The Dataset

Group 5 – ML2 Group Project

Dataset Overview

The dataset used for this project is the Glioma Grading Clinical and Mutation Features dataset, obtained from Kaggle:
https://www.kaggle.com/datasets/statjourney/glioma-grading-clinical-and-mutation-features/data.
It was originally derived from The Cancer Genome Atlas (TCGA), a large-scale genomics program providing molecular and clinical data for various cancer types.
This dataset focuses on glioma, a category of brain tumors classified mainly into:

- **Low-Grade Glioma (LGG)** — slower-growing and less aggressive tumors.
- **Glioblastoma Multiforme (GBM)** — highly malignant and fast-progressing tumors.

Each record represents a single patient and includes:

- Clinical and demographic information, such as age, gender, and race.
- Diagnostic data, including the tumor's primary histopathological category.
- Genetic mutation indicators for key oncogenes and tumor suppressor genes known to influence glioma grade.

In total, the dataset contains approximately 1,500 patient observations and over 43 features (columns).

Structure and Key Variables

The variables fall into three main categories:

Category	Description	Example Variables
Target Variable	Indicates tumor grade (our classification label).	Grade
Clinical and Demographic Variables	Describe patient background and diagnosis information.	Age_at_diagnosis, Gender, Race, Primary_Diagnosis
Genetic Mutation Variables	Binary indicators for the presence or absence of mutations in specific genes.	IDH1, TP53, ATRX, EGFR, PTEN, CIC, NOTCH1, NF1, PIK3CA, SMARCA4, etc.

1. Target Variable

Variable	Description	Type	Coding
Grade	Tumor grade classification used as the target variable for supervised learning.	Categorical (binary)	0 = Low-Grade Glioma (LGG) 1 = Glioblastoma Multiforme (GBM)

This variable represents the label the model tries to predict, based on all the other patient features.

Clinical and Demographic Variables

Variable	Description	Type	Coding / Preprocessing
Age_at_diagnosis	Age of the patient at diagnosis. Originally expressed as a string (e.g., “55 years 210 days”).	Numeric → Ordinal	Extracted numerical age in years, then binned into 5 categories: 0–20, 21–40, 41–60, 61–80, 80+.
Age_bin	Age group derived from age numeric values.	Ordinal (categorical)	One-hot encoded with 4 binary columns (e.g., Age_bin_21_40, etc.).
Gender	Patient sex.	Categorical	Encoded as one-hot columns: Male, Female, Unknown.
Race	Self-identified race or ethnicity.	Categorical	Common values: White, Black or African American, Asian, Unknown. Encoded using one-hot encoding.
Primary_Diagnosis	Histopathological tumor subtype assigned by clinicians.	Categorical	Common values: Astrocytoma, Oligodendroglioma, Glioblastoma. One-hot encoded with drop='first' to avoid redundancy.

During preprocessing, missing or non-standard entries such as '--' or 'not reported' were replaced by 'Unknown'. This choice preserves data integrity and prevents record loss while signalling missingness to the model.

Genetic Mutation Variables

These columns indicate whether specific genes are mutated or not mutated in each patient.
They are all binary categorical variables, originally expressed as text ("MUTATED" or "NOT_MUTATED") and converted to numeric codes for modelling.

Variable (examples)	Description	Type	Coding
IDH1	Mutation in the IDH1 gene — often present in LGG.	Binary	0 = NOT_MUTATED, 1 = MUTATED
TP53	Tumor suppressor gene, frequently mutated in both LGG and GBM.	Binary	Same encoding
ATRX	Associated with chromatin remodeling; often co-mutated with IDH1 in LGG.	Binary	Same encoding
EGFR	Epidermal Growth Factor Receptor gene; highly mutated in GBM.	Binary	Same encoding
PTEN	Tumor suppressor gene involved in cell regulation; often mutated in GBM.	Binary	Same encoding
CIC, FUBP1, NOTCH1, NF1, PIK3CA, SMARCA4, etc.	Additional relevant mutations contributing to glioma characterization.	Binary	Same encoding

The encoding step ensures that all mutation features are represented numerically (0 or 1), allowing models like Random Forest and XGBoost to handle them efficiently.

Data Cleaning and Preparation Steps

1. Handling Missing Values:
   - Replaced non-informative entries ('--', 'not reported') with 'Unknown'.
   - Ensured consistency across categorical fields.
2. Age Conversion:
   - Parsed text strings into numeric years.
   - Binned into age groups to improve interpretability and mitigate overfitting due to small sample variation in raw ages.
   - Unreported ages associated to ‘0’ (first bin).
3. Encoding:
   - Applied One-Hot Encoding to categorical features (Gender, Race, Primary_Diagnosis, Age_bin).
   - Applied Binary Encoding (0/1) to mutation columns.
   - Dropped one category per encoded feature (using drop='first') to avoid multicollinearity.

4. Feature Reduction and redundancy handling:
    - Variables were removed if they had a redundancy ≥ 0.99 or a relevance ≤ 0.0001.
    - Retained variables contributing unique information about tumor characteristics.
5. Target Mapping:
    - Transformed Grade to numeric form: LGG → 0, GBM → 1.

Final Dataset Summary

After cleaning and encoding:

- Number of observations: ~1,500
- Number of features used for modelling: 41 (after encoding and feature reduction)
- Feature composition:
    - 4 demographic variables (encoded into ~10 columns)
    - 1 target column (Grade)
    - 25+ mutation columns
- Format: CSV, UTF-8 encoded
- Shape: (1500, ~41) after preprocessing

2. Example of Encoded Variables (Simplified)

Patient_ID	Age_bin_21_40	Gender_Male	Race_White	Primary_Diagnosis_Glioblastoma	IDH1	EGFR	TP53	ATRX	Grade
1	0	1	1	0	1	0	1	1	0
2	0	0	0	1	0	1	0	0	1
3	1	1	1	0	1	0	1	1	0

This table illustrates how textual and categorical variables were converted into machine-readable form.

Summary

The Glioma Grading Dataset integrates both biological and clinical perspectives, making it ideal for a classification problem.

Its variables capture multiple dimensions of patient data (demographic, diagnostic, and genomic), providing a robust foundation for training models to distinguish between LGG and GBM.

The careful preprocessing and encoding ensured that all variables were consistent, interpretable, and suitable for machine learning algorithms, while maintaining their biological meaning.