AI/ML Intern Assessment for Intern at Cozii Technologies

Important: Complete one (1) assessment only!

Option 1. Personalized Recommendation System

Task: Build a recommendation system that suggests the best product based on a user's past preferences and behavior.

Dataset Sources:

- **MovieLens Dataset** (GroupLens) – Movie rating data for collaborative filtering. https://grouplens.org/datasets/movielens/
- **Amazon Product Review Dataset** – Customer reviews and ratings for different products. https://nijianmo.github.io/amazon/index.html
- **Goodreads Book Reviews Dataset** – User ratings and reviews for books. https://sites.google.com/eng.ucsd.edu/ucsdbookgraph/home
- **Retail Rocket Recommender System Dataset** – User-item interactions for retail products. https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset

Evaluation Focus:
✓ **Approach Selection:** Content-based filtering vs. Collaborative Filtering (User-User, Item-Item)
✓ **Similarity Measures:** Cosine Similarity, Pearson Correlation, Jaccard Index
✓ **Advanced Methods:** Matrix Factorization (SVD, NMF), Deep Learning (Neural Collaborative Filtering)
✓ **Data Preprocessing & Feature Engineering:** Handling missing data, normalization, encoding categorical variables
✓ **Evaluation Metrics:** Precision@K, Recall@K, NDCG (Normalized Discounted Cumulative Gain), RMSE (if predicting ratings)

Submission Requirements:

- Python code (Jupyter Notebook/Google Colab/ Anaconda)
- Short report (1-2 pages) explaining methodology, model selection, improvements, and future optimization
- Submit your project on a GitHub repository and provide a live a link

Option 2. Fraud Detection in Rent Payments

Task: Develop an ML model to detect fraudulent rental transactions using historical tenant behavior and transaction data.

Dataset Sources:

- **IEEE-CIS Fraud Detection Dataset** – Transactional data with fraud labels. https://www.kaggle.com/competitions/ieee-fraud-detection
- **PaySim (Synthetic Financial Transactions)** – Simulated mobile money transactions with fraud cases. https://www.kaggle.com/datasets/ealaxi/paysim1
- **Credit Card Fraud Detection Dataset** – Credit card transactions labeled as fraudulent or genuine. https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud
- **Synthetic Fraud Data (LoL Dataset)** – Log-based online fraud dataset. https://www.kaggle.com/datasets/dhanushnarayananr/credit-card-fraud

Evaluation Focus:
✓ Supervised vs. Unsupervised Learning:

- Logistic Regression, XGBoost (Supervised)
- Isolation Forest, One-Class SVM (Unsupervised)

✓ Handling Imbalanced Data:

- SMOTE (Synthetic Minority Over-sampling Technique)
- Class Weighting in Loss Function

✓ Feature Engineering:

- Analyzing time-based spending patterns, frequency, transaction location anomalies

✓ Precision-Recall Tradeoff:

- Focus on F1-Score, Precision, Recall, AUC-ROC

Submission Requirements:

- Python code (Jupyter Notebook/Google Colab/ Anaconda)
- Short report (1-2 pages) explaining methodology, model selection, improvements, and future optimization
- Submit your project on a GitHub repository and provide a live a link