Sample Documents -
https://drive.google.com/drive/folders/1CWK5gJHv4XLPiVFupwxV0Bu0K9y0BNIH?usp=sharing

** Attempt any 3 Questions**

** Avoid using Chat GPT to generate the code**

Please create sample datasets wherever possible for Q1 & 2 using ChatGPT

Q 1 Problem Statement: You are working on a binary classification problem to predict worker safety incidents in construction sites. The dataset needs to be more balanced with fewer incidents. Design and implement a custom loss function to handle this imbalance.

Dataset: A tabular dataset with features such as HoursWorked, SafetyTrainingHours, EquipmentUsed, WeatherConditions, and a binary target variable IncidentOccurred (0 for no incident, 1 for incident).

Tasks:

1. Load and preprocess the dataset.
2. Implement a custom loss function (e.g., focal loss) to give more importance to the minority class.
3. Train a neural network classifier using your custom loss function.
4. Evaluate the performance using appropriate metrics (e.g., precision, recall, F1-score).

Dataset Example:

```
HoursWorked,SafetyTrainingHours, EquipmentUsed,WeatherConditions,IncidentOccurred
8,10,Bulldozer,Clear,0
```

Q2

Problem Statement: You are working on predicting project delays in construction projects. The dataset has a large number of features, some of which are redundant. Use mutual information to select the top 10 most informative features.

Dataset: A tabular dataset with features such as ProjectSize, Budget, NumberOfWorkers, WeatherConditions, MaterialAvailability, TargetCompletionDate, and a binary target variable Delayed (0 for on-time, 1 for delayed).

Tasks:

1. Load the dataset and preprocess it (handle missing values, normalize, etc.).
2. Compute the mutual information between each feature and the target variable.
3. Select the top 10 features based on mutual information.
4. Train a simple classifier (e.g., logistic regression) using only the selected features and report the performance.

Sample

```
ProjectSize,Budget,NumberOfWorkers,WeatherConditions,MaterialAvailability,TargetCompletionDate,Delayed

500000,2000000,150,Rainy,High,2023-12-31,0
```

...

Q3

You are given a set of construction contract PDF documents. Each contract contains important information such as project name, start date, end date, parties involved, and contract value. Your task is to extract these key pieces of information and store them in a structured format.

Dataset: A set of PDF files containing construction contracts.

Tasks:

1. Load and preprocess the PDF documents.
2. Implement an algorithm to extract key information (project name, start date, end date, parties involved, contract value) from each contract.
3. Store the extracted information in a structured format (e.g., a CSV file or a database).

Sample Contract documents -

https://drive.google.com/drive/folders/1CWK5gJHv4XLPiVFupwxV0Bu0K9y0BNIH?usp=sharing

Q4

Problem Statement: You are given a set of PDF invoices from various construction projects. Each invoice contains line items with descriptions, quantities, unit prices, and total amounts. Your task is to extract these line items and store them in a structured format.

Dataset: A set of PDF files containing construction invoices.

Tasks:

1. Load and preprocess the PDF documents.
2. Implement an algorithm to extract line items (description, quantity, unit price, total amount) from each invoice.
3. Store the extracted information in a structured format (e.g., a CSV file or a database).

** Establish a few ways to detect fraud in the invoices - share in formation on steps, on how wll you do it

Sample Invoice -

https://drive.google.com/drive/folders/1CWK5gJHv4XLPiVFupwxV0Bu0K9y0BNIH?usp=sharing