Integrated Molecular and Clinical Feature Selection for Accurate Glioma Grading

Name:	**Vedika Vikas Jakate**
Registration No./Roll No.:	21330
Institute/University Name:	IISER Bhopal
Program/Stream:	EECS
Problem Release date:	August 17,2023
Date of Submission:	September 17,2023

Introduction

The dataset has 775 instances in the training data set and 87 instances in the test dataset. The number of features in the dataset are 24. The feature '*Age_at_diagnosis*' is of the form ' 36 years 100 days' which is continuous in nature. All the other features are categorical in nature. The features '*Age_at_diagnosis*','*Gender*' and '*Race*' have missing values in them.

```
Number of missing values before handling:
Gender                4
Age_at_diagnosis      5
Race                 21
```

Figure: Missing values

In the feature '*Gender*' there are two classes Male and Female. In the feature '*Race*' there are 4 classes white, black or african american, asian and american indian or alaska native.The feature '*Primary_diagnosis*' has various types of gliomas. All the other features have only two classes '$MUTATED$' and '$NOT\_MUTATED$'.

Methods

1.Data Loading
2.Dropping '*Primary_diagnosis*' from the data: The feature '*Primary_diagnosis*' was dropped because it was observed that whenever the class in '*Primary_diagnosis*' is '*Glioblastoma*' the class of target variable is '*GBM*'and otherwise the class of the target variable is '*LGG*'.
3.Handling of Missing values: In order to handle the missing values in the data I used the imputation method '$most\_frequent$' which replaces the missing values with the most frequent class in that feature.
4.Modification in the feature '*Age_at_diagnosis*': Initially all the instances in this feature were of type string.Instances in this feature has two parts years and days. I modified the feature to include only the years part of the feature. Hence the string '36 years 100 days' was now converted to integer 36.
5.Encoding: In order to convert categorical variables to numerical variables I performed Label Encoding on the data.
6.Feature Selection Techniques: At first I computed the correlation matrix to see which of my features were correlated with each other. But observed that my features were not correlated to each other. Therefore, I implemented Recursive Feature elimination and Selecting K-best feature techniques in order to perform feature selection.

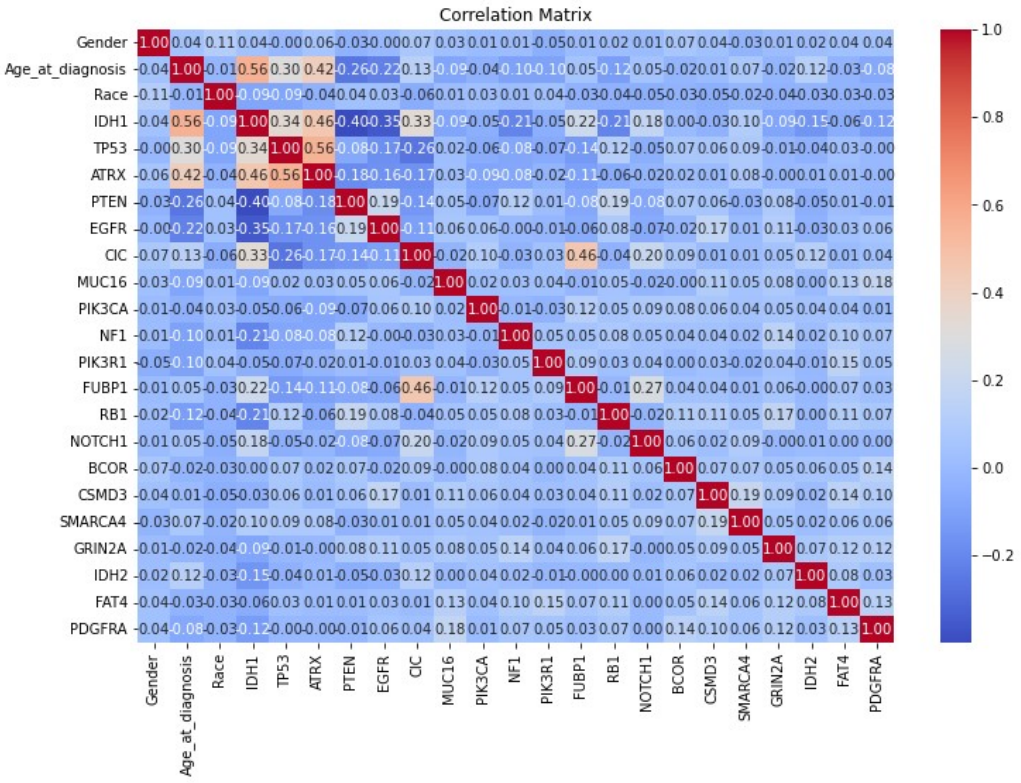

Figure: Correlation matrix

7.Splitting of data: The training data was split into two parts one was the training data and the other was the test data. The classification algorithms will be first implemented on the training data and then on the test data to check the working on the data that it has not seen before. This step can reduce the chances of data overfitting.
8.Classification Algorithms: I used classification algorithms including Decision Tree, Logistic Regression, Support Vector Machine, Multinomial Naive Bayes, Random Forest K-Nearest Neighbours and AdaBoost in order to solve the problem.
8.Hyperparameter Tuning: In order to improve the accuracy of the model, hyperparameter tuning was performed on each of the models so that best parameters are chosen. This was done with the help of Grid search. The best parameters obtained for each model after hyperparameter tuning are listed below:
1.Decision Tree: $criterion : gini, max\_depth : None, min\_samples\_leaf : 4, min\_samples\_split : 2, splitter : random$
2.Logistic Regression: $C : 1, penalty : l2, solver : liblinear$
3. Support Vector Machine: $C : 1, kernel : linear$
4.Multinomial Naive Bayes: $alpha : 0.1, fit\_prior : True$
5. Random Forest: $criterion : gini, max\_depth : None, min\_samples\_split : 10, n\_estimators : 100$
6.K-Nearest Neighbours: $n\_neighbors : 7, p : 1, weights : distance$
7. AdaBoost: $learning\_rate : 0.01, n\_estimators : 50$
9.Implementation of best model on test data and prediction of labels: The best model is chosen to be implemented on the test data and the target labels are predicted.
Link to the Github:Brain Tumor Prediction

Experimental Analysis

I conducted a comprehensive analysis by running all the classification algorithms using two feature selection techniques: Recursive Feature Elimination (RFE) and K-best Features. To ensure a thorough evaluation, I systematically altered the number of features to be selected, experimenting with values of 20, 15,10 and 5. This allowed me to assess the impact of different feature subsets on the performance

of each algorithm. The evaluation metrics were then obtained for each combination of algorithm and feature selection method, providing insights into the optimal number of features for effective classification. The performance of the model on training data and test data(split from training data) was analysed.

```
Support Vector Machine - Accuracy: 0.7483870967741936
Confusion Matrix:
 [[44 21]
 [18 72]]
Classification Report:
              precision    recall  f1-score   support

         GBM       0.71      0.68      0.69        65
         LGG       0.77      0.80      0.79        90

    accuracy                           0.75       155
   macro avg       0.74      0.74      0.74       155
weighted avg       0.75      0.75      0.75       155
```

Figure: Classification Report for SVM before feature selection and Hyperparamater Tuning

```
Selected Features:
Index(['Gender', 'Age_at_diagnosis', 'Race', 'IDH1', 'TP53', 'ATRX', 'PTEN',
       'EGFR', 'CIC', 'MUC16', 'NF1', 'PIK3R1', 'FUBP1', 'RB1', 'NOTCH1',
       'CSMD3', 'SMARCA4', 'GRIN2A', 'IDH2', 'PDGFRA'],
      dtype='object')

Hyperparameter Tuning for SVC:
Best Parameters: {'C': 1, 'kernel': 'linear'}
Best F1_score: 0.8813080639167594
```

Figure: 20 best selected features('*k_best*') and best parameters after Hyperparameter tuning

```
Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.97      0.86        65
           1       0.97      0.80      0.88        90

    accuracy                           0.87       155
   macro avg       0.88      0.88      0.87       155
weighted avg       0.89      0.87      0.87       155
```

Figure: Classification Report for SVM after Feature Selection and Hyperparameter Tuning

```
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.92      0.85        65
           1       0.94      0.82      0.88        90

    accuracy                           0.86       155
   macro avg       0.86      0.87      0.86       155
weighted avg       0.87      0.86      0.87       155
```

Figure: Classification Report for Logistic Regression after Feature selection and Hyperparameter Tuning

In the initial phase, prior to feature selection and hyperparameter tuning, the Support Vector Machine exhibited an accuracy of 0.74. However, upon implementing feature selection and fine-tuning hyperparameters, there was a significant improvement in the model's performance, resulting in an elevated F1 score of 0.87. Furthermore, during the analysis, when k-best feature selection or recursive feature elimination were implemented with varying numbers of features (20, 15, 10, 5), it was observed that the F1 score remained consistent across these different subsets. This consistency in performance

F Score	Model('*k_best*', 20 features)
0.80	Decision Tree
0.86	Logistic Regression
0.87	SVM
0.86	Multinomial Naive Bayes
0.86	KNN
0.82	AdaBoost

Table 1: F score for models

across diverse data subsets suggests that the Support Vector Machine is robust and performs reliably across various feature sets. Other than SVM, Logistic Regression, K nearest Neighbours and Multinomial Naive Bayes also provide a f score of 0.86 but these are not robust enough as the f score varies with different data subsets.

Discussions

The chosen classification algorithm for addressing the brain tumor prediction problem is Support Vector Machine (SVM) with '*k_best*' feature selection (20 features). SVMs demonstrate effectiveness in high-dimensional spaces, making them particularly suitable for problems characterized by a large number of features, such as the molecular mutation and clinical data in this context.One key strength of SVM is its proficiency in binary classification tasks. Given that the target variable in this problem comprises two classes, namely '*GBM*' and '*LGG*' SVM proves to be a well-suited choice. The algorithm aims to find a hyperplane that maximally separates these two classes, making it effective for discerning patterns and relationships in the given dataset. But SVM can be computationally expensive. Training SVM with large number of features and instances might require considerable computational resources.

References

1.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.
2.Heba Abusamra,A Comparative Study of Feature Selection and Classification Methods for Gene Expression Data of Glioma,Procedia Computer Science,Volume 23,2013,Pages 5-14