entries are converted to None. For 'Gender' and 'Race' columns, missing values are filled with mode values. The 'Age' column undergoes processing, rounding off years and converting string values to integers. Missing values are imputed using the rounded mean. 'Primary Diagnosis' values intentionally remain unfilled, as they will be addressed in subsequent one-hot encoding.

Feature selection is performed by dropping the 'Primary Diagnosis' column based on its correlation with the target column. This process ensures that all entries in our data are scaled between 0 and 1. Feature selection continues by evaluating correlation scores, enabling the removal of features with high correlation, leaving a refined dataset ready for model construction.

Following dataset preparation, a split into training and testing sets is executed. The model-building phase commences with AdaBoost, followed by additional models geared towards achieving optimal accuracy.

[1] [1, 2].

3 Experimental Setup

In assessing the performance of various machine learning models, we employ diverse evaluation criteria such as accuracy, precision, recall, and F1 score. Each criterion offers unique insights into different

[1] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC

Table 1: Performance Of Different Classifiers Using All Features

Classifier	Precision	Recall	F-measure
Adaptive Boosting	0.87	0.88	0.87
Decision Tree	0.87	0.88	0.88
K-Nearest Neighbor	0.78	0.78	0.78
Logistic Regression	0.84	0.84	0.84
Random Forest	0.88	0.89	0.88
Support Vector Machine	0.86	0.87	0.86

facets of a model's effectiveness. Precision measures the accuracy of positive predictions, recall gauges the ability to capture all relevant instances, and the F1 score provides a balanced assessment considering both precision and recall. The use of these multiple metrics ensures a comprehensive evaluation of each model's overall performance.

Initiating our analysis with the Models mentioned above, we employ grid search to optimize hyperparameters, prioritizing the f1-score.

AdaBoost: The optimal hyperparameters for AdaBoost are {'algorithm': 'SAMME', 'base_estimator': DecisionTreeClassifier(max_depth=1), 'learning_rate': 1, 'n_estimators': 100}. The resultant F1 Score (Macro) achieved by AdaBoost is 0.877.

Decision Tree: For the Decision Tree model, the identified best hyperparameters include a criterion of 'gini,' a maximum depth of 20, and a minimum sample leaf of 4. The resulting metrics indicate an accuracy of 0.877, a precision of 0.882, a recall of 0.877, and an F1 score of 0.878.

Logistic Regression: The best configuration for Logistic Regression involves a regularization parameter (C) of 1, intercept fitting (fit_intercept : True), a maximum iteration limit of 200, multinomial classification (multi_class : 'multinomial'), a penalty type of L2 (penalty : 'l2'), and solver 'sag'. The resulting Macro F1 Score for Logistic Regression stands at 0.8415.

Random Forest: For Random Forest, the best hyperparameters comprise enabling bootstrap (bootstrap : True), a 'gini' criterion (criterion : 'gini'), no specific limit on the maximum depth (max_depth : None), 'auto' for maximum features (max_features : 'auto'), a minimum samples leaf of 2 (min_samples_leaf : 2), a minimum samples split of 5 (min_samples_split : 5), and 100 estimators (n_estimators : 100). The resulting Macro F1 Score for Random Forest is 0.88306.

k-Nearest Neighbors: The identified best hyperparameters for k-Nearest Neighbors encompass {'algorithm': 'auto', 'leaf_size': 10, 'metric': 'manhattan', 'n_neighbors': 7, 'p': 1, 'weights': 'uniform'}. The corresponding performance metrics include an F1 Score (Macro) of 0.7768, an accuracy of 0.8835, a precision of 0.8863, and a recall of 0.8835.

SVM: The best hyperparameters for Support Vector Machine (SVM) are {'C': 1, 'coef0': 0.0, 'degree': 2, 'gamma': 'scale', 'kernel': 'linear', 'shrinking': True}. The resulting Macro F1 Score for SVM is 0.862.

4 Results and Discussion

The most effective model identified in the analysis is [3]AdaBoost, demonstrating superior precision and recall, as highlighted in 1. AdaBoost's prominence is further emphasized by its remarkable accuracy in predicting instances of Glioblastoma Multiforme (GBM), a pivotal consideration given its heightened malignancy compared to Lower-Grade Glioma (LGG). Notably, AdaBoost exhibits minimal misclassifications, with only two inaccuracies in GBM predictions. Additionally, the model

Table 2: Confusion Matrices of Different Classifiers

Actual Class	Predicted Class	
	GBM	LGG
GBM	61	4
LGG	16	74

Adaptive Boosting

Actual Class	Predicted Class	
	GBM	LGG
GBM	57	8
L	13	77

Decision Tree

Actual Class	Predicted Class	
	GBM	LGG
GBM	52	13
LGG	19	71

K-Nearest Neighbor

Actual Class	Predicted Class	
	GBM	LGG
GBM	54	11
LGG	13	77

Logistic Regression

Actual Class	Predicted Class	
	GBM	LGG
GBM	61	4
LGG	16	74

Random Forest

Actual Class	Predicted Class	
	GBM	LGG
GBM	57	8
LGG	13	77

SVM

demonstrates a limited rate of false predictions for LGG, misjudging it only 16 times out of 90 instances, surpassing other models delineated in 2.

The decision to designate AdaBoost as the primary classifier for predicting labels on the test data is justified by its robust accuracy, particularly in GBM classification. To maintain consistency between the training and test datasets, features present during training but absent in the test data are appended as zero arrays of size 87 (the size of the test data). This meticulous approach ensures comprehensive prediction capabilities, solidifying AdaBoost's role as the optimal classifier for this specific task.

5 Conclusion

It is important to explore the potential for overfitting by analyzing the relationship between individual attributes and target labels. It is important to exercise care even when attaining high accuracy on the validation set, as this does not always translate into trustworthy predictions on the test data. One of the challenges in addressing potential inaccuracies in model predictions is the necessity of rearranging columns.

Furthermore, it remains difficult to determine if a strongly linked characteristic indicates overfitting or functions as a reliable estimator. The analysis is made more difficult by the absence of a conclusive technique to differentiate between the two.

Analysing the data shows that the predictions from different models are similar. Because Glioblastoma Multiforme (GBM) is a severe disorder, models with greater accuracy in predicting GBM cases tend to be selected based on this criterion. This highlight aligns with the importance of accurate predictions, particularly in critical cases like GBM.

`https://github.com/r0ny4425/Machine-Learning`

References

[1] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[2] I. H. Witten, E. Frank, and M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, third edition, 2011.

[3] Robert E Schapire. Explaining adaboost. In *Empirical inference*, pages 37–52. Springer, 2013.