Glioma Grading Clinical and Mutation Features

Supervised Learning

- Gliomas are the most common primary brain tumors.
- Based on histological/imaging criteria, they can be classified as:
  - LGG (Lower-Grade Glioma)
  - GBM (Glioblastoma Multiforme)
- For the grading process, clinical and molecular/mutation factors are highly important, and molecular tests for accurately diagnosing glioma patients are costly.

Problem Description

- ❏ This is a supervised learning problem where the main goal is to leverage classification algorithms to **grade gliomas** based on **clinical and genetic mutation features**.
- ❏ More specifically, we are trying to determine whether a glioma patient has **LGG** (Lower-Grade Glioma) or **GBM** (Glioblastoma Multiforme).
- ❏ Additionally, we are also trying to **find the optimal subset of mutation genes and clinical features** for the glioma grading process to **improve performance** and **reduce costs**.
- ❏ The given dataset represents records of patients who have brain glioma. Each record is characterized by **20 molecular features**, each of which can be *mutated* or *not_mutated*, and **3 clinical features**.

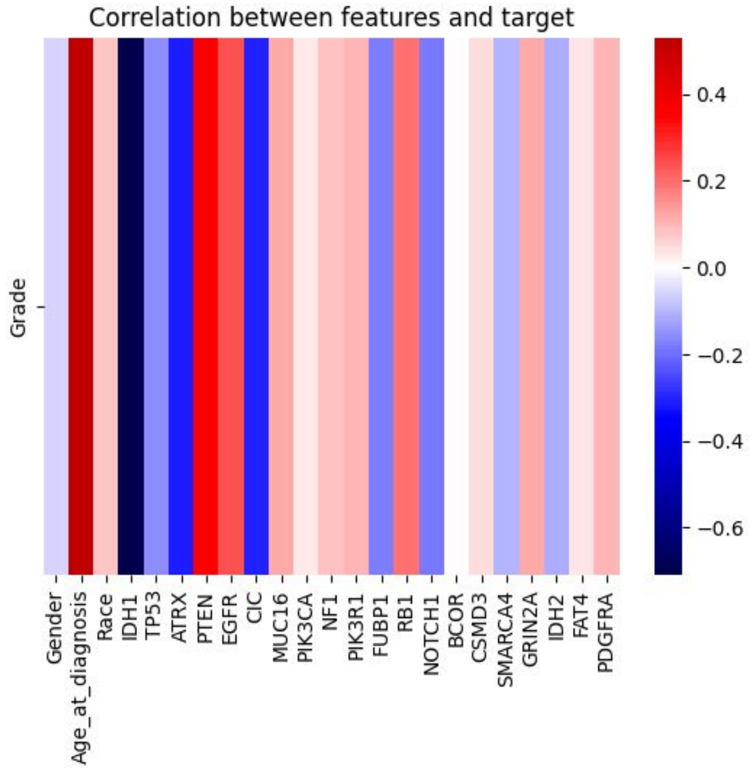
Tools and Algorithms

- **Programming Language** - Python
- **Development Environment** - Jupyter Lab
- **Libraries/Packages** - NumPy, MatPlotLib, Seaborn, Pandas, SciKit-Learn.
- **Supervised Learning Classification Algorithms:**
  - Nearest Neighbors
  - Decision Tree
  - Support Vector Machine
  - Neural Network (Multi-layer Perceptron)
  - Gaussian Naive Bayes
  - Random Forests
  - Gradient Boosting

Data Pre-processing

- **Dataset analysis**
  - Missing data
  - Redundant features
  - Data imbalances
  - Outliers

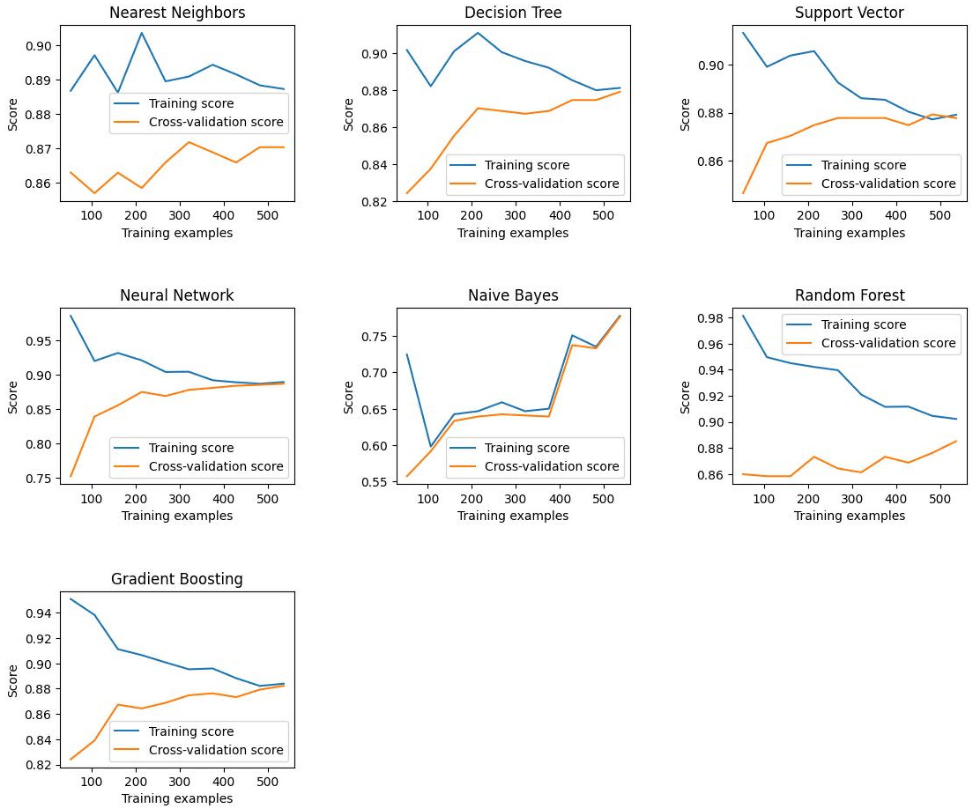
- **Data pre-processing**
  - Imputation or removal of missing data
  - Encode categorical variables
  - Normalize and standardize features
  - Remove or correct outliers
  - Feature extraction



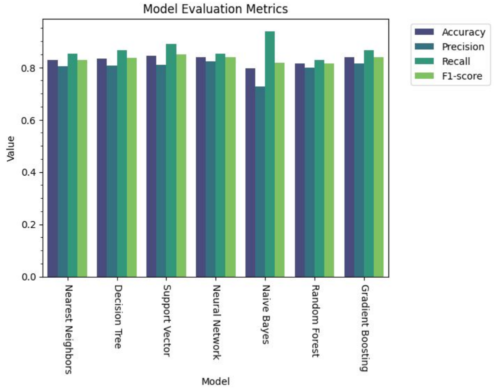
Data Splitting and Training

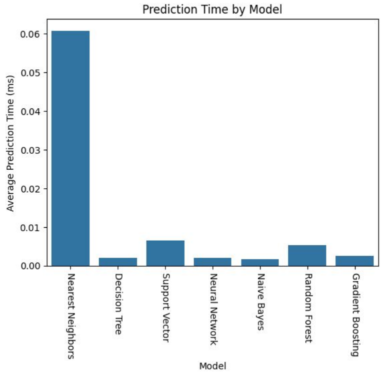
- Each model was **trained** on **80%** of the data, using the remaining **20%** for **testing**.

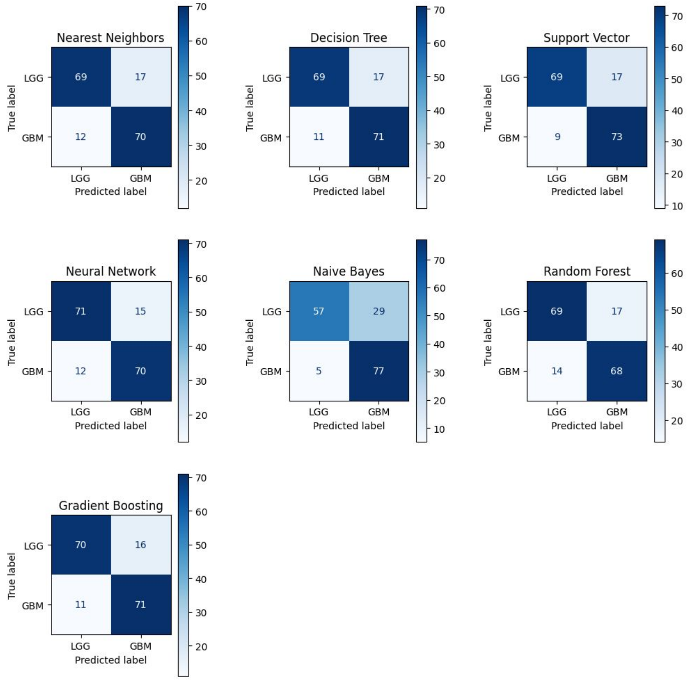
- The optimal settings for each model were determined through **Grid Search** using **Stratified K-Fold** with **10** folds for **Cross Validation**.

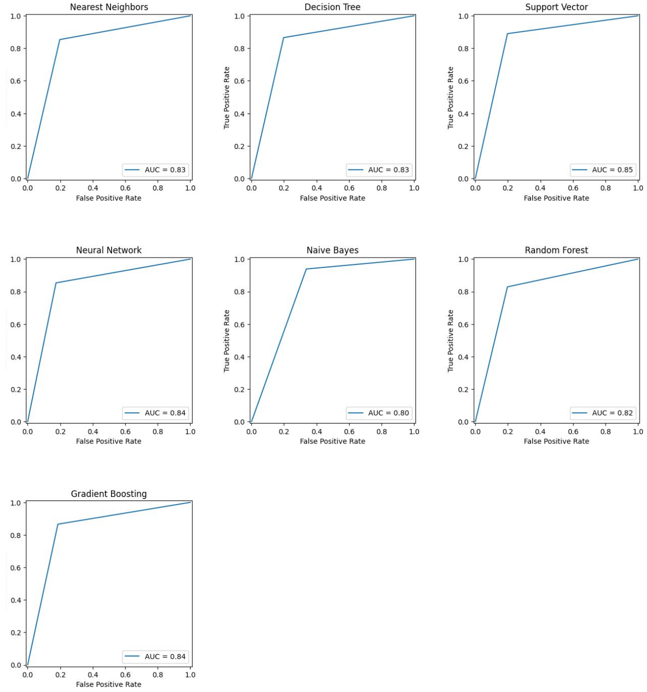


Model Evaluation and Comparison









Conclusions

- The **data pre-processing** phase of the project allowed us to simplify the original dataset and merely keep relevant data.

- The **pre-analysis** of the dataset gave us some valuable insights that denoted parallels with the conclusions drawn later.

- The **grid search** approach, with **stratified k-fold cross-validation**, allowed us to find the best possible performance for each selected model.

- The **stratified k-fold cross-validation** ensured that the models were trained and tested on **balanced data**, guaranteeing the models' **generalization** to unseen data.

Conclusions

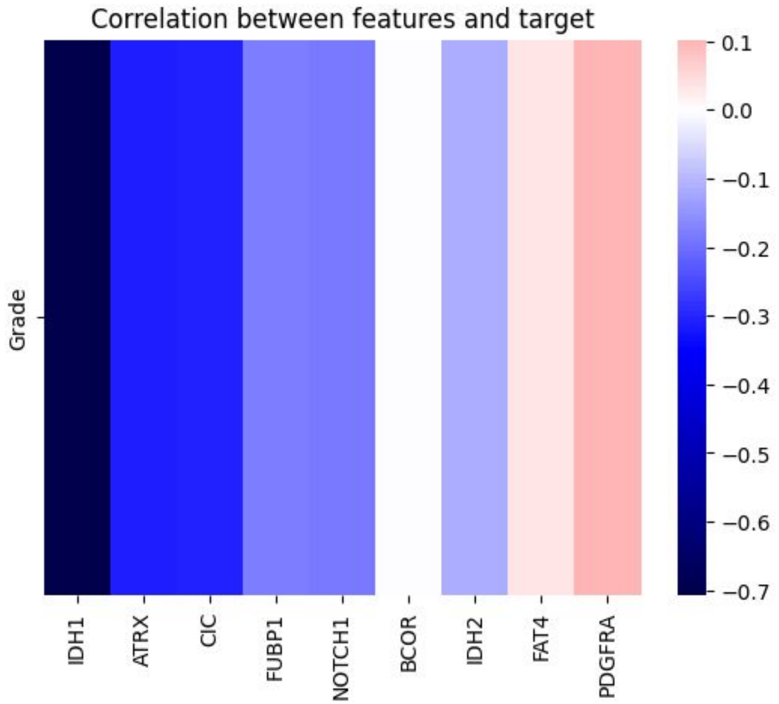
- ❏ The **Support Vector Machine**, **Neural Network** and **Gradient Boosting** models had the overall **best performance**.
- ❏ The **Nearest Neighbor** model performed the **worst**, with the **lowest accuracy** and, by far, the **highest prediction time**.
- ❏ The **Decision Tree** model performed well too, and provided a clear **insight into the importance of each feature**.
- ❏ We also computed a **weighted cost** of the entries of the confusion matrices, and the **Gaussian Naive Bayes** model **performed the best**.
- ❏ Stating that a model is the **best suited** for a **given classification task** depends on the **criteria** that are **most important for the problem** at hand.

Extra - Dimensionality Reduction

- The goal was to **reduce the dimensionality** of the dataset, allowing for **faster training and prediction times**, while still maintaining a good level of classification performance.

- We used **recursive feature elimination** with **cross-validation** on the best performing model, the Support Vector Machine.

- The accuracy remained the same and the **prediction time improved** by about **70%**.



References

- Glioma Grading Clinical and Mutation Features. Retrieved from https://archive.ics.uci.edu/dataset/759/glioma+grading+clinical+and+mutation+features+dataset. Accessed April 25, 2024.
- Hierarchical Voting-Based Feature Selection and Ensemble Learning Model Scheme for Glioma Grading with Clinical and Molecular Characteristics. Retrieved from https://www.semanticscholar.org/paper/Hierarchical-Voting-Based-Feature-Selection-and-for-Tasci-Zhuge/992bf4c0b92ef251644ac2854dd1baacd7e42dc5. Accessed April 25, 2024.
- SciKit-Learn User Guide Supervised Learning. Retrieved from https://scikit-learn.org/stable/supervised_learning.html. Accessed April 25, 2024.
- Classification: ROC Curve and AUC. Retrieved from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=pt-br. Accessed May 25, 2024.
- Russell, S., & Norvig, P. (2010). Artificial intelligence: A modern approach (3rd ed.). Prentice Hall.