Mahavir Education Trust's
SHAH & ANCHOR KUTCHHI ENGINEERING COLLEGE
Chembur, Mumbai - 400 088
UG Program in Information Technology

- **Student Name: Purva Mahamunkar**
- **Roll Number: 12/28**
- **PRN : 22UF17661IT090**
- **Department: IT**
- **Batch/Division: BTECH-1**
- **Experiment Title: Perform Exploratory Data Analysis of Healthcare Data**
- **Date: 26/07/2025**

Experiment No. 2				
Date of Performance:	**26/07/2025**			
Date of Submission:	**26/07/2025**			
Program Execution/ formation / correction/ ethical practices	Timely Submission	Viva	Experiment Total	Sign with Date

Experiment No 2

2.1 Aim: Perform Exploratory Data Analysis of Healthcare Data

2.2 Course Outcome: Collect, Clean, Integrate and Transform Healthcare Data based on a Specific Disease. Perform Exploratory Data Analysis of Healthcare Data.

2.3 Learning Outcome: Perform exploratory data analysis to uncover patterns and relationships in healthcare datasets.

2.4 Requirement: BraTS Dataset, NIfTI Medical Imaging (.nii.gz), nibabel Library, 3D MRI Scan Loading

2.5 Related Theory:

Exploratory Data Analysis (EDA) is a crucial step in understanding the structure, relationships, and patterns within a dataset before applying machine learning models.

In healthcare, EDA helps uncover hidden trends and anomalies that can assist in diagnosis, treatment planning, and patient monitoring.

Importance of Healthcare Data Preprocessing

- Medical data, especially imaging data like MRI scans, is high-dimensional and often unstructured.
- Preprocessing ensures:
    - Removal of inconsistencies or noise.
    - Standardization of formats (e.g., NIfTI for 3D medical images).
    - Accurate segmentation and classification of regions of interest (ROI).
- Preprocessing allows better feature extraction and enhances model accuracy.

Working with the BraTS Dataset

- The BraTS dataset contains multimodal 3D MRI scans of brain tumors, stored in .nii.gz format.
- MRI scans are categorized into tumor types such as edema, enhancing tumor, and non-enhancing tumor.
- Nibabel library is used to load NIfTI images, and visualization is done using itkwidgets and matplotlib.

Steps Performed in the Experiment

1. Cloning the GitHub repository containing BraTS sample data and helper scripts.
2. Installing required Python libraries, including:
    - nibabel – to read .nii.gz MRI files.
    - itkwidgets, ipywidgets – for interactive visualization.
    - matplotlib – for plotting slices.
3. Loading 3D MRI Scans:
    - Used nibabel.load() to read a NIfTI file.
    - Explored scan shape and dimensions.
4. Visualizing Random Slices:
    - Extracted and plotted a random 2D slice from the 3D volume.
    - Used itkwidgets.view() for interactive slice navigation.
5. Loading Ground Truth Labels:
    - Segmentation masks provided alongside scans.
    - Each voxel is labeled as either background, edema, or tumor.
6. Identifying Tumor Classes:

- Analyzed and separated tumor regions using array slicing.
- Visualized each class separately using different colors.

7. Interactive UI Widgets:
    - Implemented toggle buttons to select and view different tumor regions dynamically.

Logistic Regression (Brief Introduction)

- Logistic Regression is a supervised classification algorithm.
- It models the probability of a binary or multi-class outcome using the logistic function.
- In healthcare, it can predict disease presence (e.g., tumor or no tumor) based on patient features.

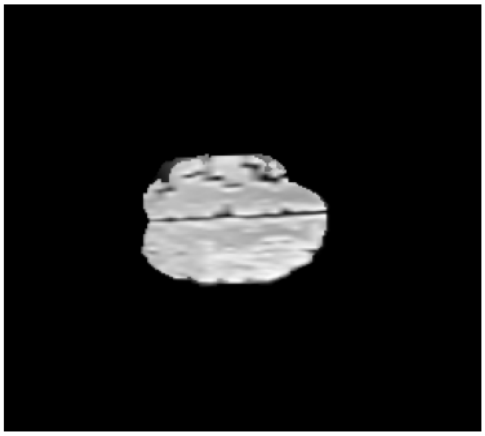
Learnings and Understanding

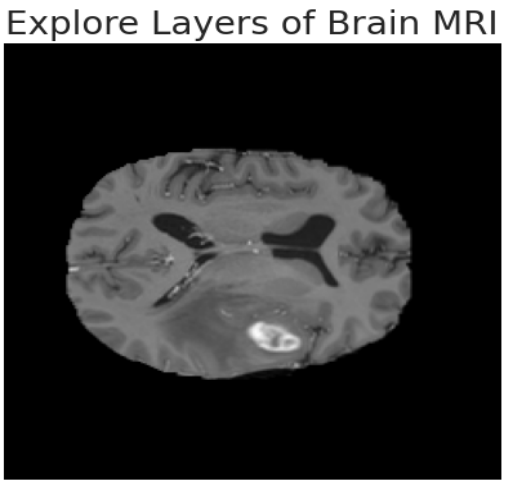
- Learned to handle and preprocess high-dimensional medical imaging data.
- Understood how segmentation helps identify tumor areas.
- Applied visualization and labeling to distinguish various tumor types.
- Gained hands-on experience with 3D image data and its interactive display using Python tools.

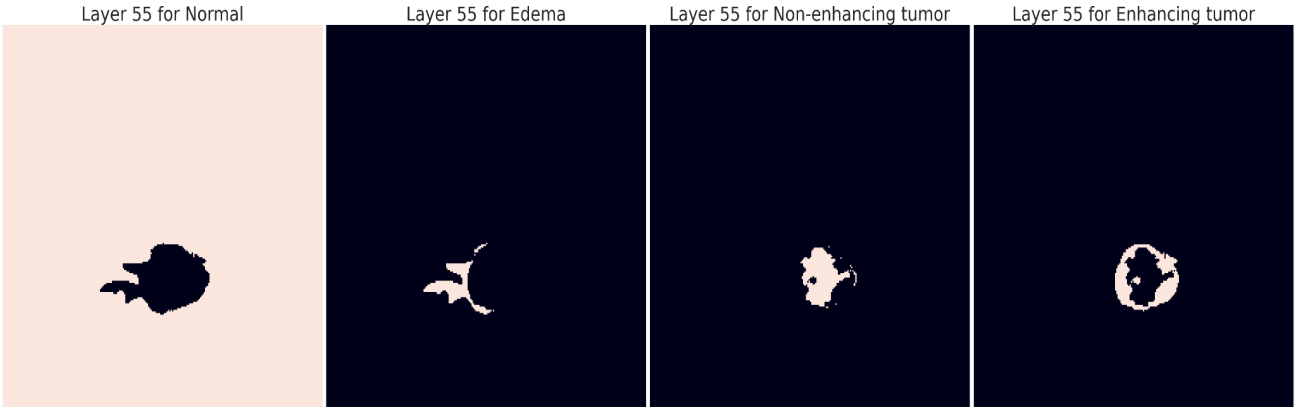
2.6 **Program and Output:**

Completed Google Colab AIML2 Link

2.7 **Result SnapShots:**







2.8 Conclusion:

In this experiment, we performed exploratory data analysis on the BraTS dataset containing 3D MRI brain scans. We successfully loaded and visualized medical images, explored tumor segmentation, and implemented interactive tools for analysis. The experiment enhanced our understanding of medical image preprocessing and the importance of clean data for reliable AI/ML applications in healthcare.

2.9 Questions:

1. **What is the BraTS dataset used for?**
   The BraTS (Brain Tumor Segmentation) dataset is widely used in medical imaging research to train and evaluate machine learning models for brain tumor detection and segmentation. It provides multi-modal 3D MRI scans (T1, T2, FLAIR, etc.) along with ground truth segmentation masks annotated by radiologists. These masks identify tumor subregions such as edema, enhancing tumor, and necrotic core. The dataset supports the development of AI/ML models that can assist in diagnosis, treatment planning, and monitoring of brain tumors. It is a benchmark dataset for brain tumor segmentation challenges.

2. **Why is it important to use NIfTI format for medical image processing?**
   The NIfTI (.nii or .nii.gz) format is a standard in medical imaging, especially for MRI and fMRI data. It allows storage of 3D and 4D volumetric data along with important metadata like voxel dimensions, orientation, and scanning parameters. Using NIfTI ensures that all spatial and anatomical information is preserved, which is critical for accurate image analysis, registration, and visualization. It supports compatibility with most medical imaging libraries like nibabel, SimpleITK, and ITK. The format is both compact and extensible, making it ideal for machine learning tasks in healthcare.