Project Title: J P Morgan classification for legal Documents

Problem Statement: “Automate the classification of various legal documents.”

Learning Outcomes:

- Convert a business problem into an analytical problem.
- Break the process down using the CRISP-DM methodology.

First of all, before breaking the process down using CRISP-DM (Cross-Industry Standard Process for Data Mining), we should know the six phases of CRISP-DM**, i.e., Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, & Deployment.**

Now we start breaking down the problem into the six phases of CRISP-DM.

1. Business Understanding

This is the initial phase of **CRISP-DM**, where the focus is on the project goals, objectives, and requirements. In this case, JP Morgan needs to automate the classification process of various legal documents to ensure more time for review and reduce errors compared to manual review, as manual review would take 360,000 hours and would have higher chances of errors.

Project Objectives:

- To reduce the amount of time spent on the analysis and classification of legal documents.
- To minimize the errors in the loan agreement service.
- To improve efficiency.
- To deal with a new algorithm using machine learning to interpret new regulations.

Project Goals:

- The system should maintain high accuracy and efficiency to reduce the possibility of errors.
- Design a system that can handle these various legal documents for classification.

2. Data Understanding:

This is one of the critical phases of CRISP-DM, where we perform various data-related activities, which include data collection, data description, data exploration, and verifying the data quality.

Data Collection:
According to the goals and objectives identified, collect all the legal documents that are required for classification and state classification of legal documents and collect samples of each type.

Data Description:
Familiarize yourself with the data characteristics to have an understanding of the analysis and understand the metadata for each document for more adequate results.

Data Exploration: Organize the collected data to interpret it, identify the missing values, and perform the first level of data analysis for identifying the general distribution of various types of documents.

Data Quality: Check data quality and evaluate inconsistent and duplicate data.

3. Data Preparation

In this phase, the model is dedicated to cleaning and transforming raw data into a suitable format for modeling, which is our next stage of CRISP-DM.

Data Cleaning and Transformation: In this, the system removes duplicate data copies, preprocesses the data, and handles the missing values properly. Cleaning of the text data in a preprocessing step involves transforming the raw data to a standard format.

Data Integration and Data Selection: Transform this huge amount of data into a single set that can be used for further modeling and select the transformed data of a suitable format for modeling.

4. Modeling

In this phase, we have to create a model that would be able to categorize different legal documents. The model name is COIN (Contract Intelligence).

Model Selection: Select an appropriate machine learning algorithm for text classification that can be applied, and we can apply Natural Language Processing (NLP) and deep learning models.

Model Training: After selecting the appropriate model, we need to train that model. After training, we need to build the model and perform certain tests on that model so that the model's accuracy will be increased and also the performance of the model will be enhanced.

5. Evaluation

In this phase, the model's performance is thoroughly evaluated, all the processes that have been carried out till now are reviewed, and it is determined what will be the next steps.

The model's performance is carefully examined at this phase, along with all the steps that have been taken so far. The next steps are then decided.

Model and Business Evaluation: Test the final model also to check whether the model is working properly or not You can take a different set of data and evaluate the performance of the system and check whether this model meets the business goals.

Reviewing Process: All the phases are reviewed and checked whether they are functioning properly or not, and also doing error analysis if an error is found at the time of reviewing. The next step is to identify possible improvements based on the identified error patterns.

6. Deployment

At this stage, the model is deployed into a real-world environment, and other processes such as planning deployment, monitoring, maintenance of the model, and finalizing the project are carried out.

Deployment Planning:
Make sure that the model easily supports a large number of documents and guides on what will be the best way to integrate the model in the real world.

Monitoring:
For this, you can make alert systems to alert when the system's performance is degrading. Before deploying, we need to finalize the project.

After performing these six stages, we can successfully deploy the model COIN (Contract Intelligence) into the real world. And now we need to train the users on how to use this model so that it can be implemented for the betterment of the company. In this way, we can create a model that will automate the classification of various legal documents.