Brain Tumor Classification Project - Insights Report

📊Problem Statement

Brain tumors are life-threatening medical conditions requiring timely and accurate diagnosis. Manual diagnosis through MRI scans can be error-prone and time-consuming. The project aims to develop a machine learning pipeline that classifies brain tumor images (represented in structured label format) into four categories: Glioma, Meningioma, Pituitary, and No Tumor. This solution supports early diagnosis and assists radiologists in making faster, more accurate decisions.

📑Data Understanding & Wrangling

- **Sources:** Three datasets - train, test, and validation - in CSV format.
- **Preprocessing Steps:**
- Stripped column names to remove whitespaces.
- Added a new column called `split` to identify source.
- Merged all three datasets.
- One-hot encoded tumor labels were converted to a single `class` column.
- **Final Dataset Columns:** `filename`, `split`, and `class`

🔍Exploratory Data Analysis (EDA) & Visualizations

Chart 1: Class Distribution (Barplot)

- Showed how balanced/unbalanced the dataset is.

Chart 2: Dataset Split Counts (Pie Chart)

- Helped visualize distribution between train, test, and validation.

Chart 3: Tumor Class Percentage (Pie Chart)

- Helped understand dominant tumor categories.

Chart 4: Class vs Dataset Type (Countplot)

- Identified class imbalance across dataset splits.

Chart 5-13:

- KDE, Boxplot, Violin plot to check class distributions
- Heatmap & Pairplot (Chart 14, 15) for correlation & pattern discovery

---

🔬Hypothesis Testing

Hypothesis 1: Chi-square Test

- Tested relationship between tumor class and dataset split.
- Result: Significant relationship found.

Hypothesis 2: ANOVA

- Compared mean distribution of image counts per class.
- Result: Statistically significant differences exist.

Hypothesis 3: Z-Test for Proportions

- Checked whether 'No Tumor' proportion varies across splits.
- Result: Supported significant imbalance.

⏳Outlier Handling & Categorical Encoding

- No numeric features available to check outliers.
- Label encoding used for converting tumor classes into numerical values.

💶Feature Engineering, Scaling & Selection

- **Scaling:** StandardScaler used to normalize encoded values.
- **Dimensionality Reduction:** Not needed due to limited feature space.
- **Feature Selection:** Only encoded class used; future integration with images may need PCA.

🔓Model Development & Evaluation

Model 1: Logistic Regression

- **Accuracy:** 1.0
- **Hyperparameters Tuned:** `C`, `solver`
- **Evaluation:** Performed well due to feature simplicity.

Model 2: Support Vector Machine (SVM)

- **Accuracy:** 1.0
- **Hyperparameters Tuned:** `C`, `kernel`
- **Evaluation:** Achieved perfect results on current dataset.

Model 3: Random Forest Classifier

- **Accuracy:** 1.0

- **Hyperparameters Tuned:** `n_estimators`, `max_depth`
- **Evaluation:** Final model chosen due to robustness and interpretability.

🕐Evaluation Metrics & Business Impact

- **Precision:** Avoids false positives; important in medical scenarios.
- **Recall:** Avoids false negatives; life-critical when diagnosing tumors.
- **F1-Score:** Balanced performance metric.
- **Confusion Matrix:** Verified class-wise prediction strengths.

📈Model Explainability

- **Random Forest** chosen for feature importance interpretation.
- **Classes correctly learned:** Shown via confusion matrix.
- Future explainability can include SHAP/LIME with image features.

💡 Conclusion

The classification pipeline performs extremely well with metadata labels. Although limited in real-world application (since it doesn't yet use image data), the structure is robust, interpretable, and scalable. This pipeline can be integrated with CNN-based image features to build a powerful brain tumor diagnosis assistant. Early detection through such tools has potential to significantly improve treatment outcomes and assist overburdened radiology teams.

Prepared by: Tharun K | Project: Brain Tumor Classification | July 2025