List of Projects
DSP315/DSP615: Data Science in Practice [40 marks]

Basic Instructions:

1) The use of generative AI (ChatGPT) to write reports is strictly prohibited.
2) All members need to make tangible contributions.
3) Late submissions are not allowed.
4) Talk to mentors on daily basis to show progress (20% weightage)

Assignment Submission Guidelines:

Phase 1 (midterm):

- Explain problem statement clearly. Search for the research work already done on this topic (Literature review). Do data analysis and look for various machine learning techniques that can be applied on the dataset.
- Submission: Code of data analysis, Report, Presentation.

Phase 2 (endterm):

- Understand and apply various machine learning techniques on the dataset. Come up with new findings.
- Submission: Code containing all the results/findings (min 2000 words), Final Report, Presentation.

Marking distribution: Code: 50% (you have to explain the code, copy pasting code from internet/LLM will not help), Report: 20%, Presentation: 10%, Interaction with mentors: 20%

Report Template:

https://drive.google.com/file/d/1t982FOGodsAjp_gIxnIILEfum2y30DfQ/view?usp=drive_link

---

1. Title: Diabetes Prediction Over Telephonic Health Survey

Problem Statement: The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the Centers for Disease Control in the USA. Each year, the survey collects responses from over 400,000 US citizens on health-related risk behaviors, chronic health conditions, and the use of preventative services. The aim here is to check whether these survey datasets from the BRFSS provide accurate predictions of whether an individual has diabetes using machine
Learning.

Dataset: https://drive.google.com/drive/folders/1XMugPBZZcVlw8-DCLXGv8JCQLoX-B3l-?usp=sharing

Tasks: Identify important features related to the health survey to identify the classes of diabetes. Design a novel feature selection framework or use the existing techniques with proper justifications to identify salient features of the given data. Report the performance of different classification techniques on the training data to demonstrate that the proposed feature selection scheme is working well. You can also propose a novel classification framework for this problem. The class labels are 0 (no diabetes or only during pregnancy), 1 (prediabetes), and 2 (diabetes). Subsequently, execute the best

framework on the test data and submit the class labels in a text file. Each row of the text file will contain the class label of an instance of the test data e.g., 0 following the order of the given test data

2. Title: Credit Card Fraud Detection

Problem Statement: Given a set of real bank transactions made by users, the goal is to identify fraudulent transactions that have not been made by the users.

Dataset: https://data.world/vlad/credit-card-fraud-detection

Tasks: Classify and predict fraudulent transactions using support vector machine classifiers. Use different preprocessing techniques and feature extraction techniques to identify salient features. Show the experimental results of the method proposed (if any) and compare them with the state of the arts. Subsequently, analyze the results and report significant findings and scopes of future works. Each row of the text file will contain the class label of an instance of the test data e.g., 0 following the order of the given test data.

3. Title: Medical Insurance Cost Prediction Using Machine Learning

Problem Statement: Medical insurance costs are a significant financial concern for many individuals. Predicting these costs accurately can help both insurance providers and customers in planning and budgeting. This project aims to develop a machine learning model to predict medical insurance costs based on various factors such as age, sex, BMI, number of children, and smoking status. By leveraging linear regression, the goal is to create a model that can estimate insurance costs with high accuracy, providing valuable insights into how different features affect insurance pricing.

Dataset: https://drive.google.com/drive/folders/1XMugPBZZcVlw8-DCLXGv8JCQLoX-B3l-?usp=sharing

Tasks: Develop a machine learning model to predict medical insurance costs based on factors such as age, sex, BMI, number of children, and smoking status. It will start by exploring and preprocessing the dataset, including handling missing values and encoding categorical variables. Then split the data into training and testing sets and train a Linear Regression model to predict insurance costs. Evaluate the model's performance using metrics like Mean Absolute Error and R-squared, interpret the model's coefficients to understand feature impacts, and document their findings. Finally, document your process, including data preprocessing, model development, evaluation results, and any deployment considerations, and prepare a comprehensive project report. You can also propose a novel framework for this problem.

4. Title: Predict Wine Quality Using Machine Learning with Python

Problem Statement: The quality of wine is a critical factor in the wine industry, influencing consumer preference and market value. Traditional methods of assessing wine quality are often subjective and time-consuming. By leveraging machine learning techniques, we aim to develop a predictive model that can objectively estimate the quality of wine based on its physicochemical properties. This project focuses on utilizing machine learning algorithms to analyze a dataset of wine characteristics and predict wine quality ratings, providing an automated and efficient tool for quality assessment.

Dataset: https://drive.google.com/drive/folders/1XMugPBZZcVlw8-DCLXGv8JCQLoX-B3l-?usp=sharing

Tasks: Explore and preprocess a dataset containing the physicochemical properties of wine to predict its quality rating. Then begin by cleaning and normalizing the data, selecting relevant features for analysis. Next, build and train various machine learning models, such as Decision Trees or Random Forests, and evaluate their performance using appropriate metrics. The task also involves interpreting the results to understand the impact of different features on wine quality and optionally deploying the model for practical use. Finally, document your process, including data preprocessing, model development, evaluation results, and any deployment considerations, and prepare a comprehensive project report. You can also propose a novel framework for this problem.

5. Title: Iris Flower Classification

Problem Statement: Given a set of features corresponding to Iris flowers. Classify the features using standard classification techniques to categorize the flowers into 3 categories, setosa, versicolor, and Virginia, so that you can achieve at least 80% accuracy.

Dataset: https://archive.ics.uci.edu/dataset/53/iris

Tasks: Classify and predict the type of Iris flower based on its features. Accurately identify the species of iris flowers based on their measurements. Automate the classification process and develop an accurate practical method for identifying iris species.

6. Title: Email Spam Detection

Problem Statement: Given a dataset containing labels about emails, implement a machine learning model for accurately detecting spam emails. Use the Pandas library to read and preprocess the dataset.
Dataset: https://drive.google.com/drive/folders/1XMugPBZZcVlw8-DCLXGv8JCQLoX-B3l-?usp=sharing
Tasks: Read the dataset using the Pandas library, and perform some preprocessing tasks if required. Subsequently, use the appropriate machine learning model to detect spam emails. Use the standard training, testing, and validation framework and use sci-kit learn functions. Properly report the accuracy, precision, recall, and F1 score measures.

7. Title: Rainfall in Indian States

Problem Statement: The dataset contains the monthly rainfall values in various Indian states. Plot the rainfall values in MP from 2020. Detect whether the rainfall after 2020 in MP and Jharkhand has any correlation. Visualize these two variables using the scatter plot, and quantify the statistical correlation between these two variables.
Dataset: https://www.kaggle.com/datasets/rajanand/rainfall-in-india
Tasks: Read the rainfall dataset using the pandas library and appropriately store the values in data frames. Only consider the data after Jan 2020. Plot monthly rainfall values in the MP from 2020 using an appropriate graph/plot using the Matplotlib library. Use appropriate statistical measures to check if there is any correlation between the rainfall in MP and Chattisgarh. Use a scatterplot to visualize these aforementioned variables.

8. Title: Sales Prediction over time

Problem Statement: The sales dataset contains several features. Display how the sales and discounts change with time. Compare the standard machine learning models for time series prediction tasks. Evaluate the methods and display the results using a plot and table.
Dataset: https://www.kaggle.com/datasets/farshadtofighi/sales-prediction-dataset
Tasks: Read and store the features in the sales dataset using the Pandas data frame. Use time-series visualization to display the sales and discounts over time. Predict future sales and discounts using machine learning methods. Display the predicted values using time series plots. Also, compare the performance of these time-series methods.

19. Title: Marketing Campaign Analysis

Problem Statement: A response model can provide a significant boost to the efficiency of a marketing campaign by increasing responses or reducing expenses. The objective is to predict who will respond to an offer for a product or service
Dataset: https://data.world/data-society/bank-marketing-data
Tasks: Determine relevant marketing campaign data features to identify the campaign's success or failure. Design a novel feature selection framework or use the existing techniques with proper justifications to identify essential features of the given data. Report the performance of different classification techniques on the training data to demonstrate that the proposed feature selection scheme is working well. Find the missing data and use feature engineering techniques to create new features and columns, such as age, total number of campaigns, average spend, total spend, and non-relevant columns. Use exploratory data analysis to find the distribution of income, age, marital status, and educational level. Plot graphs for the distribution of the number of children and teenagers in a household, total campaigns accepted, average spend per purchase, spending distribution by marital status, spending distribution by education level, spending distribution by is_parent, distribution of online purchase ratio, distribution of the number of web purchases, distribution of the number of catalog purchases, and then scatter plots.

10. Title: Dimensionality Reduction of University Dataset

Problem Statement: A university dataset contains information related to several aspects such as application, acceptance, student information, etc. Ignoring the type of university (Private University) reduces the other 17 features using PCA and visualizes the results.
Dataset: https://www.kaggle.com/code/karthickaravindan/k-means-clustering-project
Tasks: Read the university information dataset given in the link, and ignore the information related to the type of university. The other features may contain some correlation. Reduce the number of features using PCA so that minimum information is lost. Identify the proper number of PCs, and plot the transformed data obtained after PCA using a scatterplot. Plot PC1 Vs PC2, PC3 Vs PC4, etc.

11. Title: University Categorization

Problem Statement: A university dataset contains information related to several aspects such as application, acceptance, student information, etc. Use linear discriminant analysis to reduce the dimensionality of the data so the type of university is segregated.
Dataset: https://www.kaggle.com/code/karthickaravindan/k-means-clustering-project
Tasks: Read the university information dataset given in the link, and ignore the information related to the type of university. Consider the type of university as the classes (Private and government universities are the two classes). Use LDA for dimensionality reduction. Also, ensure that the classes are well separated. Display the reduced dimensional representation obtained by the LDA method. Report the classification performance using standard measures.

12. Title: Flat Price Estimation

Problem Statement: The objective of this project is to predict the prices (in lakhs) of flats in various cities of India based on different factors.
Dataset: https://drive.google.com/drive/folders/1XMugPBZZcVlw8-DCLXGv8JCQLoX-B3l-?usp=sharing
Tasks: Identify relevant features of the given training data to estimate the flat prices (in lakhs) in India. Design a suitable feature selection framework or use the existing techniques with proper justifications to identify salient features of the given data. Report the performance of different regression techniques on the training data to demonstrate that the proposed feature selection scheme is working well. Subsequently, execute the best framework on the test data and submit the estimated flat price in a text file. Each row of the text file will contain the price of an instance of the test data e.g., 22.5 following the order of the given test data.

13. Title: Face Recognition from Features

Problem Statement: The Yale Face Database contains 165 grayscale images in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink. Extract standard features from the images and identify the persons using common machine learning methods.
Dataset: http://cvc.cs.yale.edu/cvc/projects/yalefaces/yalefaces.html
Tasks: Explore the dataset and extract hand-crafted features such as local binary pattern, Gabor filter, Laplacian of Gaussian (LoG), and Gray-Level Co-occurrence Matrix (GLCM). Combine these features and apply standard machine learning models to the feature set. The classifier aims to detect persons accurately. Report the results in detail.

14. Title: Fake Face Detection

Problem Statement: The dataset contains both real and fake face images. Use standard machine learning models to distinguish these two types of faces. Report and compare the performance of these methods.
Dataset: https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection
Tasks: Read and save the images from the dataset into training, testing, and validation sets. Use standard machine learning dataset split. Categorize real and fake images using machine learning algorithms. Compare the accuracy, precision, recall, F1 score etc. of these models.

15. Title: Understanding Real and Fake Faces

Problem Statement: The dataset contains both real and fake face images. Since the individual images contain many pixels, the dataset has high dimensionality. Use PCA and LDA to reduce the dimensionality of the dataset. Visualize the plots obtained by PCA and LDA, and illustrate how PCA differs from LDA.
Dataset: https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection
Tasks: Read and save the images from the dataset into training, testing, and validation sets. Use standard dimensionality reduction techniques, namely PCA and LDA to reduce the image dimension. Visualize the reduced data using a scatterplot. Display the PCA plots and LDA plots.

16. Title: Classification of Satellite Images Using Deep Learning Models

Problem Statement: The objective of this project is to classify satellite images.
Dataset: https://www.kaggle.com/datasets/mahmoudreda55/satellite-image-classification
Tasks: The objective of this project is to classify satellite images into four categories: cloudy, desert, green area, and water. To achieve this, use machine learning models such as Random Forest, Decision Tree classifier. The task involves developing these models to accurately classify the images while also employing different preprocessing techniques and feature extraction methods to enhance model performance. Experimental results will be analyzed and compared with state-of-the-art methods to identify the most effective models and features. The findings will be reported, along with insights into potential directions for future research. Each classified image will be labeled according to the specified categories to ensure accurate classification.

17. Title: Indian Crime Analysis

Problem Statement: The focus of this project is to analyze crime patterns in India.
Dataset: https://drive.google.com/drive/folders/1XMugPBZZcVlw8-DCLXGv8JCQLoX-B3l-?usp=sharing
Tasks: Analysis will involve identifying trends, hotspots, and correlations within the crime data. Various statistical and machine-learning techniques will be employed to uncover insights and make predictions about crime rates and patterns. The findings from this analysis will be used to provide actionable recommendations and to identify areas for further investigation.

18. Title: Electricity Consumption Using Time Series Analysis

Problem Statement: A statistical technique called time series analysis is used to forecast future events by analyzing historical data over a specified period. It is made up of data in an organized sequence with equal spacing between each piece. Let's look at an example better to grasp the time series data and the analysis. Take airline passenger statistics as an example. It contains the number of passengers over a specified duration. A time series is a collection of observations made at specific time intervals, usually equal ones. We can forecast future values using the series' analysis and past observed values. Time and the variable we wish to forecast are the only two variables in a time series.
Dataset: https://drive.google.com/drive/folders/1XMugPBZZcVlw8-DCLXGv8JCQLoX-B3l-?usp=sharing
Tasks: Build a model to forecast the electricity power consumption value. The data is classified by date, time, and value of consumption. The goal is to predict electricity consumption for the next 6 years i.e. till 2024.

19. Title: Marketing Campaign Analysis

Problem Statement: A response model can provide a significant boost to the efficiency of a marketing campaign by increasing responses or reducing expenses. The objective is to predict who will respond to an offer for a product or service
Dataset: https://data.world/data-society/bank-marketing-data
Tasks: Determine relevant marketing campaign data features to identify the campaign's success or failure. Design a novel feature selection framework or use the existing techniques with proper justifications to identify essential features of the given data. Report the performance of different classification techniques on the training data to demonstrate that the proposed feature selection scheme is working well. Find the missing data and use feature engineering techniques to create new features and columns, such as age, total number of campaigns, average spend, total spend, and non-relevant columns.

Use exploratory data analysis to find the distribution of income, age, marital status, and educational level. Plot graphs for the distribution of the number of children and teenagers in a household, total campaigns accepted, average spend per purchase, spending distribution by marital status, spending distribution by education level, spending distribution by is_parent, distribution of online purchase ratio, distribution of the number of web purchases, distribution of the number of catalog purchases, and then scatter plots.

20. Title: Fake Face Detection

Problem Statement: The dataset contains both real and fake face images. Use standard machine learning models to distinguish these two types of faces. Report and compare the performance of these methods.

Dataset: https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection

Tasks: Read and save the images from the dataset into training, testing, and validation sets. Use standard machine learning dataset split. Categorize real and fake images using machine learning algorithms. Compare the accuracy, precision, recall, F1 score etc. of these models.
