Project Title: J P Morgan classification for legal documents

## Goal:

Classification and interpretation of legal documents (e.g., loan agreements, credit-default swaps) to reduce manual effort, increase accuracy, and ensure compliance using proprietary Technology called **COIN** or Contrast Intelligence in short, using **CRISP-DM.**

### 1. Business Understanding (Objective Clarification)

**Business Objectives:**

- Reduce human effort and time taken due to manual legal reviews.
- Minimize errors in legal documentation.
- Enable faster response to regulatory changes and contract handling.

**Project plan:**

- Build model with High classification accuracy (>95%):

  The algorithm should classify clauses into one of about one hundred and fifty different “attributes” of credit contracts. For example, it may note certain patterns based on clause wording or location in the agreement.

- Reduction in processing time from weeks to seconds.
- Reduction in loan-servicing mistakes.
- Scalable solution for various contract types.

### 2. Data Understanding (Exploratory Analysis)

**Data Gathering (Unstructured data and Structured Data)**

- Historical legal contracts (e.g., commercial loans, credit-default swaps).
- Labelled clauses (if available).
- Meta-data (document types, version history, timestamps, responsible departments).

**Data Description:**

- Identify types of contracts (PDFs, scanned images, DOCs).
- Analyse distribution of clauses and key legal terms.
- Understand document structure: headers, paragraphs, sections, clause patterns.
- Check for OCR needs (image-to-text conversion for scanned docs).

**Data exploration:**

- Certain clauses appear in consistent formats/locations.
- Legal language patterns are repetitive and can be learned by models.
- This step can be done using tools like python, SQL, MS power BI for visualization.

**Data Verification:**

- Ensuring collected data has no data type inconsistencies and null values, like different file format or difference in structure of the data set compared to required ones.

## 3. Data Preparation:

**Data Selection:**

- Selection of files which are most relevant to the current project and can be easily done by the COIN Software
- Selection only relevant clauses that are most repeated or most come across.

**Data Cleaning:**

- **OCR processing** for scanned documents using tools like Tesseract or AWS Textract.
- **Text normalization:** removing stop-words, correcting OCR errors, tokenization, lowercasing.
- **Section and clause segmentation** using NLP-based chunking.
- **Labelling clauses** into one of 150+ known attributes.
- **Feature engineering:** note certain patterns based on clause wording or location in the agreement, keyword density.
- **Resampling** if needed.

**Data Integration and Formatting:**

- Integrate any similar file formats into one file or bring it to the similar Data structure.

**Outcome:**
A structured and clean clause-level dataset ready for training ML models.

## 4. Modelling (Algorithm Development)

**Selection of Modelling techniques (Machine Learning):**

- **Text Classification Algorithms:** Logistic Regression, SVM, Random Forests.
- **Deep Learning:** BERT, Roberta for context-aware clause classification.
- **Clustering:** For discovering new or emerging clause types.
- **Ensemble Models:** To boost precision/recall in multi-class classification.

**Designing of Tests:**

- Check for Error in the model and its data sources.
- Check for any outliers than are skewing the results etc….
- Check for most commonly occurring clauses and if they are classified accurately
- Check for rare but critical clauses (e.g., risk-related clauses) and if they are handled well.

**Building of Model:**

- Train model to classify clauses into 150+ legal categories.
- Check which ML tool that is suitable and fits according to our business requirements and data sets.

**Assessing the model:**

- Use cross-validation to tune hyperparameters such as number of iterations required before validating the test or the sample size of the data to be considered for the testing, etc….
- Evaluate multiple models using F1-score, accuracy, and confusion matrix which categorizes the predicted and actual outcomes.

**Tools** than can be used are**:** Python with libraries like Scikit-learn, TensorFlow, Hugging Face etc…

## 5. Evaluation (Validation of Results)

**Result Evaluation:**

- **Accuracy & F1 Score**: Balanced view of performance across classes.
- Elimination of the need for 360,000+ hours of manual legal review.

**Process Review:**

- **Precision & Recall**: Especially important due to high impact of false positives/negatives.
- **Interpretability**: Is model interpretable to legal teams?
- **Business Validation**: Do legal experts agree with model outputs?
- **Error Analysis**: Manually review of misclassified clauses.

**Result:**
High-performing model validated through legal expert feedback and test datasets.

**Next Steps:**

- If the model has any discrepancy and does not align with business objectives, we go back to redefining our goals and select different data sets.
- If the model meets our requirements, it is ready for deployment.

## 6. Deployment (Real-world Implementation)

**Deployment Plan:**

- Integrate model with JP Morgan's internal document management system.
- Design a **user interface** for legal teams to review AI-suggested clause classifications.
- Implement **human-in-the-loop** review for sensitive contracts.
- Continuous **model retraining** with feedback loop from legal reviewers.

**Monitoring KPIs:**

- Accuracy drift over time.
- Volume of documents processed daily.

- Frequency of manual overrides.
- Time saved per document review.

**Review and Long-Term Scalability:**

- Expand to new legal document types (e.g., NDAs, custody agreements).
- Incorporate adaptive learning for evolving regulations.
- Allow multilingual contract processing.

**Finalization:**

By following the **CRISP-DM methodology**, JP Morgan's COIN project is systematically developed from business need to deployed solution and deployed.

**CONCLUSION:**

The structured **CRISP-DM** approach provides a comprehensive and systematic roadmap for automating the classification of legal documents at JP Morgan using the **COIN** software. Beginning with a clear understanding of the business objectives—reducing manual workload, improving accuracy, and enhancing regulatory compliance—we translated the problem into a data science framework. Through careful data exploration, cleaning, and preparation, we built a solid foundation for reliable model training. By applying advanced modelling techniques and refining them using cross-validation, we ensured the development of a high-performance, scalable solution. The evaluation phase validated the model's effectiveness, while the deployment plan focused on seamless integration, continuous learning, and measurable impact. CRISP-DM-driven methodology establishes a robust foundation for legal document automation.

Video Link: https://drive.google.com/file/d/14RJsx-umT5E686EsQ6gcdD6JOOszW-xY/view?usp=sharing