CareerMatch AI - Project Report

Detailed Data Mining Project Documentation

Project: CareerMatch AI
Version: 2.1
Institution: IULM University - Academic Year 2025-2026
Course: Data Mining & Text Analytics
Professor: Prof. Alessandro Bruno
License: PolyForm Noncommercial License 1.0.0

Live Demo: https://dataminingiulm.streamlit.app/
Repository: https://github.com/Giacomod2001/datamining

Report Date: January 11, 2026

---

1. Executive Summary

CareerMatch AI is an intelligent analytics platform that leverages Machine Learning and Natural Language Processing (NLP) to evaluate the compatibility between candidate CVs and job descriptions. The system provides actionable insights through match scoring, skill gap analysis, and personalized career recommendations.

Primary Objectives

The platform is designed to address four key challenges in the recruitment and career development process:

- **Match Score Calculation:** Quantify the alignment between a candidate's CV and job description on a 0-100% scale
- **Skill Gap Identification:** Detect missing skills while recognizing transferable competencies from the candidate's background
- **Learning Path Generation:** Provide personalized course recommendations to address identified skill gaps
- **Career Guidance:** Suggest alternative career roles that align with the candidate's existing profile

---

2. System Architecture

CareerMatch AI implements a modular three-tier architecture that separates presentation, business logic, and data management concerns.

Frontend Layer (app.py)

The user interface is built using Streamlit and provides an interactive dashboard experience:

- **Interactive Dashboard:** Real-time analysis and results visualization
- **Data Visualizations:** Plotly-based charts including gauge meters, scatter plots, and bar charts
- **Design System:** Premium CSS styling with glassmorphism effects for modern aesthetics

Backend Layer (ml_utils.py)

The core analytical engine implements multiple machine learning and text mining algorithms:

- **Random Forest Classifier:** Supervised learning model for skill matching predictions
- **K-Means Clustering:** Unsupervised learning for grouping similar skills and profiles
- **Hierarchical Clustering:** Agglomerative clustering for skill taxonomy
- **LDA Topic Modeling:** Latent Dirichlet Allocation for thematic analysis
- **Named Entity Recognition (NER):** Extraction of key entities from unstructured text

Knowledge Base Layer (constants.py)

A structured repository of domain knowledge that powers the system's reasoning capabilities:

- **Hard Skills Dictionary:** Comprehensive skill database with synonyms and variants
- **Inference Rules:** Logical relationships mapping skills to related competencies
- **Skill Clusters:** Tool equivalence groups (e.g., similar technologies or methodologies)
- **Job Archetypes:** Predefined career profiles for the Career Compass feature

---

3. Data Mining Methodology

Knowledge Discovery in Databases (KDD) Process

The application implements a complete seven-stage KDD pipeline:

1. **Data Cleaning:** Text preprocessing including normalization, tokenization, and noise removal
2. **Data Integration:** Consolidation of multiple data sources (CV, job description, portfolio)
3. **Data Selection:** Extraction of relevant sections and features from integrated data
4. **Data Transformation:** Feature engineering through TF-IDF vectorization and N-gram extraction
5. **Data Mining:** Application of classification, clustering, and topic modeling algorithms
6. **Pattern Evaluation:** Calculation of match scores and confidence intervals
7. **Knowledge Presentation:** Generation of dashboard visualizations and PDF reports

Machine Learning Algorithms

The system employs five complementary machine learning techniques:

Random Forest Classifier

- Configuration: 150 decision trees, maximum depth of 15 levels
- Class balancing: Weighted to handle imbalanced skill distributions
- Purpose: Primary classification model for skill matching

K-Means Clustering

- Dynamic cluster selection: min(N/3, 5) clusters, minimum of 2
- Algorithm: Elkan variant for improved computational efficiency
- Purpose: Grouping similar candidate profiles and skills

Hierarchical Clustering

- Linkage method: Ward's minimum variance method
- Visualization: Dendrogram representation of skill hierarchies
- Purpose: Understanding skill relationships and taxonomy

Latent Dirichlet Allocation (LDA)

- Topic range: 3-5 latent topics per document
- Learning mode: Batch learning for consistency
- Purpose: Thematic analysis of job descriptions and CVs

Principal Component Analysis (PCA)

- Dimensionality: Reduction to 2D space
- Purpose: Data visualization and pattern exploration

Text Mining Techniques

Four specialized text mining approaches extract insights from unstructured content:

TF-IDF Vectorization

- Weighting: Sublinear term frequency normalization
- Purpose: Converting text to numerical feature vectors

N-gram Analysis

- Range: Unigrams, bigrams, and trigrams
- Purpose: Capturing multi-word phrases and skill combinations

Fuzzy String Matching

- Algorithm: Levenshtein distance calculation
- Threshold: 85% similarity for skill matching
- Purpose: Handling spelling variations and synonyms

Named Entity Recognition

- Framework: NLTK-based entity extraction
- Purpose: Identifying key entities (companies, technologies, certifications)

---

4. Feature Set

Core Analytical Features

Match Score Engine

- Calculates CV-to-job-description compatibility as a percentage (0-100%)
- Considers both exact and fuzzy skill matches

Transferable Skills Recognition

- Identifies relevant skills from different domains using SKILL_CLUSTERS
- Maps equivalent tools and technologies

Gap Analysis

- Prioritizes missing skills based on job requirements
- Distinguishes critical vs. nice-to-have competencies

Cover Letter Analyzer

- Evaluates structure, clarity, and personalization
- Provides improvement recommendations

Project Portfolio Evaluation

- Assesses quality and relevance of candidate projects
- Identifies demonstrated competencies

Career Compass

- Recommends alternative career paths based on existing skills
- Uses job archetype matching

Learning Path Generator

- Curates personalized course recommendations
- Provides direct links to learning resources

PDF Report Export

- Generates comprehensive analysis reports
- Professional formatting for sharing with stakeholders

CV Builder Module (Version 2.1)

A newly introduced guided workflow for CV creation and optimization:

Four-Step Process

1. Profile information entry
2. Skills inventory
3. Experience documentation
4. Export and finalization

Key Capabilities

- Real-time optimization scoring against target job descriptions
- Demo mode controls (Load Demo / Exit Demo)
- Multiple export formats (PDF/TXT)
- Professional formatting templates
- AI-powered suggestions for skill enhancement

Demonstration Mode

Pre-configured sample data for immediate platform testing:

- **Target Score:** Approximately 93% match (13 out of 14 skills matched)
- **Sample CV Profile:** Marco Bianchi, Marketing Data Analyst
- **Sample Job Description:** Senior Marketing Analyst position
- **Sample Project:** E-commerce Analytics Dashboard

---

5. Technical Specifications

Project Metrics

- **Codebase Size:** 6,000+ lines of code
- **Machine Learning Algorithms:** 5 implemented (Random Forest, K-Means, Hierarchical Clustering, LDA, PCA)
- **Text Mining Techniques:** 4 implemented (TF-IDF, N-gram, Fuzzy Matching, NER)
- **Language Support:** 5 languages for stop word filtering (English, Italian, Spanish, French, German)

Development Team

Authors:

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

AI Development Tools:

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

---

6. Conclusion

CareerMatch AI represents a comprehensive application of data mining and machine learning techniques to solve real-world challenges in recruitment and career development. Through its combination of supervised and unsupervised learning, advanced text mining, and a knowledge-driven approach, the platform delivers actionable insights for both job seekers and recruiters.

The system's modular architecture and extensive feature set demonstrate practical implementation of academic concepts from the Data Mining & Text Analytics course at IULM University.

CareerMatch AI v2.1 - IULM University, Academic Year 2025-2026