GILOMA GRADING

Artificial Intelligence | T11 | A2_110

Adriano Machado – 202105352
Diogo Fernandes – 202108752
João Torre Pereira - 202108848

The problem

Context:

- Gliomas are the most common primary tumors of the brain. They can be graded as LGG (Lower-Grade Glioma) or GBM (Glioblastoma Multiforme) depending on the histological/imaging criteria. Clinical and molecular/mutation factors are also very crucial for the grading process. Molecular tests are expensive to help accurately diagnose glioma patients.

Goal:

- The prediction task is to determine whether a patient is LGG or GBM with a given clinical and molecular/mutation features. Our goal is to develop a machine learning classifier model to detect LGG or GBM based on gene molecular mutation, information which can be obtained from a DNA test.

Related work

In our work we drew inspiration from the following sources:

- Tasci E, Jagasia S, Zhuge Y, Camphausen K, Krauze AV. GradWise: A Novel Application of a Rank-Based Weighted Hybrid Filter and Embedded Feature **Selection Method for Glioma Grading with Clinical and Molecular Characteristics**. Cancers (Basel). 2023 Sep 19;15(18):4628. doi: 10.3390/cancers15184628. PMID: 37760597; PMCID: PMC10526509.

Dataset

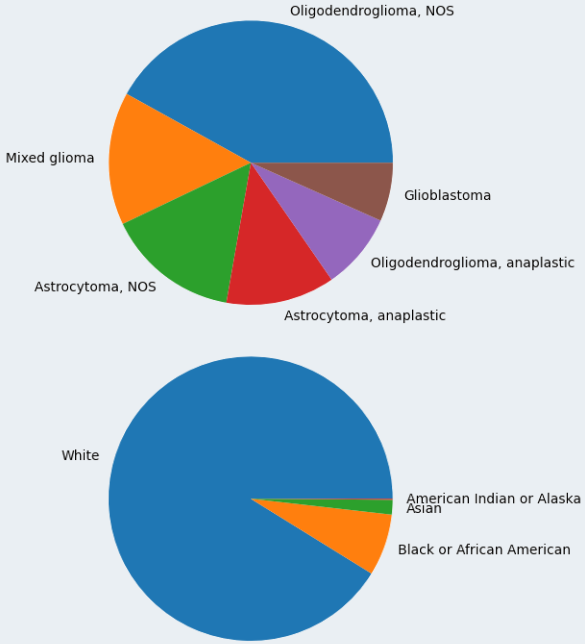
The Dataset is composed of two widely employed genome atlas databases, the Cancer Genome Atlas (TCGA) and the Chinese Glioma Genome Atlas (CGGA). TCGA consists of 3 clinical features, the most frequently mutated 20 molecular / mutation features plus the class labels. The only feature missing from CGGA is the race feature, as the dataset is derived from a Chinese population.

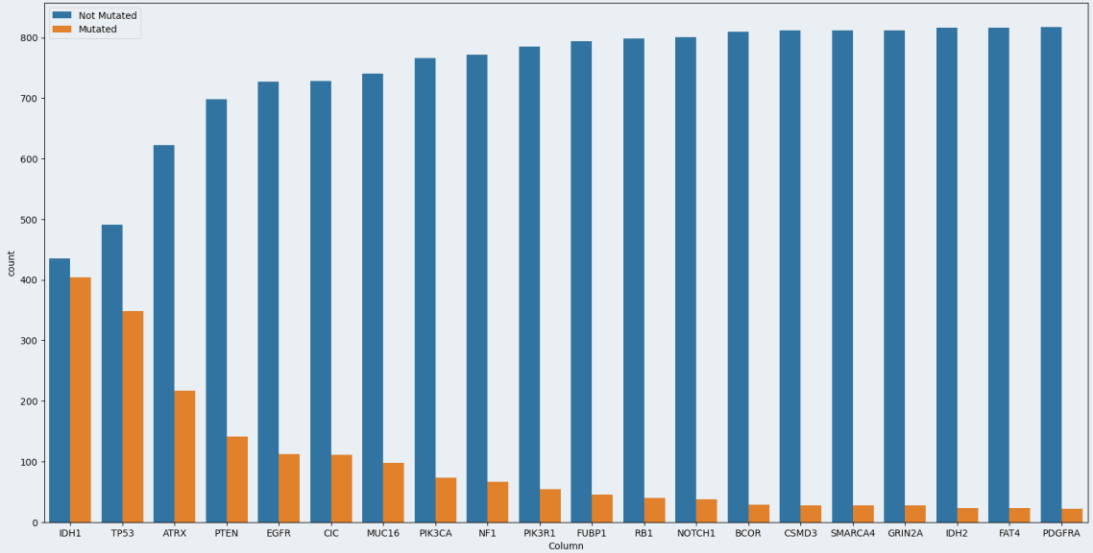
Features:

Clinical features: Gender, Age at diagnosis, Race
Molecular features: IDH1, TP53, ATRX, PTEN, EGFR, CIC, MUC16, PIK3CA, NF1, PIK3R1, FUBP1, RB1, NOTCH1, BCOR, CSMD3, SMARCA4, GRIN2A, IDH2, FAT4, PDGFRA
Class feature: Grade (LGG or GBM)

The molecular features tells us if the specific molecular gene is mutated or not.

Dataset analysis





Tools and algorithms

We selected **Python** as our programming language due to its powerful packages such as Scikit-Learn, Pandas, and Matplotlib. To minimize the number of features in the dataset, we applied feature selection techniques, including tree-based methods, LASSO L1-based selection, and recursive feature elimination.

For classifying glioma patients as LGG or GBM, we decided to implement the following machine learning algorithms:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- K Nearest Neighbors (KNN)
- Random Forest (RF)
- AdaBoost
- Neural Network (NN)

We also used soft voting-based model selection algorithms to evaluate how the models perform together.

Data pre-processing

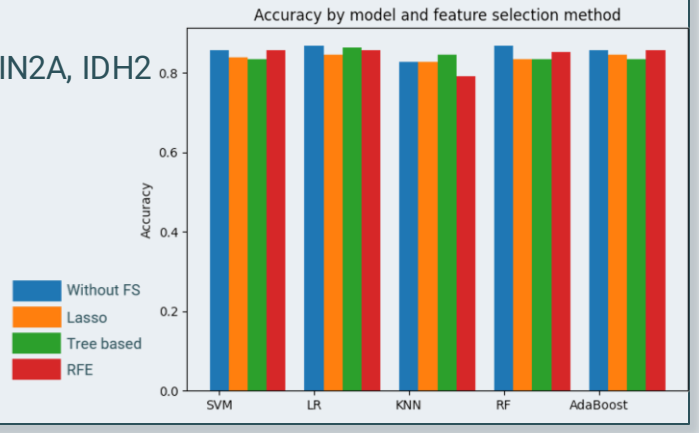
We employed three different feature selection approaches to identify the most important features in the dataset. Here are the results:

LASSO L1-based feature selection (12/22)
Age_at_diagnosis, IDH1,TP53, PTEN, CIC, MUC16, NF1, PIK3R1, RB1, NOTCH1, GRIN2A, IDH2

Tree based feature selection (4/22)
Age_at_diagnosis, IDH1, PTEN, CIC

Recursive feature elimination (11/22)
IDH1, TP53, PTEN, CIC, MUC16, NF1, PIK3R1, NOTCH1, SMARCA4, GRIN2A, IDH2

Comparing the three approaches, we observed that, in most cases, not using feature selection generally improved accuracy. However, there were some exceptions: tree-based feature selection yielded the highest accuracy for KNN, and recursive feature elimination resulted in the highest accuracy for AdaBoost.



Work carried out

After performing data preprocessing, where we eliminated certain incomplete rows and selected features, we implemented the following algorithms. You can see the results on the right:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- K Nearest Neighbors (KNN)
- Random Forest (RF)
- AdaBoost
- Neural Network (NN)

	LR	SVM	KNN	RF	AB	NN
Accuracy	0.869	0.857	0.827	0.869	0.857	0.839
Precision	0.875	0.868	0.829	0.873	0.868	0.841
Recall	0.869	0.857	0.827	0.869	0.857	0.839

Ensembling

We evaluated different combinations of these models using a VotingClassifier, testing combinations of up to five models to determine the best performing ensemble.

The performance was measured using accuracy on the test data. The best combination, consisting of SVM, KNN, and NN, achieved an accuracy of 0.88. This demonstrates that ensembling can significantly enhance model performance by leveraging the strengths of multiple algorithms.

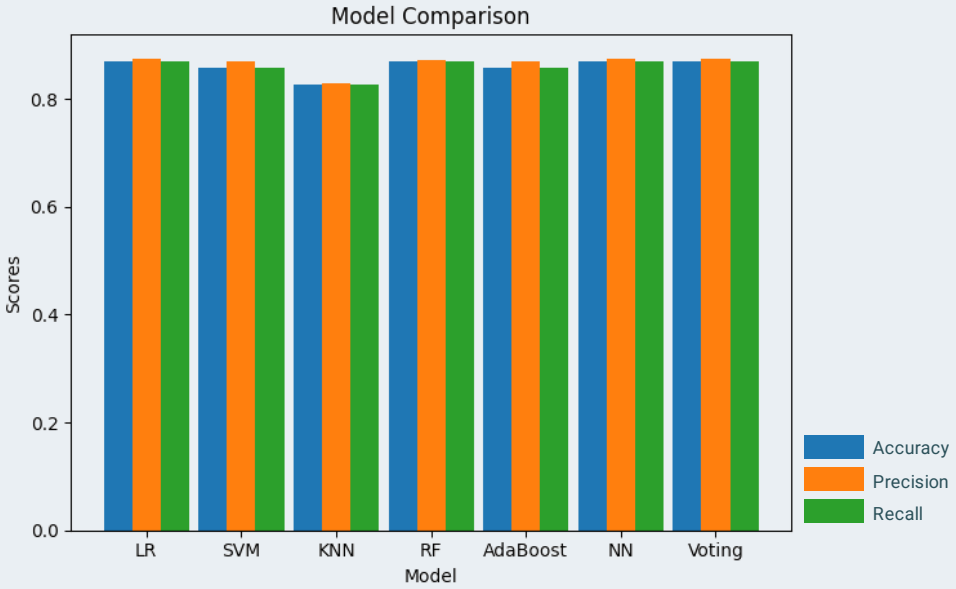
	Ensemble
Accuracy	0.88
Precision	0.88
Recall	0.88

Algorithms comparasion

By analyzing the results, we reached the following conclusions:

- Neural Network, Random Forest, and the voting ensemble model consistently perform well across accuracy, precision, and recall metrics.

- KNN is the least effective model among those compared, with relatively low scores across all three metrics.



Conclusion

In this analysis of glioma grading using the TCGA-LGG and TCGA-GBM datasets, we explored and pre-processed clinical and molecular features. We applied several classification algorithms (Logistic Regression, SVM, KNN, Random Forest, AdaBoost, Neural Network) and evaluated them using metrics like accuracy and confusion matrices.

Ensemble modeling with a Voting Classifier showed improved accuracy and robustness.

Our study highlights optimal feature subsets and ensemble strategies, paving the way for better clinical decision-making and personalized treatments for glioma patients.