CareerMatch AI

Detailed Data Mining Project Documentation

Version 2.1 | January 2026

1. GENERAL OVERVIEW

CareerMatch AI is an intelligent analytics platform that uses Machine Learning and NLP (Natural Language Processing) to analyze the compatibility between CVs and job postings.

Main Objectives

- Calculate a match score between CV and Job Description
- Identify missing and transferable skills
- Provide a personalized learning path
- Suggest alternative roles based on candidate profile

Live Demo: https://dataminingiulm.streamlit.app/

Repository: https://github.com/Giacomod2001/datamining

2. APPLICATION ARCHITECTURE

The application follows a 3-tier architecture:

Frontend (app.py)

- Interactive Streamlit Dashboard
- Plotly Visualizations (gauge, scatter, bar charts)
- Premium CSS with glassmorphism

Backend (ml_utils.py)

- Random Forest Classifier for skill matching
- K-Means & Hierarchical Clustering
- LDA Topic Modeling
- Named Entity Recognition (NER)

Knowledge Base (constants.py)

- Hard Skills with synonyms and variants
- Inference Rules (skill -> related skills)
- Skill Clusters (tool equivalences)
- Job Archetypes for Career Compass

3. DATA MINING TECHNIQUES

Knowledge Discovery Process (KDD)

- Step 1: Data Cleaning - Text preprocessing, tokenization
- Step 2: Data Integration - Merge CV + JD + Portfolio
- Step 3: Data Selection - Relevant section extraction
- Step 4: Data Transformation - TF-IDF Vectorization, N-grams
- Step 5: Data Mining - Classification, Clustering, Topic Modeling
- Step 6: Pattern Evaluation - Match score, confidence calculation
- Step 7: Knowledge Presentation - Dashboard, PDF reports

Machine Learning Algorithms

Random Forest Classifier: 150 trees, max_depth=15, class_weight=balanced

K-Means Clustering: n_clusters=max(2, min(N/3, 5)), elkan algorithm

Hierarchical Clustering: Ward linkage, dendrogram visualization

LDA Topic Modeling: 3-5 topics, batch learning

PCA: 2D dimensionality reduction for visualization

Text Mining Techniques

- TF-IDF Vectorization with sublinear TF
- N-gram Analysis (unigram, bigram, trigram)
- Fuzzy String Matching (Levenshtein, 85% threshold)
- Named Entity Recognition with NLTK

4. APPLICATION FEATURES

Core Features

- Match Score: CV-JD compatibility (0-100%)
- Transferable Skills Recognition via SKILL_CLUSTERS
- Gap Analysis with skill prioritization
- Cover Letter Analysis (structure, personalization)
- Project Portfolio Evaluation
- Career Compass with alternative roles
- Learning Path with course links
- PDF Report Export

CV Builder (NEW v2.1)

- 4-step guided workflow: Profile, Skills, Experience, Export
- Real-time JD optimization scoring
- Load Demo / Exit Demo controls
- PDF/TXT export with professional formatting
- AI suggestions for missing skills

Demo Mode

Sample data optimized for ~93% match (13/14 skills)

5. TEAM & CREDITS

Authors

- Giacomo Dellacqua - Project Design (UI/UX & Architecture)
- Luca Tallarico - Machine Learning & NLP/Text Mining
- Ruben Scoletta - Testing, QA & Documentation

AI Tools Used

- Claude Opus 4 (thinking) - Primary AI assistant
- Gemini 3 Pro High - Debugging support
- Antigravity - Agentic development support

Project Metrics

Total lines of code: ~6000+

ML algorithms: 5 (RF, K-Means, Hierarchical, LDA, PCA)

Text Mining techniques: 4 (TF-IDF, N-gram, Fuzzy, NER)

Stop words languages: 5 (EN, IT, ES, FR, DE)

Report updated: January 11, 2026

CareerMatch AI v2.1 - IULM University