# **Data Mining Capstone Project Guide**

## **Data Science Diploma Program**

**Total Score: 100 Points**

## **Project Overview**

This capstone project demonstrates your mastery of data mining concepts and techniques covered throughout the course. You will complete a comprehensive data science project that spans the entire data mining pipeline from data collection to insights presentation.

## **Project Structure & Scoring Rubric**

### **1. Project Proposal & Planning (10 points)**

* **Problem Definition (5 points)**: Clear articulation of business problem and objectives
* **Dataset Description (3 points)**: Comprehensive description of data sources and characteristics
* **Methodology Overview (2 points)**: High-level approach and techniques planned

### **2. Data Wrangling (25 points)**

#### **2.1 Data Collection (8 points)**

* **Data Sources (4 points)**: Multiple sources (files, web scraping, APIs, databases)
* **Data Quality Assessment (4 points)**: Initial data quality evaluation

#### **2.2 Data Integration & Preprocessing (12 points)**

* **Data Integration (4 points)**: Combining multiple data sources effectively
* **Missing Value Handling (3 points)**: Appropriate strategies for missing data
* **Outlier Detection & Treatment (3 points)**: Statistical and visual outlier analysis
* **Data Transformation (2 points)**: Scaling, discretization, and feature engineering

#### **2.3 Exploratory Data Analysis (5 points)**

* **Statistical Analysis (2 points)**: Descriptive statistics and correlation analysis
* **Data Visualization (3 points)**: Comprehensive visual exploration using multiple tools

### **3. Data Analysis & Modeling (35 points)**

#### **3.1 Classification Tasks (10 points)**

Choose and implement **minimum 2** classification algorithms:

* K-Nearest Neighbors (KNN/RNN)
* Decision Trees
* Support Vector Machines (SVM)
* Naive Bayes
* Logistic Regression
* Random Forests

#### **3.2 Regression Analysis (8 points)**

Implement **minimum 1** regression technique:

* Simple/Multiple Linear Regression
* Regularization techniques (Ridge, Lasso)
* Ensemble methods

#### **3.3 Clustering Analysis (8 points)**

Apply **minimum 1** clustering method:

* K-Means (Partition-based)
* Hierarchical Clustering
* DBSCAN (Density-based)
* Grid-based Clustering

#### **3.4 Advanced Analytics (9 points)**

Choose **minimum 1** advanced technique:

* **Frequent Pattern Mining (5 points)**: Apriori or FP-Growth algorithms
* **Outlier Detection (4 points)**: Advanced outlier detection methods
* **Principal Component Analysis (4 points)**: Dimensionality reduction
* **Association Rules (5 points)**: Market basket analysis

### **4. Model Evaluation & Validation (15 points)**

* **Cross-Validation (5 points)**: Proper validation techniques
* **Performance Metrics (5 points)**: Appropriate metrics for each task type
* **Model Comparison (5 points)**: Comparative analysis of different approaches

### **5. Results & Business Insights (10 points)**

* **Interpretation (5 points)**: Clear explanation of results and their significance
* **Business Recommendations (5 points)**: Actionable insights and recommendations

### **6. Presentation & Documentation (5 points)**

* **Code Quality (2 points)**: Clean, well-documented, reproducible code
* **Report Quality (2 points)**: Professional presentation of findings
* **Technical Communication (1 point)**: Clear explanation of technical concepts

## **Suggested Project Topics**

### **Business Analytics Projects**

1. **Customer Segmentation & Recommendation System**
   * Retail/e-commerce customer behavior analysis
   * Clustering customers and building recommendation engines
   * *Techniques*: K-Means, Association Rules, Collaborative Filtering
2. **Financial Risk Assessment**
   * Credit scoring and fraud detection
   * Loan default prediction
   * *Techniques*: Logistic Regression, SVM, Decision Trees, Outlier Detection
3. **Marketing Campaign Optimization**
   * Customer response prediction
   * Market basket analysis
   * *Techniques*: Classification algorithms, Frequent Pattern Mining

### **Healthcare Analytics Projects**

1. **Disease Prediction & Diagnosis**
   * Medical diagnosis support systems
   * Health risk assessment
   * *Techniques*: Multiple classification algorithms, Feature selection
2. **Drug Discovery Analytics**
   * Molecular activity prediction
   * Clinical trial optimization
   * *Techniques*: SVM, Neural Networks, Clustering

### **Technology & Web Analytics Projects**

1. **Social Media Sentiment Analysis**
   * Opinion mining and trend analysis
   * Influencer identification
   * *Techniques*: Text mining, Classification, Clustering
2. **Recommendation Systems**
   * Movie/Music/Product recommendations
   * Content-based and collaborative filtering
   * *Techniques*: Association Rules, Clustering, Classification

### **Operations & Supply Chain Projects**

1. **Demand Forecasting**
   * Sales prediction and inventory optimization
   * Seasonal pattern analysis
   * *Techniques*: Regression, Time series analysis, Clustering
2. **Quality Control & Anomaly Detection**
   * Manufacturing defect prediction
   * System monitoring and alerting
   * *Techniques*: Outlier Detection, Classification, Statistical Process Control

### **Environmental & Social Projects**

1. **Climate Data Analysis**
   * Weather pattern prediction
   * Environmental impact assessment
   * *Techniques*: Regression, Clustering, Time series analysis

## **Technical Requirements**

### **Programming & Tools**

* **Primary Language**: Python (recommended) or R
* **Required Libraries**:
  + Data Manipulation: Pandas, NumPy
  + Visualization: Matplotlib, Seaborn, Plotly
  + Machine Learning: Scikit-learn
  + Database: SQLite, SQL connectors
  + Web Scraping: Beautiful Soup, Scrapy (if applicable)

### **Data Requirements**

* **Minimum Dataset Size**: 1,000+ records
* **Multiple Data Sources**: At least 2 different data sources
* **Data Variety**: Structured and/or semi-structured data
* **Real-world Data**: Encourage use of real business datasets

### **Documentation Requirements**

1. **Jupyter Notebook**: Complete analysis with markdown documentation
2. **Technical Report**: 15-20 page comprehensive report
3. **Presentation**: 15-20 minute presentation with slides
4. **Code Repository**: Well-organized GitHub repository

# **Danh Sách Dataset Phù Hợp Cho Đồ Án Khai Thác Dữ Liệu**

## **1. Phân Khúc Khách Hàng & Hệ Thống Gợi Ý (Customer Segmentation & Recommendation)**

### **Dataset 1: Online Retail Dataset (UCI)**

* **Nguồn**: UCI Machine Learning Repository
* **Link tải**:<https://doi.org/10.24432/C5BW33>
* **Mô tả**: Dữ liệu giao dịch từ công ty bán lẻ trực tuyến tại Anh (12/2010 - 09/2011) [Uci](https://archive-beta.ics.uci.edu/dataset/352/online+retail)[UCI Machine Learning Repository](https://archive.ics.uci.edu/datasets)
* **Kích thước**: 541,909 giao dịch, [Eagronin](https://eagronin.github.io/market-basket-acquire/) 8 đặc trưng
* **Đặc điểm chính**: Dữ liệu thực tế hoàn hảo cho phân tích RFM, phân khúc khách hàng, và luật kết hợp [UCI Machine Learning Repository](https://archive.ics.uci.edu/datasets)
* **Độ khó**: Cơ bản đến Trung bình - cần làm sạch dữ liệu nhưng tài liệu đầy đủ
* **Phù hợp cho sinh viên**: Xuất sắc cho học phân khúc khách hàng và clustering

### **Dataset 2: Online Retail II Dataset (UCI)**

* **Nguồn**: UCI Machine Learning Repository
* **Link tải**:<https://doi.org/10.24432/C5CG6D>
* **Mô tả**: Phiên bản mở rộng với 2 năm dữ liệu (12/2009 - 12/2011) [Medium](https://medium.com/@mayureshrpalav/clustering-customers-online-retail-dataset-516e961e7bc)
* **Kích thước**: Dataset lớn hơn với 2 năm dữ liệu, 8 đặc trưng
* **Đặc điểm chính**: Chuỗi thời gian dài hơn để phân tích xu hướng, mùa vụ, vòng đời khách hàng
* **Độ khó**: Trung bình - phức tạp hơn do kích thước lớn
* **Phù hợp cho sinh viên**: Tốt cho phân tích nâng cao và time-based analysis

### **Dataset 3: E-Commerce Customer Behavior Dataset (Kaggle)**

* **Nguồn**: Kaggle
* **Mô tả**: Dữ liệu giao dịch và hành vi khách hàng thương mại điện tử
* **Kích thước**: Thay đổi theo phiên bản
* **Đặc điểm chính**: Bối cảnh e-commerce hiện đại, sẵn sàng cho collaborative filtering
* **Độ khó**: Cơ bản đến Trung bình
* **Phù hợp cho sinh viên**: Hoàn hảo cho phát triển hệ thống gợi ý

## **2. Đánh Giá Rủi Ro Tài Chính (Financial Risk Assessment)**

### **Dataset 1: Credit Card Fraud Detection**

* **Nguồn**: Kaggle
* **Link tải**:<https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>
* **Mô tả**: Giao dịch thẻ tín dụng châu Âu (09/2013) với nhãn gian lận [XGBoosting](https://xgboosting.com/xgboost-for-the-kaggle-credit-card-fraud-detection-dataset/)[GitHub](https://github.com/pankush9096/kaggle-Credit-Card-Fraud-Detection)
* **Kích thước**: 284,807 giao dịch, 30 đặc trưng [XGBoosting](https://xgboosting.com/xgboost-for-the-kaggle-credit-card-fraud-detection-dataset/)[GitHub](https://github.com/pankush9096/kaggle-Credit-Card-Fraud-Detection)
* **Đặc điểm chính**: Mất cân bằng cao (0.172% gian lận), 28 features đã PCA [XGBoosting](https://xgboosting.com/xgboost-for-the-kaggle-credit-card-fraud-detection-dataset/)[GitHub](https://github.com/pankush9096/kaggle-Credit-Card-Fraud-Detection)
* **Độ khó**: Trung bình - học về imbalanced datasets
* **Phù hợp cho sinh viên**: Tốt cho học fraud detection và xử lý dữ liệu mất cân bằng

### **Dataset 2: German Credit Data (Statlog)**

* **Nguồn**: UCI ML Repository
* **Link tải**:<https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data>
* **Mô tả**: Phân loại người vay là rủi ro tín dụng tốt hoặc xấu [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data)
* **Kích thước**: 1,000 mẫu, 20 đặc trưng [SpringerOpen](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-024-00882-0)
* **Đặc điểm chính**: 700 tín dụng tốt, 300 tín dụng xấu, [SpringerOpen](https://journalofbigdata.springeropen.com/articles/10.1186/s40537-024-00882-0) có ma trận chi phí [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data)
* **Độ khó**: Cơ bản đến Trung bình
* **Phù hợp cho sinh viên**: Xuất sắc cho học cost-sensitive learning

### **Dataset 3: Taiwan Default Credit Card Clients**

* **Nguồn**: UCI ML Repository
* **Link tải**:<https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients>
* **Mô tả**: Dự đoán vỡ nợ thẻ tín dụng tại Đài Loan (04-09/2005) [Kaggle +2](https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset)
* **Kích thước**: 30,000 mẫu, 23 đặc trưng [GitHub](https://github.com/MatteoM95/Default-of-Credit-Card-Clients-Dataset-Analisys)[Medium](https://medium.com/@agrawal.sourabh644/default-of-credit-card-clients-dataset-a31ec2ffdea4)
* **Đặc điểm chính**: 22.1% tỷ lệ vỡ nợ, [GitHub](https://github.com/MatteoM95/Default-of-Credit-Card-Clients-Dataset-Analisys) 6 tháng lịch sử thanh toán [GitHub](https://github.com/MatteoM95/Default-of-Credit-Card-Clients-Dataset-Analisys)
* **Độ khó**: Trung bình
* **Phù hợp cho sinh viên**: Phù hợp cho mô hình rủi ro tín dụng toàn diện

## **3. Tối Ưu Hóa Chiến Dịch Marketing (Marketing Campaign Optimization)**

### **Dataset 1: Bank Marketing Dataset (UCI)**

* **Nguồn**: UCI ML Repository
* **Link tải**:<https://doi.org/10.24432/C5K306>
* **Mô tả**: Chiến dịch marketing qua điện thoại của ngân hàng Bồ Đào Nha [RStudio +2](https://rstudio-pubs-static.s3.amazonaws.com/122867_8b77eae02a5240d5a06307e11de09f4f.html)
* **Kích thước**: 41,188 mẫu (full), 4,119 mẫu (10% subset), 20 đặc trưng [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/222/bank+marketing)
* **Đặc điểm chính**: Dữ liệu telemarketing thực, [RStudio](https://rstudio-pubs-static.s3.amazonaws.com/122867_8b77eae02a5240d5a06307e11de09f4f.html) imbalanced, có biến kinh tế vĩ mô [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/222/bank+marketing)
* **Độ khó**: Trung bình - cần xử lý imbalanced data và feature engineering
* **Phù hợp cho sinh viên**: Xuất sắc cho phân loại và business analytics

### **Dataset 2: Marketing Campaign Response Dataset (Kaggle)**

* **Nguồn**: Kaggle
* **Mô tả**: Dữ liệu hiệu suất chiến dịch marketing qua các phân khúc khách hàng
* **Kích thước**: Thay đổi theo phiên bản
* **Đặc điểm chính**: Metrics hiệu suất chiến dịch, patterns phản hồi khách hàng
* **Độ khó**: Cơ bản đến Trung bình
* **Phù hợp cho sinh viên**: Tốt cho dự đoán phản hồi và tối ưu chiến dịch

### **Dataset 3: Portuguese Bank Marketing Dataset (Kaggle)**

* **Nguồn**: Kaggle (phiên bản thay thế của UCI)
* **Mô tả**: Giống UCI Bank Marketing với preprocessing bổ sung
* **Kích thước**: Giống phiên bản UCI
* **Đặc điểm chính**: Tài liệu tốt với community notebooks
* **Độ khó**: Cơ bản đến Trung bình
* **Phù hợp cho sinh viên**: Nhiều tutorial resources có sẵn

## **4. Dự Đoán & Chẩn Đoán Bệnh (Disease Prediction & Diagnosis)**

### **Dataset 1: Heart Disease Dataset (Kaggle)**

* **Nguồn**: Kaggle
* **Link tải**:<https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>
* **Mô tả**: Kết hợp từ 5 dataset tim mạch phổ biến [IEEE DataPort](https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive) (Cleveland, Hungary, Switzerland, v.v.)
* **Kích thước**: 1,025 records, 11 features + target
* **Đặc điểm chính**: Binary classification (có/không bệnh tim)
* **Độ khó**: Cơ bản đến Trung bình
* **Phù hợp cho sinh viên**: Xuất sắc cho projects, tài liệu đầy đủ, dataset sạch

### **Dataset 2: UCI Breast Cancer Wisconsin (Diagnostic)**

* **Nguồn**: UCI ML Repository
* **Link tải**:<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)>
* **Mô tả**: Features từ ảnh số hóa FNA của khối u vú [UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic)[GeeksforGeeks](https://www.geeksforgeeks.org/machine-learning/breast-cancer-wisconsin-diagnostic-dataset/)
* **Kích thước**: 569 mẫu, 30 features [GeeksforGeeks](https://www.geeksforgeeks.org/machine-learning/breast-cancer-wisconsin-diagnostic-dataset/)
* **Đặc điểm chính**: Binary classification (ác tính/lành tính) [GeeksforGeeks](https://www.geeksforgeeks.org/machine-learning/breast-cancer-wisconsin-diagnostic-dataset/)
* **Độ khó**: Trung bình
* **Phù hợp cho sinh viên**: Tốt cho sinh viên nâng cao, benchmark dataset phổ biến

### **Dataset 3: Stroke Prediction Dataset (Kaggle)**

* **Nguồn**: Kaggle
* **Link tải**:<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>
* **Mô tả**: Dataset y tế để dự đoán đột quỵ [IEEE DataPort](https://ieee-dataport.org/documents/stroke-prediction-dataset)
* **Kích thước**: 5,110 observations, 11 features + target [IEEE DataPort](https://ieee-dataport.org/documents/stroke-prediction-dataset)
* **Đặc điểm chính**: Highly imbalanced, có missing values
* **Độ khó**: Cơ bản đến Trung bình
* **Phù hợp cho sinh viên**: Tốt cho học preprocessing và imbalanced classification

## **5. Phân Tích Khám Phá Thuốc (Drug Discovery Analytics)**

### **Dataset 1: ChEMBL Bioactivity Database**

* **Nguồn**: European Bioinformatics Institute (EBI) [Wikipedia](https://en.wikipedia.org/wiki/ChEMBL)
* **Link tải**:<https://www.ebi.ac.uk/chembl/>
* **Mô tả**: Database bioactivity quy mô lớn cho drug discovery [Wikipedia](https://en.wikipedia.org/wiki/ChEMBL)
* **Kích thước**: 2.4 triệu cấu trúc hóa học, [Oxford Academic](https://academic.oup.com/nar/article/52/D1/D1180/7337608) 1.6 triệu assays, 17,000+ targets [Oxford Academic](https://academic.oup.com/nar/article/52/D1/D1180/7337608)
* **Đặc điểm chính**: Molecular descriptors, bioactivity measurements (IC50, EC50, Ki, Kd)
* **Độ khó**: Nâng cao
* **Phù hợp cho sinh viên**: Sinh viên nâng cao với nền tảng cheminformatics

### **Dataset 2: COVID-19 Drug Repurposing Dataset**

* **Nguồn**: Kaggle
* **Link tải**:<https://www.kaggle.com/datasets/andrewmvd/covid19-drug-repurposing>
* **Mô tả**: Dữ liệu IC50 của phân tử hóa học chống COVID-19 [Kaggle](https://www.kaggle.com/datasets/divyansh22/drug-discovery-data)
* **Kích thước**: 104 phân tử hóa học [Kaggle](https://www.kaggle.com/datasets/divyansh22/drug-discovery-data) với activity data
* **Đặc điểm chính**: Molecular descriptors, IC50 values, chemical structures
* **Độ khó**: Trung bình đến Nâng cao
* **Phù hợp cho sinh viên**: Tốt cho sinh viên trung bình học drug discovery concepts

### **Dataset 3: Drug Discovery Analytics Dataset (Kaggle)**

* **Nguồn**: Kaggle
* **Link tải**:<https://www.kaggle.com/datasets/tags/drug-discovery>
* **Mô tả**: Bộ sưu tập molecular datasets cho computational drug discovery
* **Kích thước**: Thay đổi (thường 1,000-10,000 compounds)
* **Đặc điểm chính**: Molecular fingerprints, physicochemical properties, bioactivity data [UCI Machine Learning Repository](https://archive.ics.uci.edu/)
* **Độ khó**: Trung bình đến Nâng cao
* **Phù hợp cho sinh viên**: Sinh viên có kiến thức hóa học cơ bản

## **6. Phân Tích Cảm Xúc Mạng Xã Hội (Social Media Sentiment Analysis)**

### **Dataset 1: Stanford Sentiment Treebank (SST)**

* **Nguồn**: Stanford NLP Group
* **Link tải**:<https://nlp.stanford.edu/sentiment/>
* **Mô tả**: Câu đánh giá phim với nhãn sentiment chi tiết [Papers with Code](https://paperswithcode.com/datasets?task=sentiment-analysis)
* **Kích thước**: 11,855 câu, 215,154 unique phrases [Papers with Code](https://paperswithcode.com/datasets?task=sentiment-analysis)
* **Đặc điểm chính**: 5-class sentiment labeling, hierarchical annotation [Papers with Code](https://paperswithcode.com/datasets?task=sentiment-analysis)
* **Độ khó**: Trung bình
* **Phù hợp cho sinh viên**: Lý tưởng cho học NLP fundamentals

### **Dataset 2: TweetEval - Sentiment Analysis**

* **Nguồn**: Cardiff NLP (HuggingFace) [GitHub +2](https://github.com/cardiffnlp/tweeteval)
* **Link tải**: load\_dataset("tweet\_eval", "sentiment")
* **Mô tả**: Twitter-specific sentiment analysis benchmark [Papers with Code](https://paperswithcode.com/dataset/tweeteval)
* **Kích thước**: ~67,000 tweets với train/validation/test splits
* **Đặc điểm chính**: 3-class sentiment (positive, neutral, negative) [Hugging Face](https://huggingface.co/datasets/cardiffnlp/tweet_eval)
* **Độ khó**: Trung bình
* **Phù hợp cho sinh viên**: Hoàn hảo cho social media NLP

### **Dataset 3: Social Media Sentiments Analysis Dataset**

* **Nguồn**: Kaggle
* **Mô tả**: Mixed social media posts với sentiment labels
* **Kích thước**: 10,000-100,000+ posts (tùy phiên bản)
* **Đặc điểm chính**: Multi-platform data, timestamps, engagement metrics
* **Độ khó**: Cơ bản đến Trung bình
* **Phù hợp cho sinh viên**: Tốt cho beginners, nhiều notebooks trên Kaggle

## **7. Hệ Thống Gợi Ý (Recommendation Systems)**

### **Dataset 1: MovieLens 1M Dataset**

* **Nguồn**: GroupLens Research, University of Minnesota
* **Link tải**:<https://files.grouplens.org/datasets/movielens/ml-1m.zip>
* **Mô tả**: 1 triệu ratings từ 6,000 users cho 4,000 movies [Harvard Dataverse](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/NRDJCB)[GroupLens](https://grouplens.org/datasets/movielens/)
* **Kích thước**: 1,000,209 ratings, [Harvard Dataverse](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/NRDJCB) có demographics
* **Đặc điểm chính**: Stable benchmark dataset, [GroupLens](https://grouplens.org/datasets/movielens/) user demographics included
* **Độ khó**: Cơ bản đến Trung bình
* **Phù hợp cho sinh viên**: Classic dataset cho collaborative filtering tutorials

### **Dataset 2: MovieLens 25M Dataset**

* **Nguồn**: GroupLens Research
* **Link tải**:<https://files.grouplens.org/datasets/movielens/ml-25m.zip>
* **Mô tả**: Phiên bản mới nhất với 25 triệu ratings [Grouplens](https://files.grouplens.org/datasets/movielens/ml-25m-README.html)[GroupLens](https://grouplens.org/datasets/movielens/)
* **Kích thước**: 25,000,095 ratings, 62,423 movies, 162,541 users [Grouplens](https://files.grouplens.org/datasets/movielens/ml-25m-README.html)[GroupLens](https://grouplens.org/datasets/movielens/)
* **Đặc điểm chính**: Comprehensive dataset, [GroupLens](https://grouplens.org/datasets/movielens/) tag genome data included [GroupLens](https://grouplens.org/datasets/movielens/)
* **Độ khó**: Nâng cao - cần xử lý hiệu quả
* **Phù hợp cho sinh viên**: Lý tưởng cho advanced recommendation algorithms

### **Dataset 3: MovieLens 100K Dataset**

* **Nguồn**: GroupLens Research
* **Link tải**:<https://files.grouplens.org/datasets/movielens/ml-100k.zip>
* **Mô tả**: Classic benchmark với 100,000 ratings [D2l](https://classic.d2l.ai/chapter_recommender-systems/movielens.html)[GroupLens](https://grouplens.org/datasets/movielens/)
* **Kích thước**: 100,000 ratings, 1,000 users, 1,700 movies [GroupLens](https://grouplens.org/datasets/movielens/)[GitHub](https://github.com/rposhala/Recommender-System-on-MovieLens-dataset)
* **Đặc điểm chính**: Dataset nhỏ, pre-computed splits included
* **Độ khó**: Cơ bản
* **Phù hợp cho sinh viên**: Hoàn hảo cho học basic recommendation concepts

## **8. Dự Báo Nhu Cầu (Demand Forecasting)**

### **Dataset 1: Store Sales - Time Series Forecasting (Kaggle)**

* **Nguồn**: Kaggle Competition
* **Link tải**:<https://www.kaggle.com/competitions/store-sales-time-series-forecasting>
* **Mô tả**: Dự đoán doanh số grocery cho nhiều product families ở Ecuador [dotData](https://dotdata.com/blog/practical-guide-for-feature-engineering-of-time-series-data/)
* **Kích thước**: 3,000,888 training records (4 năm), 28,512 test records [Medium](https://medium.com/@humzahmalik/a-beginners-approach-to-time-series-with-working-example-c6bff9c24928)
* **Đặc điểm chính**: Hierarchical forecasting, [dotData](https://dotdata.com/blog/practical-guide-for-feature-engineering-of-time-series-data/) external factors (oil prices, holidays) [dotData](https://dotdata.com/blog/practical-guide-for-feature-engineering-of-time-series-data/)
* **Độ khó**: Trung bình đến Nâng cao
* **Phù hợp cho sinh viên**: Xuất sắc cho học demand forecasting và time series

### **Dataset 2: Walmart Store Sales Forecasting**

* **Nguồn**: Kaggle
* **Mô tả**: Historical sales data của Walmart stores
* **Kích thước**: Nhiều năm weekly sales data
* **Đặc điểm chính**: Economic indicators included (Temperature, Fuel Price, CPI)
* **Độ khó**: Trung bình
* **Phù hợp cho sinh viên**: Industry-standard forecasting problem

### **Dataset 3: Rossmann Store Sales**

* **Nguồn**: Kaggle
* **Mô tả**: Historical sales từ Rossmann drug stores
* **Kích thước**: 1,017,209 training records, 41,088 test records
* **Đặc điểm chính**: European retail data, store-specific factors, promotions
* **Độ khó**: Trung bình
* **Phù hợp cho sinh viên**: Tốt cho retail forecasting và time series analysis

## **9. Kiểm Soát Chất Lượng & Phát Hiện Bất Thường (Quality Control & Anomaly Detection)**

### **Dataset 1: Severstal Steel Defect Detection**

* **Nguồn**: Kaggle Competition
* **Link tải**:<https://www.kaggle.com/c/severstal-steel-defect-detection>
* **Mô tả**: Phát hiện và phân loại defects trên steel sheets
* **Kích thước**: Hàng nghìn high-resolution steel surface images
* **Đặc điểm chính**: 4 defect classes, segmentation masks, industrial-grade images
* **Độ khó**: Nâng cao - cần computer vision và deep learning
* **Phù hợp cho sinh viên**: Projects nâng cao với computer vision

### **Dataset 2: CWRU Bearing Dataset**

* **Nguồn**: Case Western Reserve University
* **Link tải**:<https://engineering.case.edu/bearingdatacenter>
* **Mô tả**: Accelerometer data từ motor bearings bình thường và lỗi [GitHub](https://github.com/srigas/CWRU_Bearing_NumPy)[Case](https://engineering.case.edu/bearingdatacenter)
* **Kích thước**: 161 records với nhiều bearing conditions [Medium](https://medium.com/@NameerAkhter/all-you-need-to-know-about-cwru-dataset-8d391577d8f2)
* **Đặc điểm chính**: 3 fault types, multiple fault sizes, [GitHub](https://github.com/srigas/CWRU_Bearing_NumPy)[Case](https://engineering.case.edu/bearingdatacenter) time-series vibration data [GitHub](https://github.com/srigas/CWRU_Bearing_NumPy)
* **Độ khó**: Trung bình
* **Phù hợp cho sinh viên**: Tốt cho time-series analysis và anomaly detection

### **Dataset 3: Manufacturing Quality Control Datasets Repository**

* **Nguồn**: GitHub
* **Link tải**:<https://github.com/WenjinTao/Manufacturing-Datasets>
* **Mô tả**: Bộ sưu tập manufacturing datasets cho quality control
* **Kích thước**: Nhiều datasets với kích thước khác nhau
* **Đặc điểm chính**: AITEX textile defects, PCB defects, magnetic tile defects
* **Độ khó**: Cơ bản đến Nâng cao
* **Phù hợp cho sinh viên**: Đa dạng options cho mọi skill levels

## **10. Phân Tích Dữ Liệu Khí Hậu (Climate Data Analysis)**

### **Dataset 1: NASA GISS Temperature Data (GISTEMP v4)**

* **Nguồn**: NASA Goddard Institute for Space Studies [nasa](https://data.giss.nasa.gov/gistemp/)
* **Link tải**:<https://data.giss.nasa.gov/gistemp/>
* **Mô tả**: Global surface temperature anomaly analysis [nasa](https://data.giss.nasa.gov/gistemp/)
* **Kích thước**: Monthly data từ 1880-hiện tại, 2°×2° resolution [nasa](https://data.giss.nasa.gov/gistemp/)
* **Đặc điểm chính**: Temperature anomalies, global/hemispheric/zonal time series [nasa](https://data.giss.nasa.gov/gistemp/)
* **Độ khó**: Trung bình
* **Phù hợp cho sinh viên**: Xuất sắc cho time series và climate trend studies

### **Dataset 2: NOAA Global Surface Temperature Dataset**

* **Nguồn**: NOAA National Centers for Environmental Information
* **Mô tả**: Blended land-ocean surface temperature analysis
* **Kích thước**: Monthly gridded data (5°×5°) từ 1850-hiện tại
* **Đặc điểm chính**: Combines SST và land air temperature
* **Độ khó**: Trung bình đến Nâng cao
* **Phù hợp cho sinh viên**: Tốt cho advanced climate variability studies

### **Dataset 3: Kaggle Climate Change Temperature Dataset**

* **Nguồn**: Kaggle
* **Mô tả**: Processed climate data với temperature, precipitation, emissions
* **Kích thước**: Thường 1750-2023 với nhiều variables
* **Đặc điểm chính**: Pre-processed data, country-level aggregations
* **Độ khó**: Cơ bản đến Trung bình
* **Phù hợp cho sinh viên**: Hoàn hảo cho beginners trong climate data analysis