PRML Major Projects

Minimum expectations:

- ➢ Implement an end-to-end machine learning pipeline for the task given in the project
- ➢ Use the concepts from the course (Supervised and unsupervised learning techniques)
- ➢ 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:

COVID Detection using X-ray Images

The Coronavirus Disease 2019 (COVID-19) has brought a worldwide threat to the living society. One of the areas where machine learning can help is detecting the COVID-19 cases from chest X-ray images.

The task is a simple classification problem where given an input chest X-ray image, the machine learning-based model must detect whether the subject of study has been infected or not.

Dataset Link: COVID-19

Project 2:

The purpose of a recommendation system basically is to search for content that would be interesting to an individual. Recommendation systems are Artificial Intelligence based algorithms that skim through all possible options and create a customized list of items that are interesting and relevant to an individual. These results are based on their profile, search/browsing history, what other people with similar traits/demographics are watching, and how likely you are to watch those movies.

Your aim will be to recommend similar movies if a type of movie is given.

Dataset:Movie Recommendation System

Project 3:

Brain Stroke Prediction

A stroke is a medical condition caused by poor blood flow to the brain, leading to cell death and the impairment of brain function. There are two main types of stroke: ischemic, due to lack of blood flow, and hemorrhagic, due to bleeding. The symptoms of a stroke may include an inability to move or feel on one side of the body, problems understanding or speaking, dizziness, or loss of vision to one side. The main risk factor for stroke is high blood pressure, but other risk factors include high cholesterol, smoking, obesity, diabetes, and previous TIAs or heart conditions. Prevention includes reducing risk factors and taking medication such as aspirin or statins. Emergency care is required for stroke or TIA, and treatment to attempt recovery of lost function is called stroke rehabilitation.
Your task will be classifying heart disease.

Dataset Link: Brain Stroke Dataset

Project 4:

Histopathologic Cancer Detection

In this dataset, you are provided with a large number of small pathology images to classify. Files are named with an image id. The train_labels.csv file provides the ground truth for the images in the train folder. You are predicting the labels for the images in the test folder. A positive label means that there is at least one pixel of tumor tissue in the center 32x32px area of a patch. Tumor tissue in the outer region of the patch does not influence the label. This outer region is provided to enable fully-convolutional models that do not use zero-padding, to ensure consistent behavior when applied to a whole-slide image.
Dataset Link: Histopathologic Cancer Detection

Project 5:

Speech Emotion Recognition

Speech Emotion Recognition, abbreviated as SER, is the act of attempting to recognize human emotion and affective states from speech. This is capitalizing on the fact that voice often reflects underlying emotion through tone and pitch. This is also the phenomenon that animals like dogs and horses employ to be able to understand human emotion.
Dataset Link: speech-emotion-recognition

Project 6:

Accident Detection

Accidents can be used for numerous applications such as real-time car accident prediction, studying car accidents hotspot locations, casualty analysis and extracting cause and effect rules to predict car accidents, and studying the impact of precipitation or other environmental stimuli on accident occurrence.

You can perform a detailed EDA for the dataset and then perform predictions for the severity and the location.
Dataset Link: Accidents

Project 7:

Fruit and Vegetable Recognition
The idea was to build an application which recognizes the food item(s) from the captured photo and gives its user different recipes that can be made using the food item(s).
Dataset Link: Fruits and Vegetables Image Recognition Dataset

Project 8:

Football Club Logo Detection
Your aim is to detect the Club logos.
Dataset Link: Football Club Logos

Project 9:

Detecting Parkinson's Disease
This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds to one of 195 voice recordings from these individuals ("name" column).
The main aim of the data is to discriminate healthy people from those with PD, according to the "status" column which is set to 0 for healthy and 1 for PD.
Dataset Link: Parkinson's Disease

Project 10:

Metro Data
The Indian Metro Data Project dataset is a collection of data about the metro systems in various cities in India. The project aims to provide insights and analysis of the metro systems in India, including ridership data, revenue data, and information about the metro stations and lines. The dataset includes information about the Delhi Metro, Mumbai Metro, Bangalore Metro, Chennai Metro, and Kolkata Metro, among others.
The data can be used to analyze the performance of the metro systems in different cities, compare the ridership and revenue of different lines, and explore trends in metro usage over time.
You can predict the traffic volume using different ML models and find out which is performing the best.
Dataset Link: Metro Data

Project 11:

Diabetes is one of the fastest growing chronic life-threatening diseases that have already affected 422 million people worldwide according to the report of World Health Organization (WHO), in 2018. Due to the presence of a relatively long asymptomatic phase, early detection of diabetes is always desired for a clinically meaningful outcome. Around 50% of all people suffering from diabetes are undiagnosed because of its long-term asymptomatic phase.

This dataset contains 520 observations with 17 characteristics, collected using direct questionnaires and diagnosis results from the patients in the Sylhet Diabetes Hospital in Sylhet, Bangladesh.

Dataset Link: Classification of Diabetes

Project 12:

Text Classification on Emails

If you are working, then you are bound to face the problem of reading all the emails that are cluttered in your inbox. Some may be relevant and some may just try to loot you. Now our client is an editor for a major newspaper who is fed up of reading the emails that his/her journalists send and segregating them in their categories.

Your aim will be to process the data, store it properly and apply different models to predict the category of an email. You should also show testing results on some random email taken by you which is not in the dataset and predict its result based on the models you have taken.

Try using Decision Tree, KNeighbors, XGBoost along with Ensemble Methods as well and compare the performances.

Dataset Link: Text Classification on Emails

Project 13:

Toxic Comment Classification Challenge

You are provided with a large number of Wikipedia comments which have been labeled by human raters for toxic behavior. The types of toxicity are toxic, severe_toxic, obscene, threat, insult and identity_hate. You must create a model which predicts a probability of each type of toxicity for each comment.

Dataset Details: Toxic Comment Classification

Project 14:

Customer Support Classification

You have been given a dataset that includes demographic information and web session records for a group of users from the USA, along with summary statistics. Your task is to predict which country a new user's first booking destination will be, out of a list of 12 possible outcomes including 'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL', 'DE', 'AU', 'NDF' (no destination found), and 'other'. It's important to note that 'NDF' indicates that the user did not book a destination, while 'other' means they did book, but the destination country is not included in the list.

The dataset is split into training and test sets based on dates. For the test set, you will be predicting the first booking destination for all new users who have their first activity after July 1st, 2014. Keep in mind that the sessions dataset only dates back to January 1st, 2014, while the user's dataset goes back to 2010.

Dataset: User Bookings

Project 15:

Hate Speech Detection

The objective of this task is to build a machine learning model that can accurately detect hate speech. Hate speech is defined as any tweet with a racist or sexist sentiment associated with it. Given a dataset of tweets and their corresponding labels (0 for non-hate speech, 1 for hate speech), the task is to predict the labels for a test dataset of tweets.

The goal of this project is to develop a classifier that can accurately distinguish between tweets containing hate speech and those that do not. This model can be used to report tweets that could be harmful or offensive, making the online community safer and more welcoming for everyone.

To solve this problem, we will use machine learning techniques. Metrics like accuracy, precision, recall, and F1 score will be used to judge how well the model works.

Dataset: Sentiment Analysis

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