DataByte

The Official Machine Learning and Data Science Club of NITT

FIRST-YEAR INDUCTION

TASK-1

MACHINE LEARNING

All the machine learning models are to be built from scratch without taking the help of sklearn or any similar library. Otherwise, the submission would not be considered. All the problem statements involve very basic steps to approach them, so we expect Innovation and creativity in approach, more than implementing a model.

General Instructions for Submission:

- The submission must include all the EDA performed and the results from them.
- There must be at least 3 very different models implemented and results verified (say decision trees, KNN, Logistic regression)
- Apart from the given evaluation metric, explore others and suggest why F1 or any other metric is better for the problem statement you are working on
- Comment on the code correctly and make sure they are readable, and highlight wherever you have tried something creative or used an innovative approach.
- All the labels have been given for the problem statements given below, but what do you do if you just have a bunch of numbers without labels? Try implementing that along with your task for exploring your creativity.

Problem Statements:

(Choose any one problem statement)

1. **Liver Disease Prediction (Classification):**

   The task is to develop prediction models utilising 18 clinical characteristics to forecast the stage of liver cirrhosis. Cirrhosis causes liver harm for a number of reasons, which results in scarring and liver failure. These types of applications are very crucial to use AI in healthcare.

   This is basically a classification problem where the goal is to categorise the stage of the disease so that measures can be taken according to it. Before modelling, use suitable processing techniques and present visually the inference you get from the EDA. Then proceed to develop the prediction models and compare them using evaluation metrics, and justify the results statistically to support your model's performance.

   About data:
   Train Dataset - It consists of a total of 6801 data points.
   Test Dataset - You must predict the stage of cirrhosis of 3201 data points.

   Evaluation: F1 Score

   Dataset: Liver_disease_pred

   (or)

2. **Taxi Fare Prediction (Regression):**

   Have you ever felt taxis or autos just get you fooled at times, asking for an insanely high amount, now you will have to approach this as a budding machine learning enthusiast. Your task is to delve into the world of transportation economics and build a regression-based model to predict the fare using features like the trip duration, distance travelled, etc. Problems like this are solved better using domain knowledge so feel free to take an intuitive approach to this mystery.

About data:
Train: 20967 x 8
Test: 89861, 8

Evaluation Metric: Choose the best metric to evaluate for the dataset with statistical reason

Dataset: Fare_prediction

(or)

3. **Decode AQI (Classification):**

   The Air Quality Index (AQI) is a widely-used measure of air pollution that provides information on the quality of air in a city on a daily basis. The task is to develop a model which can predict the AQI_Bucket based on features like PM2.5, CO, etc. This dataset contains air quality measurements for several cities, with observations taken at different dates and times. The variables represent different pollutants and their concentrations in the air, as well as the AQI bucket, which provides an overall measure of air quality."

About data:
Train: 495512 x 15
Test: 212363 x 14

Evaluation: F1 Score

Dataset : AQI

(or)

4. **Wine Quality Prediction (Classification):**
   Embark on a journey into the world of wine, where predicting quality takes an unexpected twist. Your mission is to develop a model that accurately forecasts wine quality using freely available datasets from the internet. But here's the catch: the dataset you'll be working with lacks clear information on the usual factors used to assess wine quality.

   As a curious data scientist, your task is to uncover hidden connections within this mysterious dataset. By harnessing the power of machine learning, you'll need to uncover insights that could hold the key to predicting wine quality. Dive deep into the unknown, examining hidden variables and patterns that might reveal the secrets of excellence. Your model must adapt and discover the unspoken features that truly influence wine quality, exploring new territory in the world of data-driven wine assessment.

Dataset: Wine Quality - UCI Machine Learning Repository

(or)

5. **Smart Botanist (Classification):**
   As a botanist, the task at hand involves developing a machine learning model to effectively classify different species of flowers. This model will utilise available features from the dataset to accurately determine the species of a given flower. The ultimate objective is to create a reliable and robust classification model that can generalise well to new samples of iris flowers. Take into consideration the effects of seasonal and climatic changes i.e environmental conditions.(Hint> Soil types, rainfall etc). Various geographic locations differ in the types of iris flowers present, the objective is to provide accurate predictions for their species(Do research into incorporating these points into your dataset while building your model).

   Now, consider that the model is able to perform well on the training data but does not perform well while testing on unseen input features of flowers. What will you do in order to prevent this? Furthermore, the model's performance will be assessed using appropriate evaluation metrics such as precision, recall, and F1 score, ensuring a comprehensive understanding of its effectiveness across various classes. Suggest a way you can automate the learning process of the model.
   Dataset: Iris - UCI Machine Learning Repository

   (or)

6. **Real Estate Price Prediction (Regression):**
   Melbourne, being the evergreen city in the beautiful country of Australia, has a variety of households for the average real estate buyer to choose from. Your objective is to correctly predict the price of a respective real estate property in Melbourne, taking into consideration the various factors such as land area, the number of bedrooms, etc. The model must utilise all factors that may affect the price of a house in Melbourne with the help of regression techniques, train and test your dataset so as to evaluate the model with regressive metrics such as (MSE) Mean Squared Error, RMSE, etc.
   Your dataset may have outliers, so do the needful and make sure it can adapt and improvise to correctly predict the price of a house for a prospective buyer. Be sure to produce realistic predictions.

   Brownie points will be given to those who can do an appropriate EDA on the dataset and derive meaningful conclusions.

Dataset: House_Pricing

Brownie points:

Devise a deep learning algorithm to improvise the accuracy scores of above-implemented models developed using machine learning techniques.

(NOTE: This task is optional. But it is highly recommended to do so. Implementing this task will increase your chance of selection into the club)

Also,

- Try implementing these algorithms from scratch without using libraries
- To further enhance your skills and exposure, try executing the clustering classification and regression algorithms in different environments. This will allow you to explore various features and libraries simultaneously.

Expectations:

- Implement it with the help of PyTorch (preferred) or Tensorflow for the task.
- Use Jupyter Notebooks as the coding environment.
- Consider setting up Jupyter in VS Code to execute the algorithms.

Guidelines:

- The read.me file of Github must include the accuracies of algorithms by adding suitable evaluation metrics scores, graphs of epochs, loss, and accuracy.
- Utilize various evaluation methods such as confusion matrix, ROC curve, or precision-recall curve to assess the performance of each approach.
- Provide insights into the strengths and limitations of each method based on the evaluation results.

Github link submission:

Include your code implementation, evaluation metric scores, graphs, and report. (NOTE: Failing to include any of these will abstain evaluation of your work)

Resources:

Setting up Google Colab Notebooks

- Google Colab Tutorial for Beginners | Get Started with Google Colab

Setting up Jupyter Notebooks

- Install Miniconda (Python) with Jupyter Notebook and Setting Up Virtual Environments on Windows 10

Setting up VS Code

- VSCode Tutorial For Beginners - Getting Started With VSCode

Python

- https://www.python.org/

Virtual Environment

- Create your own python virtual environment using the link below
- https://www.youtube.com/watch?v=ohlRbcasPAc

Machine Learning

- Machine Learning Tutorial Python | Machine Learning For Beginners - YouTube
- https://trainings.internshala.com/machine-learning-course/?tracking_source=trainings-search-tags
- Practical Machine Learning Tutorial with Python Intro p.1
- https://www.coursera.org/learn/machine-learning or refer to Andrew NG videos on youtube with the link below#1 Machine Learning Specialization [Course 1, Week 1, Lesson 1]
- https://www.kaggle.com/learn/intro-to-machine-learning for extensive notes on various concepts on the basics of ML

Deep Learning

- Go through @patloeber , his courses for deep learning techniques and machine learning from scratch.
- Deep Learning playlist overview & Machine Learning intro
- PyTorch Prerequisites - Syllabus for Neural Network Programming Course
- https://youtube.com/playlist?list=PLhhyoLH6IjfxVOdVC1P1L5z5azs0XjMsb
- Pytorch

  Learn basics of Pytorch and its implementation using the link below:
  https://pytorch.org/tutorials/index.html#

Github tutorials

- GitHub Tutorial - Beginner's Training Guide