CMPS 261 – Machine Learning
Spring 2025 – Course Project

Due April 10th, 2025

In this project, you will apply machine learning techniques to real-world datasets, gaining hands-on experience in data preprocessing, model selection, and evaluation. You have the flexibility to choose from three different projects, each designed to challenge your analytical skills and enhance your understanding of key ML concepts.

1 Brain-tumor classification

Train AI to detect brain tumors with MRI/CT data for early diagnosis and better treatment planning.

1.1 About the Dataset

A brain tumor detection dataset consists of medical images from MRI or CT scans, containing information about brain tumor presence, location, and characteristics. This dataset is essential for training computer vision algorithms to automate brain tumor identification, aiding in early diagnosis and treatment planning.

1.2 Example Label Format:

```
1 0.344484 0.342723 0.221831 0.176056
```

The first number represents the classification (1 for tumor, 0 for no tumor), and the subsequent numbers define the bounding box surrounding the tumor in the image, where the second and third numbers correspond to the center coordinates (x, y) of the bounding box, and the fourth and fifth numbers represent the width and height of the bounding box, respectively.

1.3 Applications

The application of brain tumor detection using computer vision enables early diagnosis, treatment planning, and monitoring of tumor progression. By analyzing medical imaging data like MRI or CT scans, computer vision systems assist in accurately identifying brain tumors, aiding in timely medical intervention and personalized treatment strategies.

2 Molecules Toxicity Classification

Classify molecules based on their toxicity to aid in drug discovery and safety assessments.

2.1 About the Dataset

The dataset includes molecules designed for functional domains of a core protein, with some labeled as toxic and others as non-toxic. Each instance in this dataset represents a small molecule.

2.2 Applications

Toxicity classification of molecules is crucial in pharmaceutical and biomedical research. By applying machine learning to predict molecule toxicity, researchers can accelerate drug development, reduce experimental costs, and enhance safety in drug discovery.

3 Sentiment Analysis for Mental health

Develop a machine learning model capable of accurately classifying text statements into the seven mental health categories.

3.1 About the Dataset

The dataset consists of text statements representing an individual's thoughts or emotions, and their corresponding mental health status labels. The data is sourced from diverse platforms including social media posts, Reddit posts, Twitter posts, and more. Each statement is tagged with one of the following seven categories:

1. Normal
2. Depression
3. Suicidal
4. Anxiety
5. Stress
6. Bi-Polar
7. Personality Disorder

3.2 Applications

Sentiment analysis for mental health has diverse applications including early detection of depression, anxiety and suicidal tendencies from social media posts, chatbot conversations and online forums. It enhances mental health chatbots by providing supportive responses and crisis resources, aids therapists in tracking patient progress and helps universities and workplaces monitor the employees' or students' well-being.

Scoring

Your model will be evaluated on a hidden dataset, using the Accuracy metric.

Deliverables

In addition to submitting all the required files through Moodle, you are required to maintain a GitHub repository with properly documented code and include the link to that repository in your submission. You are also expected to submit a report detailing your work, including any pre-processing steps performed on the data, all experiments you performed during the training phase, and how you validated your methods. You should also mention any challenges encountered as well as any assumptions made. You will be expected to present the report orally. You are also expected to submit a saved machine learning model that can be used for inference to be evaluated for generalization on new data.

Grading

Your grade will include the GitHub repository, code, documentation, report, and presentation. In addition to your GitHub submission, add all relevant files (e.g., code, report) to one folder: `groupname.zip` (or `.rar`) and submit it to Moodle.