2-Day Project

You are given 3 questions in this area. Work on 1 question of your choice.

1. Quality Assessment System (Fall 2023)

Problem Description. In a facility, quality assessment using machine learning can be a critical task to ensure the efficiency, safety, and reliability of operations. Here's an example of how machine learning can be applied to quality assessment in a facility: Imagine a manufacturing facility that produces electronic components, and we want to ensure that the manufactured products meet the highest quality standards. Therefore, we aim to use machine learning for quality assessment.

In this project, our objective is to construct a quality assessment system based on machine learning algorithms capable of measuring the quality of produced electronic components. We assume the system gathers observations, each containing up to 3,000 feature values, from manufacturing operations on individual products. Based on this setting, we have a dataset collected from various products manufactured in the facility with multiple production lines. Each product has a label (class) y and a value $\tilde{y}$ indicating the quality of it ($y = 0$ indicates below the standard, $y = 1$ indicates meeting the standard, and $y = 2$ indicates above the standard). This information can be used to train a model in a supervised manner.

By leveraging this dataset, we aim to build a reliable quality assessment system that can accurately measure the product quality to monitor a manufacturing facility.

- Please refer to the **dataset** in the link below:
  `https://tinyurl.com/45r6xtze`
  - `Product_Code`: product type
  - `X_i`: i^{th} feature value
  - `Line`: production line
  - `Y_Class`: product quality class
  - `Y_Quality`: product quality value

(a) **Data Preprocess.** Describe how you preprocess the data into a format to build your model and how you split the dataset to properly build your model.

(b) **Exploratory Data Analysis.** Explore the dataset with descriptive statistics and visualization to understand the problem.

(c) **Learning methods.** Build *multiple models* to achieve the goal explained above using available data and properly compare them. Describe your learning algorithm (or statistical method) to build your models, and explain how they are implemented.

(d) **Evaluation Metric.** Define your evaluation metric, and explain why it is appropriate for the problem.

(e) **Result and Discussion.** Present your work at the oral exam, including (but not limited to) problem setting, hypotheses, how you preprocess data, description and justification of your learning algorithm, prediction results, interesting observations or insights learned, and potential future work.

2. Assessing ESG Risk Ratings and Sectoral Variability Among S&P 500 Companies

Problem Description. Environmental, Social, and Governance (ESG) considerations have gained prominence in investment decisions and policy-making. Stakeholders increasingly evaluate companies based on their ESG performance, and the metrics can influence the overall valuation of companies in the market. The S&P 500 ESG Risk Ratings dataset provides a comprehensive overview of the ESG profiles for companies listed on the S&P 500 index.

Please refer to the **dataset** in the link below:

`https://tinyurl.com/sp500esg`

The dataset includes the following fields:

- `Symbol`: Stock symbol
- `Name`: Company name
- `Address`: Headquarters address
- `Sector`: Economic sector
- `Industry`: Specific industry
- `Full Time Employees`: Count of full-time employees
- `Description`: Company description
- `Total ESG Risk Score`: Aggregate ESG risk score
- `Environment Risk Score`: Environmental risk score
- `Governance Risk Score`: Governance risk score
- `Social Risk Score`: Social risk score
- `Controversy Level`: Level of ESG-related controversy
- `Controversy Score`: Numerical representation of controversy level
- `ESG Risk Percentile`: ESG risk percentile rank
- `ESG Risk Level`: Categorical ESG risk level

To complete these tasks, proceed with the following steps:

(a) **Descriptive Analysis.** Summarize the datasets with descriptive statistics, focusing on sectoral and industry-specific trends in ESG risk scores.

(b) **Predictive Modeling for ESG Risk Percentile.** Utilize appropriate machine learning or statistical models to forecast a company's ESG Risk Percentile. Justify the choice of algorithms and evaluate model performance.

(c) **Employee-ESG Relationship.** Employ correlation and regression methods to investigate the influence of full-time employee count on ESG risk scores.

(d) **Controversy Impact Analysis.** Examine how the Controversy Level and Controversy Score variables influence the Total ESG Risk Score, using suitable statistical methods.

(e) **Cluster Analysis.** Utilize clustering techniques to segment companies into different groups based on their ESG profiles. Discuss the characteristics that define each cluster.

(f) **Text Mining.** Apply text mining techniques (e.g., term frequency-inverse document frequency; feel free to use more advanced methods) on the `Description` field to improve and enhance your results and findings.

(g) **Results and Discussion.** During the oral examination, elucidate your problem setting, research methodology, data preprocessing steps, choice of algorithms, key findings, and potential avenues for future research.

3. Crime Analytics

Problem Description. Suppose you have been selected as the leader of the crime analytics team of your city. You are given the crime history of the city for about 14 years. These data contain 530K incidents of crime, along with the information about the crime type, date, and location for each incident. Your task is to analyze the crime patterns of the city over years and to predict crime types missing in some incidents.

In the crime history files (`crime-train.csv` and `crime-test.csv`), rows are crime incidents and columns have the following information:

- **TYPE:** Type of crime (e.g., Break and Enter Commercial, Homicide, Mischief)
- **YEAR, MONTH, DAY, HOUR, MINUTE:** Year, month, day, hour, and minute of the incident.
- **HUNDRED_BLOCK:** Generalized location of the incident (e.g., 40XX W 19TH AVE).
- **NEIGHBOURHOOD:** Neighbourhood of the incident (e.g., Moonbeam Glade, Elfwood Nook).
- **Latitude, Longitude:** Latitude and longitude of the incident.

To enrich your analysis and prediction, you are also given the city's census data (census.csv). This file includes statistics about each neighbourhood of the city, such as ages, marital statuses, numbers of family members, and languages.

Files. You can find the above files in `https://www.dropbox.com/scl/fo/3rfp7csi2nfqy1xcicwye/h?rlkey=svwh591479efuvvm3iaathj7o&dl=0`

Tasks. Complete the following tasks.

(a) **Exploratory Data Analysis.** Explore the data with descriptive statistics and visualization. Identify interesting and insightful patterns of crime. Include your findings about how certain crime types are associated with the time of the day, time of the year, location, neighbourhood, and characteristics of a neighbourhood (you do not have to cover all crime types or all neighbourhoods).

(b) **Prediction Model for Crime Types:** Build a classification model that takes information about an incident and predicts its crime type (i.e., the TYPE column in crime-train/test.csv). As input, you can use any information available in crime-train/test.csv (except the TYPE column) and census.csv. Use crime-train.csv for training and crime-test.csv for evaluation.

You should answer the following questions.

- What information is used for model input? Justify the choice.
- What model architecture or learning algorithm did you use? Justify the choice.
- How did you split the data in crime-train.csv into training and validation? How did you use the validation data?

(c) **Evaluation Metrics:** Choose or define evaluation metrics, and explain why they are appropriate for this problem.

(d) **Result and Error Analysis:** Analyze the result of the model and error cases. Discuss interesting findings. Your analysis should include:

- What is the overall accuracy of the model in terms of the metrics you chose?
- What crime types does the model predict well? What crime types does the model predict poorly?
- Do you have any interesting or insightful findings?
- Do you think this model can be used in practice? Justify why. How can you improve the model?

Present your work at the oral exam. Your presentation should explain all the above steps and associated questions.