Expense Categorizer

AI-Based Approaches:

- Machine Learning Classifiers
- Natural Language Processing (NLP)
- Behavioral Analysis (spending patterns)
- Data extraction with OCR

SOTA:

- Decision Tree
- Random Forests
- Support Vector Machines (SVM)
- Text Preprocessing (tokenization, stopword removal, etc.
- Feature Extraction
- Sentiment Analysis
- Feedforward Neural Network
- LSTM (for sequential spend patterns)

Possible Datasets:

- https://catalog.data.gov/dataset/nyc-independent-budget-office-ibo-agency-expenditures-fy-1980-2018 or find others from https://data.gov/
- https://www.kaggle.com/datasets?sort=votes&tags=11108-finance
- Or we can generate our own random

Sources:

- https://pmc.ncbi.nlm.nih.gov/articles/PMC9527075/
- https://arxiv.org/pdf/2312.07730
- https://arxiv.org/pdf/2102.07635
- https://www.researchgate.net/publication/387721964_REAL-TIME_AI_OPTICAL_CHARACTER_RECOGNITION_ENHANCING_DATA_PROCESSING_WITH_SPEED_AND_ACCURACY

Medical Document Classifier (Challenging due to the limited amount of data sets but interesting topic)

AI-Based Approaches:

- Machine Learning Classifiers
- Natural Language Processing (NLP)
- Document Image Analysis (OCR + CV)

SOTA:

- Decision Trees
- SVM
- Random Forest
- K-means (for clustering note types)
- Text Preprocessing
- Named Entity Recognition (NER)
- Medical Term Extraction (using BioBERT, ClinicalBERT)
- LSTM
- CNN (for scanned documents)
- LayoutLMv3 (for form/document structure)

Possible Datasets:

- MIMIC-III Clinical Notes Dataset https://paperswithcode.com/dataset/mimic-iii
- https://www.kaggle.com/competitions/nbme-score-clinical-patient-notes/data
- Places where we can look for data sets
- https://www.shaip.com/blog/healthcare-datasets-for-machine-learning-projects/

Sources:

- https://arxiv.org/html/2503.01159v1
- https://www.314e.com/engineering-hub/cracking-the-code-ai-native-intelligent-document-processing-for-medical-records/
- https://arxiv.org/pdf/2310.07282
- https://www.cambridge.org/core/services/aop-cambridge-core/content/view/BEF81FDE6E12B9DC5AD4906AE67CDDEB/S1351324923000542a.pdf/lightweight-transformers-for-clinical-natural-language-processing.pdf

Chemical Inventory Tracker (Doesnt need to be chemicals could be anything)

AI-Based Approaches:

- Machine Learning Classifiers
- NLP for label/name recognition
- Computer Vision for image-based inventory scans
- Time Series Analysis for usage patterns

SOTA:

- Random Forests
- SVM
- Text Preprocessing
- Entity Recognition (chemical names, CAS numbers)
- CNN (detect bottle labels, volumes)
- LSTM (for tracking depletion patterns)

Possible Datasets:

- Internal lab inventory logs (structured or semi-structured CSVs)
- ChEMBL or PubChem(for compound info)
- Custom dataset from lab images (labeled chemical containers)
- or if we changed the inventory type we can look for other inventory logs

Sources: Not chemical focused more generalized

- **https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5199941**
- **https://www.researchgate.net/publication/388245304_Artificial_Intelligence-Driven_Inventory_Management_Optimizing_Stock_Levels_and_Reducing_Costs_Through_Advanced_Machine_Learning_Techniques**
- **https://www.researchgate.net/publication/390755582_AI-Powered_Smart_Inventory_Management_Enhancing_Efficiency_Through_Predictive_Analytics_and_Automation**