CRISP-DM Methodology

CRISP-DM stands for Cross-Industry Standard Process for Data Mining. It's a widely used framework for organizing and structuring data mining projects, particularly in data science and machine learning. It provides a structured approach to guide the data mining process, ensuring that projects are well-planned and executed, and that the results are relevant and actionable.

Phases of CRISP-DM Methodology

There are 6 phases of CRISP-DM Methodology:

1.Business Understanding: It is the first phase of the CRISP-DM process, and it focuses on gaining a clear understanding of the project objectives from a business perspective, then converting them into a data science or analytics problem. The key tasks are:

- Identify business objectives.
- Assess the current business situation.
- Define data mining goals (e.g., classify documents, detect patterns).
- Produce a project plan outlining steps, resources, and timelines.

2. Data Understanding: It focuses on collecting, exploring, and analyzing yourself with the data available for the project. The key tasks are:

- Collect data from relevant sources.
- Describe data
- Explore the data for initial patterns or insights.
- Identify data quality issues such as missing values or noise.

3. Data Preparation: It is also referred to as "data munging" , where the raw data is cleaned, transformed, and organized into a suitable format for modelling. The key tasks are:

- Select the relevant data for analysis.
- Clean data
- Construct new features
- Integrate data from multiple sources.

4. Modelling: In this phase, various models are built and assessed based on several modelling techniques. The key tasks are:

- Choose suitable algorithms
- Split data into training/testing sets.
- Fine-tune model parameters
- Compare model performance.

5. Evaluation: It is a phase where it ensures the model meets business objectives and performs well. The key tasks are:

- Evaluate model using performance metrics
- Review results with stakeholders or domain experts.
- Perform error analysis and validate results.
- Decide whether the model is ready for deployment.

6. Deployment: This is the final phase of the CRISP-DM process, where the results of the data analysis or model are implemented into a real-world environment for use by stakeholders or systems. The key tasks are:

- Deploy model into production
- Generate final reports or visualizations.
- Monitor model performance over time.
- Schedule model updates or re-training if needed.

J P Morgan-Introduction

J.P. Morgan & Co. is a prominent American financial institution specializing in investment banking, asset management, and private banking. Founded in 1871 by J.P. Morgan, the company is now a subsidiary of JPMorgan Chase, a global financial services firm.

J.P. Morgan & Co.	
J.P.Morgan	
Company type	Subsidiary
Industry	Investment banking Asset management Private banking
Founded	1871; 154 years ago
Founder	J. P. Morgan Anthony Drexel
Headquarters	New York City, U.S.
Number of employees	296,000 (2022)
Parent	JPMorgan Chase
Website	www.jpmorgan.com

JP Morgan COIN : Revolutionizing Legal Document Review

COIN is an AI-based software developed by JP Morgan that **automatically reads and analyses legal documents**, especially commercial loan contracts, to **save time and reduce human error**. **COIN** is capable of performing following tasks:

- **Contract Clause Classification:** COIN uses machine learning and natural language processing (NLP) to recognize patterns in wording, structure, and context.
- **Task Assignment:** The software analyses contracts to maintain accuracy and legal compliance while improving speed.
- **Potential for Expansion:** JP Morgan's COIN (Contract Intelligence) system has significant potential for expansion beyond its current use in analyzing commercial loan contracts.

Automate the classification of various legal documents with the help of CRISP-DM

1.Business Understanding

- **Objective:** Define the project objectives from a business perspective and convert them into data science goals.

- **Business Goals:**
  - **Automate the review of legal documents** such as commercial loan contracts.
  - **Reduce manual workload** for legal staff (currently consuming over 360,000 hours/year).
  - **Minimize human error** in contract analysis.
- **Data Mining Goals:**Develop a model to classify clauses in legal documents into predefined attributes accurately.

2.Data Understanding:

- ### Data Collection:
  - Gather historical legal documents such as commercial loan agreements and other contracts.
  - Include both **scanned documents** (PDFs/images) and **digitally available texts**.
- ### Initial Data Exploration:
  - Perform exploratory analysis to assess to understand data structure and distribution.
  - Identify common features and patterns in the document.
  - Assess data quality, noting any gaps or inconsistencies.
- ### Data Quality:
  - Check for OCR errors in scanned documents.
  - Identify **missing or incomplete clauses**.
  - Detect **inconsistent terminology** or formatting.

3. Data Preparation:

- ### Data Cleaning:
  - Remove non-relevant content
  - Standardize **punctuation, spacing**, and **legal terminology**.
- ### Data Transformation:
  - Transform data using tools like MS Excel, SQL or Python.
- ### Data Integration:
  - Merge data from various sources to create a comprehensive data format.

4. Modelling:

- ### Model Selection:
  - Choose appropriate algorithms for text classification, to use in AI and Deep learning models.
- ### Model Training:

- Split data into training, validation and testing.
- Train multiple model and finetune hyperparameters using cross validation.

5. Evaluation:

- **Model Robustness:**
  - Check model behaviour on unseen contracts from different sources or jurisdictions.
  - Evaluate how well it generalizes to new document types or formats.
- **User Feedback Integration:**
  - Gather feedback from potential end-users (e.g., legal staff) and adjust the model or outputs accordingly.
- **Risk Assessment:**
  - Identify any potential risks of false positives or false negatives, especially in high-stakes clauses (e.g., termination, collateral).

6. Deployment:

- **Model Integration:**
  - Embed the trained clause classification model into JP Morgan's document management or contract review system.
  - Ensure seamless workflow with existing legal tools or platforms.
- **Monitoring:**
  - Setup monitoring to track model performance and detect any issue over time.
  - Regularly update the data with new model to improve accuracy.
- **Documentation and User Training:**
  - Train legal staff on how to use the system effectively.
  - Provide documentation on interpreting outputs and flagging errors.
- **Communication:**
  - Regularly update key stakeholders on model progress, results, and deployment readiness through meetings, reports, or dashboards.
  - Establish a system for end-users to provide feedback on clause classification accuracy, usability, or errors, ensuring continuous improvement and trust in the system.

CONCLUSION

The JP Morgan COIN project demonstrates the transformative potential of artificial intelligence and data analytics in the legal and financial services sector. By automating the classification of

legal clauses in complex contracts, COIN significantly reduces manual effort, improves accuracy, and enhances operational efficiency.