Machine Learning Projects_2025

Team Size:

Members: 6-7

Description:

As a team you will be implementing:

1. Linear regression & KNN as regressors on a numerical dataset.

2. Logistic regression & kmeans as classifiers on an image dataset (5 classes at maximum).

Grading:

Total: 20 marks

- **8 marks: Numerical dataset.**
- **8 marks: Image dataset.**
- **4 marks: Individual Assessment.**

Deliverables:

1) **Source Code + Datasets: Uploaded to GitHub, to be filled by the team leader later before discussion.**
2) **Project Cover Sheet:**
   It should include Faculty name, course name, team number, team members' IDs and names.
3) **Project Description Document:**
   For each model, you should specify:
   a. ***General Information on dataset:*** the name of dataset used, number of classes and their labels, the total number of samples in dataset and the size of each (in case of images), and finally the number of samples used in training, validation and testing.
   b. ***Implementation details:***
      - At feature extraction phase, how many features were extracted, their names, the dimension of resulted features.
      - Is cross-validation is used in any of implemented models? If yes, specify the number of fold and ratio of training/validation.
      - Hyperparameters used in your model, as initial learning rate, optimizer, regularization, batch size, no. of epochs, etc…

   c. Results details:
   For each model you should show all these results for your model on testing data (loss curve, accuracy, confusion matrix, ROC curve)

Datasets:

A) **numerical dataset of your selection (examples)**

1. **Healthcare**
   - **Idea:** Developing healthcare predictive model to predict the medical test results (Normal, Abnormal, or Inconclusive)
   - **Link:** *https://www.kaggle.com/datasets/prasad22/healthcare-dataset*
2. **Healthcare Insurance Expenses**
   - **Idea:** Develop predictive models for estimating healthcare expenses.
   - **Link:** *https://www.kaggle.com/datasets/arunjangir245/healthcare-insurance-expenses/data*
3. **Credit Card Fraud Detection Dataset 2023**
   - **Idea:** Develop a model to identify potentially fraudulent transactions.
   - **Link:** *https://www.kaggle.com/datasets/nelgiriyewithana/credit-card-fraud-detection-dataset-2023*
4. **Bank Loan Approval**
   - **Idea:** Develop a model to predict the load approval.
   - **Link:** *https://www.kaggle.com/datasets/vikramamin/bank-loan-approval-lr-dt-rf-and-auc*
5. **Android Malware Detection**
   - **Idea**: Develop a model to detect whether the app is malware or goodware.
   - **Link:** *https://www.kaggle.com/datasets/joebeachcapital/tuandromd*

B) **Choose image dataset from one of the following:**

1) **Cell Images for Detecting Malaria**
   a) **Link**: https://www.kaggle.com/datasets/iarunava/cell-images-for-detecting-malaria

2) **Columbia Object Image Library (COIL-100) Dataset**
   a) **Link**: http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php

3) **Flower Species Recognition**
   a) **Link:** http://www.robots.ox.ac.uk/~vgg/data/flowers/102/index.html

4) **Fruits 360 Dataset**
   a) **Link :** https://www.kaggle.com/moltean/fruits

5) **Caltech-UCSD Birds-200 2011**
   a) **Link:** http://www.vision.caltech.edu/datasets/cub_200_2011/

6) **Oxford-IIIT Pet Dataset**
   a) **Link:** https://www.robots.ox.ac.uk/~vgg/data/pets/

7) **STL-10 dataset**
   a) **Link:** https://www.kaggle.com/jessicali9530/stl10

8) **Stanford Dogs Dataset**
   a) **Link:** http://vision.stanford.edu/aditya86/ImageNetDogs/main.html

9) **Age estimation**
   a) **Link:** https://susanqq.github.io/UTKFace/

10) **Plant Pathology 2020 – FGVC7**
   a) **Link:** https://www.kaggle.com/c/plant-pathology-2020-fgvc7/data

11) **Plant Disease Classification:**
   a) **Idea:** Develop a model to classify plant diseases based on images of leaves.
   b) **Dataset:** PlantVillage Dataset

12) **Food Recognition:**
   a) **Idea:** Create a system that can identify different types of food from images.
   b) **Dataset:** Food-101

13) **Medical Image Diagnosis:**
   a) **Idea:** Develop a model to classify medical images for conditions like diabetic retinopathy or pneumonia.
   b) **Dataset:** Diabetic Retinopathy Detection

14) **Fashion Item Classification:**
   a) **Idea:** Build a model to classify fashion items, such as clothing and accessories.
   b) **Dataset:** Fashion MNIST

15) **Scene Classification:**
   a) **Idea:** Develop a model to classify scenes, such as beach, city, forest, etc., from images.
   b) **Dataset:** MIT Scene Parsing Benchmark

16) **Traffic Sign Recognition:**
   a) **Idea:** Create a model to recognize and classify traffic signs from images.
   b) **Dataset:** German Traffic Sign Recognition Benchmark

17) **Art Style Classification:**
   a) **Idea:** Develop a model to classify images based on artistic style (e.g., impressionism, cubism).
   b) **Dataset:** Painter by Numbers

18) **Facial Expression Recognition:**
   a) **Idea:** Develop a model to recognize facial expressions (happy, sad, angry, etc.) from images or video frames.
   b) **Dataset:** Facial Expression Recognition Challenge (FER2013)

19) **Tomato Detection:**

a) **Idea:** Develop a model to determine tomato Fresh or Rotten.
b) **Dataset:** *https://www.kaggle.com/datasets/nexuswho/tomato-detect*

20) Eye Diseases Classification

a) **Idea:** Develop a model to classify eye diseases (Normal, Diabetic Retinopathy, Cataract and Glaucoma retinal)
b) **Dataset:** *https://www.kaggle.com/datasets/gunavenkatdoddi/eye-diseases-classification/data*