ASSIGNMENT - 1

Course 1: Introduction to Data Analytics

Project Title: **J P Morgan Classification for legal documents**

Video Link : Drive Link (https://drive.google.com/file/d/1O09aMZYLofsVetZeKYcJK_Ce8-15dZLm/view?usp=drive_link)

Introduction About CRISP- DM:

It stands for Cross Industry Standard Process for Data Mining. It is a cyclic process that provides structured approach to plan, organize and deploy a Data Mining Project. There are 6 phases in CRISP – DM

1. **Business Understanding**: *Focuses on understanding the objective and requirements of project from business perspective. Define Objective => Accessing current situation => Determine goal => Producing Project plan*
2. **Data Understanding**: *The Data Scientists begin the initial data collection and familiarizes themselves with the data. Gather Data => Describe Data =>Explore Data => Verify Data Quality*
3. **Data Preparation***: For Data cleaning and Transforming raw data into suitable format for modelling. Selecting Data => Cleaning Data =>Integrating Data =>Formatting Data.*
4. **Modelling:** *With clean data in hand various modelling techniques are applied. Each may require specific formats, If it doesn't suit well then it is looped back to data preparation phase. Selecting Model => Design Tests =>Build the model =>Assessing the model.*
5. **Evaluation**: *Model is thoroughly evaluated, this ensures it meets the business objectives mentioned. Evaluating results => Review the process => Determining the next steps.*
6. **Deployment***: It deploys the model into real world environment. Planning Deployment => Monitoring and maintenance => Review the project => Finalize the project.*

Problem Statement :

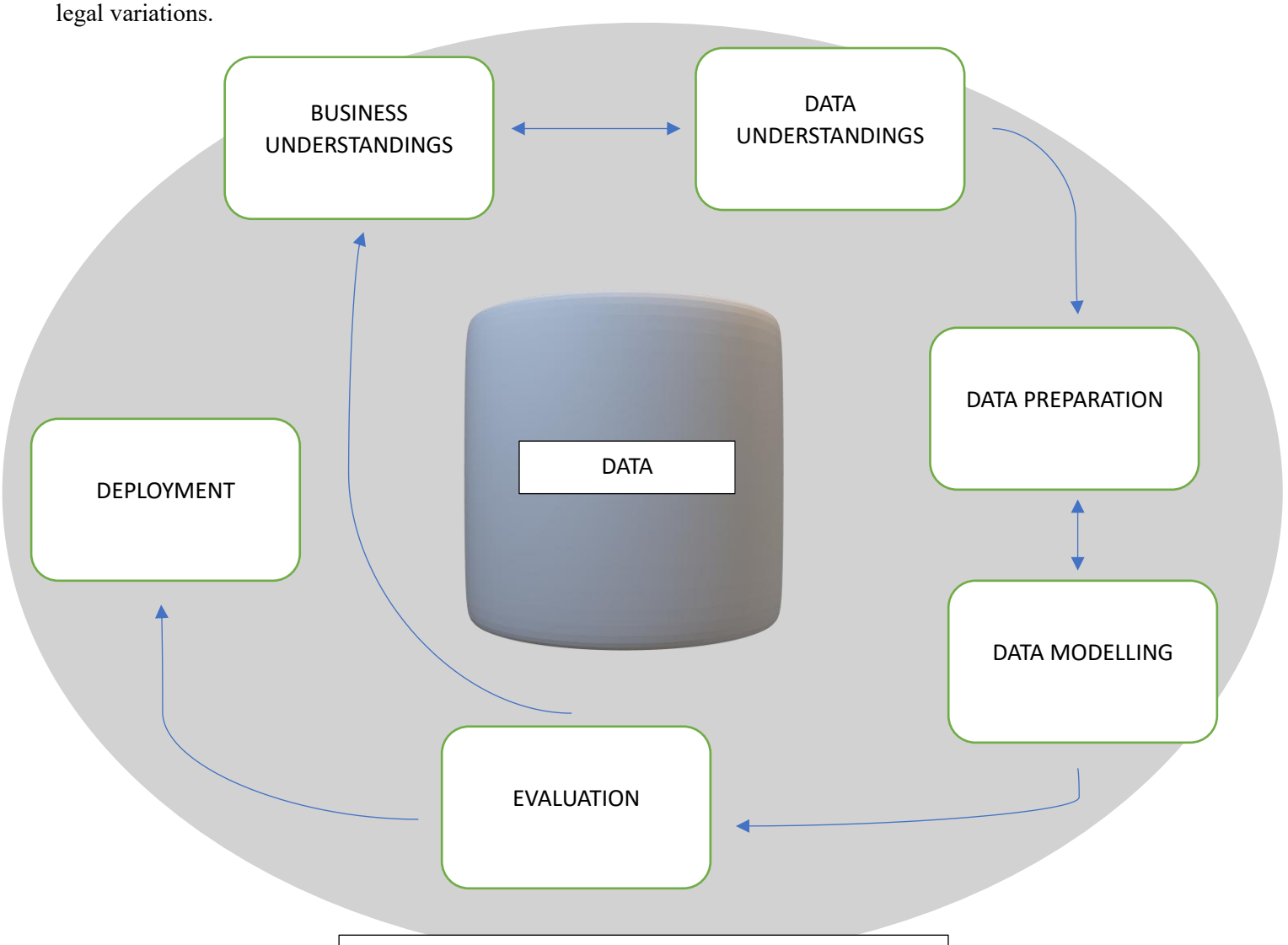
Breaking the problem/scenario down with the help of CRISP-DM methodology: "Automate the classification of various legal documents".

Let's Explore the problem statement in step by step procedure, Mapping each phase with specific project requirements.

1) Business Understanding:
   - Objective : JP Morgan wants to automate the classification of various legal documents for verification to save time and efforts.
   - Current Scenario : Lawyers and Loan Officers take 360,000 hours to complete the task manually.
   - Goal :
     i. To reduce long 360,000 hours and complete the task within seconds.
     ii. To increase the Accuracy and remove the man made errors like loan-servicing mistakes.
     iii. Enhance the model for more complex filings in the future.
   - Project Plan : Must make sure that model performs well, so regular maintenance and updates to be done. As it handles sensitive information of the applicants it must follow the security protocols and must follow the regulatory standards.
2) Data Understanding:
   - Gather the relevant legal documents: like loan agreements, Credit-default swaps and other necessary agreements like custody agreements.
   - Explore the data: identify common clauses and attributes across documents. And try to understand the patterns, terminology and structure of these documents
   - Access Data Quality : Check for the incomplete data or missing values, check for the anomalies or the outlier values that don't fit the standard patterns.
3) Data Preparation:
   - Select Data : Identify and select the elements which are necessary to investigate or perform any operations.
   - Clean Data : If documents are in different versions convert it to one readable file format (Eg: PDF)
   - Integrating Data : we can apply Feature Engineering (Text tokenization, Normalization and attribute tagging) and transform the data into one suitable format (Vectorize)
   - Formatting data : To evaluate a model we must split the data into Train set and Test set.

4) Modelling:
   - Select Model : Choose the model which performs the best in case. Select from Machine Learning models like Naïve Bayes, SVM, Deep Learning (CNN, RNN) as mentioned for image processing, include AI features to automate and enhance the model.
   - Test: Consider small amount of data given and test the working of the model. Run multiple test to identify the performance of the model.
   - Building model : Train models on the prepared dataset and experiment with different algorithms.
   - Access the model: Model Evaluation check for Accuracy , Precision , Recall and F1 Score.
5) Evaluation :
   - Evaluation results : Check if the model classify the documents accurately or not and if there are any clauses or documents types where the model underperforms.
   - Review the process : review it by analysing the business impact and validate with the stakeholders
   - Determine : Determine if the model is ready to deploy or not and if there is any other enhancement to be done.
6) Deployment:
   - Planning Deployment : Integrate this model into JP Morgan's Existing system and develop a UI (User Interface) for legal team to interact with the tool.
   - Monitoring and Maintenance : Continuously track performance over the time and if there are any issues create a feedback mechanism for understanding the user experiences and errors.
   - Review the project : By reviewing the project create a documentation and support resources in case if any stakeholder requires it. Or conduct training with the interface to familiarize them with the application.
   - Finalize the project : After successful evaluation when the model is fit and working well against all circumstances and is error free then it is set to deploy into the real world environment.

By using Feedback Loops we can collect user feedback and identify the areas of improvement and update the model accordingly and also look after the future enhancements that can be made to incorporate additional languages or regional legal variations.



CRISP -DM Framework