PRML Minor Projects

Minimum expectations:

- ➢ Implement an end-to-end machine learning pipeline for the task given in the project
- ➢ Use three to four concepts from the course (PCA, LDA, KMeans clustering, KNN,ICA, SVD, Hierarchical clustering)
- ➢ Performance evaluation and analysis of the entire pipeline/algorithm and comparison of multiple concepts implemented (as part of the previous point)
- ➢ Three-four page single column (single spacing) brief report of your ideas, experiments and results
- ➢ Codes with proper documentation

Note: *You can select any other dataset of your choice apart from the mentioned one for implementing the projects*

Project 1:

This dataset contains videos from 4 different YouTubers and all the comments made on those videos. The primary objective of this dataset is to cluster the comments to identify a cluster that contains all the spam comments and fix the issue once and for all.

Dataset : Youtube_Comment_Classification

Project 2:

Credit risk is associated with the possibility of a client failing to meet contractual obligations, such as mortgages, credit card debts, and other types of loans. The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions. These datasets are hard to handle. You have to predict whether, given the details about the credit card, it is real or fake.

Dataset: Credit Card

Project 3:

Healthcare fraud is considered a challenge for many societies. Health care funding that could be spent on medicine, care for the elderly, or emergency room visits is instead lost to fraudulent activities by materialistic practitioners or patients. With rising healthcare costs, healthcare fraud is a major contributor to these increasing healthcare costs. Your task is to find the anomalies in the data using unsupervised learning techniques.

Dataset : Anomaly Detection

Project 4:

A company that sells some of the product, and you want to know how well the selling performance of the product. You have the data that we can analyze, but what kind of analysis can we do? Well, we can segment customers based on their buying behavior on the market. Your task is to classify the data into the possible types of customers which the retailer can encounter.

Dataset : Retail

Project 5:

HELP International has been able to raise around $ 10 million. Now the CEO of the NGO needs to decide how to use this money strategically and effectively. So, the CEO has to make a decision to choose the countries that are in the direst need of aid. Hence, your Job as a Data scientist is to categorize the countries using some socio-economic and health factors that determine the overall development of the country. Your objective is to categorize the countries using socio-economic and health factors that determine the overall development of the country.

Dataset : Country_Data

Project 6:

In this project, you are given a dataset where each row belongs to a particular cluster. Your job is to predict the cluster each row belongs to. You are not given any training data, and you are not told how many clusters are found in the ground truth labels.

Dataset : Ensemble_Clustering

Bonus : Students choosing a relatively difficult project will have an upper edge.