Project Number: 31

Project Title: Ontology-Enhanced PCA Models Using Eurofidai ESG Dataset

Project Clients: Mingqin Yu

Project specializations: Software Development;Computer Science and Algorithms;Big data Analytics and Visualization;

Number of groups: 3 groups

Main contact: Mingqin Yu

Background:

The integration of ontology-driven methodologies with Principal Component Analysis (PCA) can significantly enhance ESG metric interpretation and risk assessment. The Eurofidai ESG dataset, containing 105 raw risk data metrics across global companies, provides an ideal dataset for this study. The primary objective of this project is to develop an ontology-enhanced PCA model that improves data structuring, dimensionality reduction, and interpretability in ESG analysis.

This project aligns with the ontology-driven architecture proposed by Yu, Rabhi, & Bandara (2024), which emphasizes semantic standardization and structured ESG knowledge representation. Through this study, students will gain hands-on experience with machine learning, ontology engineering, and sustainability analytics.

Requirements and Scope:

2.1 Scope

Phase 1: ESG Data Preprocessing and Ontology Integration

This phase focuses on structuring the Eurofidai ESG dataset using ontology-based ESG categorization and preparing the data for PCA analysis.

Task 1: Data Exploration

1) Load and explore the Eurofidai ESG dataset, identifying key features, missing values, and anomalies.

2) Conduct summary statistics and visual analysis to understand metric distributions and data variability.

Task 2: Data Preprocessing and Normalization

1) Handle missing values using data imputation techniques.

2) Standardize numerical ESG metrics to ensure comparability across dimensions.

3) Convert categorical ESG features into ontology-compatible representations.

Task 3: Ontology Integration

1) Utilize ESG ontology models from Yu et al. (2024) to categorize ESG metrics into Environmental, Social, and Governance (E, S, G) groups.

2) Use SPARQL queries to retrieve structured ESG data and enrich metric relationships.

3) Map ESG indicators to ontology-based feature hierarchies for improved interpretability.

Phase 2: Principal Component Analysis (PCA) for ESG Data

This phase applies PCA to extract the most relevant ESG features while integrating ontology-based knowledge to enhance explainability.

Task 4: Applying PCA

1) Compute the covariance matrix and extract eigenvalues & eigenvectors.

2) Determine principal components (PCs) that retain the most variance in ESG data.

3) Reduce the dimensionality of ESG metrics, ensuring minimal information loss.

Task 5: PCA Interpretation & Feature Engineering

1) Visualize PCA results using scree plots, biplots, and heatmaps.

2) Interpret the significance of each principal component (PC) in ESG risk assessment.

3) Compare traditional PCA vs. ontology-enhanced PCA models for interpretability and feature selection.

Task 6: ESG Risk Factor Analysis

1) Identify which ESG categories contribute most to principal components.

2) Link PCA-extracted ESG factors to corporate sustainability risks and decision-making.

3) Validate findings using external ESG reports and industry benchmarks.

Phase 3: Model Implementation, Evaluation, and Reporting

This phase implements, evaluates, and presents the ontology-enhanced PCA model through an interactive dashboard and technical report.

Task 7: Model Validation and Comparison

1) Compare ontology-enhanced PCA with traditional PCA to assess model accuracy and interpretability.

2) Evaluate performance using variance explained, clustering accuracy, and ESG metric influence.

Task 8: ESG Dashboard

1) Develop a visual dashboard (Dash, Streamlit, or Tableau) to present key PCA-driven ESG insights.

Task 9: Final Presentation and Documentation

1) Prepare a presentation summarizing project outcomes and key findings.

2) Submit documented code, datasets, and results in a reproducible format.

Required Knowledge and skills:

The system will be developed as a data processing and ESG analytics platform, integrating ontology-based knowledge representation with PCA-driven feature selection and dimensionality reduction. The following features and specifications define the software/system requirements:

A. ESG Data Processing and Ontology Integration

- Feature Extraction

1) Load and preprocess Eurofidai ESG dataset from multiple reporting sources.

2) Handle missing data and ensure data consistency before PCA transformation.

- Ontology-Based ESG Categorization

1) Assign ESG metrics to predefined ontology-based categories (E, S, G).

2) Retrieve structured ESG relationships using SPARQL queries from the ontology knowledge base.

3) Link ESG metrics to industry-specific sustainability concepts.

- Data Transformation for PCA

1) Convert categorical ESG attributes into ontology-enriched numerical values.

2) Standardize ESG risk metrics for PCA processing.

B. PCA-Based ESG Feature Selection and Risk Assessment

- PCA Model Implementation

1) Compute eigenvalues, eigenvectors, and explained variance for principal component selection.

2) Apply dimensionality reduction while maintaining ESG interpretability.

- Visualization & Interpretability

1) Generate heatmaps, biplots, and component contribution charts for ESG insights.

2) Link PCA results to ontology concepts to enhance explainability.

3) Implement anomaly detection using PCA to identify ESG reporting inconsistencies.

C. ESG Dashboard

- Interactive Dashboard

1) Develop a real-time visualization tool for PCA-based ESG analytics.

2) Display principal components, risk scores, and ESG trend analysis.

Expected outcomes/deliverables:

1) Preprocessed ESG dataset with ontology-enhanced feature categorization.

2) PCA-based ESG factor analysis, identifying key sustainability risk drivers.

3) Ontology-driven feature engineering applied to ESG data.

4) Python-based implementation, including Jupyter Notebook documentation.

5) Comparative analysis of traditional PCA vs. ontology-enhanced PCA models.

6) Final project report & presentation, detailing methodology, findings, and business applications.