CAREER & SKILLS INTELLIGENCE SYSTEM

CRISP-DM Data Analysis Report

Data Science Project Report

Executive Summary

This report documents the end-to-end data science process behind the Career & Skills Intelligence Recommendation System, a multi-component AI platform designed to help learners and professionals across the globe discover suitable career paths, understand their skill gaps, assess automation risk, and receive personalised learning roadmaps.

The project follows the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology across six structured phases: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. Each phase is documented in detail in the sections below, drawing directly from the seven Jupyter notebooks that constitute the project codebase.

Metric	Value
Total Occupations Covered	894 (O*NET)
Total Courses in Catalogue	8,050 (Coursera + Udemy + edX)
LinkedIn Job Postings Processed	119,320
Skill Dimensions	35 standardised O*NET dimensions
User Types Supported	6 (CBC, 8-4-4, Graduate, Postgrad, TVET, Professional)
Career Families	12 (Tech, Healthcare, Engineering, Finance, etc.)
Pipeline Notebooks	7 (Data → Recommendation → Gap → Risk → Courses → Integration)
Primary ML Model	Two-stage: NearestNeighbors (retrieval) + GradientBoostingClassifier (ranking)

Phase 1: Business Understanding

1.1 Project Background & Objectives

The Career & Skills Intelligence System was conceived to address a critical challenge in the education and labour market: the disconnect between what learners study and what the job market demands. The platform targets students graduating under both the legacy 8-4-4 curriculum and the new Competency-Based Curriculum (CBC), as well as TVET diploma holders, graduates, and working professionals seeking career transitions.

Primary Business Goals

- Provide personalised career recommendations aligned to a user's skills, educational background, and market demand.
- Quantify the risk of AI/automation replacing recommended careers, enabling future-proof decision-making.
- Identify skills gaps between a user's current profile and target career requirements.
- Recommend structured learning paths (courses) to close those gaps using globally available MOOCs.
- Incorporate Kenyan context — mapping CBC pathways and KCSE subjects to international O*NET occupational standards.

1.2 Success Criteria

Criterion	Target	Measurement Approach
Career Relevance	Top-5 recommendations meaningful to user	User-type zone alignment, semantic similarity
Skill Alignment	Skill match score > 50% for top recommendation	Cosine similarity on O*NET skill vectors
AI Risk Accuracy	Risk distribution plausible across sectors	BLS data concordance + skill-based cross-validation
Course Coverage	>80% of identified gaps have matching courses	TF-IDF match rate across 8,050 courses
Pipeline Latency	Full recommendation in < 10s per user	End-to-end timing in Notebook 07

1.3 Constraints & Assumptions

- BLS employment projection data is limited to 102 occupations; risk scores for remaining 792 occupations are derived from O*NET skill-based modelling.
- LinkedIn postings dataset (119,320 records) does not include parsed skill abbreviations (skills_abr column absent), limiting direct skill-demand linking.

- The system assumes users can accurately self-report skills in natural language; the synonym map (250+ entries) and semantic embedding partially compensate for vocabulary variability.
- Course duration data was unavailable in the unified catalogue, preventing precise learning-time estimates.

Phase 2: Data Understanding

2.1 Dataset Architecture

The system integrates eleven distinct datasets spanning occupational standards, labour market signals, educational content, and Kenya-specific curriculum mappings.

Dataset	Source	Records	Role in System
O*NET Occupation Data	O*NET / BLS	1,016	Career backbone — 894 unique occupations used in modelling
O*NET Skills	O*NET	62,580 rows	35 standardised skill dimensions; Level (LV) scores 0–6 per occupation
O*NET Job Zones	O*NET	923	Maps education levels (Zones 1–5) to occupations
O*NET Education & Training	O*NET	37,125 rows	Required education level per occupation (878 occupations covered)
BLS Employment Projections	Bureau of Labour Statistics	102	Demand level, employment change, automation risk — 12 career families
LinkedIn Job Postings	LinkedIn	119,320	Real-world demand signal; job title frequency as posting_demand_norm
edX Courses	edX	975	MOOC catalogue — quality and skill coverage signals
Udemy Courses	Udemy	3,678	MOOC catalogue — broad coverage, price point variety
Coursera Courses	Coursera	3,404	MOOC catalogue — highest quality signal via review data
Coursera Reviews	Coursera	107,018	Sentiment data for course quality scoring
CBC Pathways	Kenya MoE	22 pathways	Maps Kenya CBC tracks to O*NET job zones and implicit skills

2.2 Initial Data Exploration Findings

O*NET Occupations

The dataset contains 1,016 rows, reducing to 894 unique occupations after accounting for SOC code variants. No missing values or duplicates were detected. The dataset spans all major occupational groups as defined by the US Standard Occupational Classification (SOC) system, providing comprehensive coverage for career mapping.

O*NET Skills — 35 Standardised Dimensions

62,580 skill-occupation rows with 15 columns. Level (LV) score scale runs from 0.00 to 6.00. Key skill dimensions include Active Learning, Critical Thinking, Complex Problem Solving, Mathematics, Programming, Social Perceptiveness, and Systems Analysis. The pivot operation (LV scores only) produces a 894 × 37 feature matrix (occupation + 35 skills).

Skill Category	Representative Dimensions	Relevance
Cognitive	Critical Thinking, Active Learning, Complex Problem Solving	Universal across most career families
Technical	Programming, Mathematics, Equipment Operation, Installation	Key differentiators for Tech & Engineering
Interpersonal	Social Perceptiveness, Coordination, Instructing, Speaking	Healthcare, Education, Management
Physical	Equipment Maintenance, Operations Monitoring, Quality Control	Manufacturing, Agriculture, Transport
Creative	Originality, Active Listening, Judgment & Decision Making	Arts, Law, Leadership roles

Job Zones — Education-Occupation Mapping

923 occupation-zone mappings across five education zones. Zone distribution is heavily weighted toward Zones 3–4, corresponding to TVET/Diploma (Zone 3) and Graduate/Bachelor's degree (Zone 4) holders, which aligns with the primary target user groups in the Kenyan context.

Job Zone	Description	Kenya User Type	Occupations
Zone 1	Little preparation needed	N/A	Minimal
Zone 2	Some preparation	CBC / 8-4-4 High School	Low
Zone 3	Medium preparation	TVET / Diploma	Moderate
Zone 4	Considerable preparation	Graduates (Bachelor's)	High
Zone 5	Extensive preparation	Postgraduates / Professionals	Moderate

BLS Employment Projections

102 occupations with 12 fields covering employment figures for 2022 and 2032, automation risk probability, demand level classification (Low/Medium/High), and career family assignment across 7 families: Technology (17), Engineering (17), Healthcare (17), Business and Finance (18), Arts and Media (8), Education (6), and Law and Public Service (6).

LinkedIn Job Postings

119,320 postings loaded in memory-efficient chunks (50,000 rows/chunk). After processing, 71,580 unique job titles were identified. The top demand roles were Customer Service Representative (449 postings), Sales Manager (438), and Project Manager (358), providing a normalised posting_demand_norm signal used in career scoring.

Course Catalogues

The three platforms contribute a combined 8,050 courses. Udemy dominates by volume (3,672), followed by Coursera (3,404), and edX (974). All 8,050 records have complete data for the key fields: course_title, subject, skills_covered, platform, level, quality_score, and url. Duration hours were unavailable across all platforms. Coursera reviews (107,018 records) power quality scoring through sentiment-weighted ratings.

Phase 3: Data Preparation

3.1 Cleaning Strategy

Data cleaning was performed in Notebook 02, applying a standardised audit function (clean_and_audit) across all datasets. The cleaning pipeline applied the following transformations:

- Column standardisation: lowercased, stripped, spaces replaced with underscores across all dataframes.
- Text columns: NaN values filled with "Unknown", stripped of whitespace.
- Numeric columns: NaN values replaced with column median.
- Duplicate removal: applied to course datasets (Udemy: 6 exact duplicates removed; O*NET datasets: zero duplicates).
- Outlier detection: flagged (not removed) for course datasets — Udemy had 1,533 flagged; Coursera 783; Reviews 4,720. Outliers were retained as they represent valid high-engagement courses.

Cleaning Audit Summary

Dataset	Raw Rows	Cleaned Rows	Nulls Resolved	Dupes Removed	Outliers Flagged
O*NET Occupations	1,016	1,016	0	0	0
O*NET Skills	62,580	62,580	0	0	0
O*NET Job Zones	923	923	0	0	0
O*NET Education	37,125	37,125	0	0	0
BLS Projections	102	102	0	0	0
edX Courses	975	974	780	1	0
Udemy Courses	3,678	3,672	0	6	1,533
Coursera Courses	3,404	3,404	0	0	783
Coursera Reviews	107,018	107,018	0	0	4,720

3.2 Feature Engineering

Master Occupation Profile Table

The central engineered artifact is the Master Occupation Profile Table (master_occupation_profiles.parquet) — a 894 × 58 column feature matrix built by joining five O*NET datasets:

- Skills Pivot: Level scores transposed to wide format — one row per occupation, 35 skill_* feature columns.
- Job Zone Join: Education zone appended via O*NET SOC code key.
- BLS Enrichment: Demand level, employment projections, and automation_risk joined for 102 matched occupations; median/default values used for unmatched occupations.
- Career Family Assignment (SOC Prefix): Fixed deterministic mapping using the first 2 digits of each SOC code — eliminated fuzzy-match errors from the prior BLS title-based approach (e.g., "Security Guards" was incorrectly mapped to Technology).
- LinkedIn Demand Signal: Normalised posting volume (posting_demand_norm) joined by title string matching.

SOC Code Career Family Mapping (Key Design Decision)

Design Decision: SOC-Based Career Family Labels

Prior approach used fuzzy title matching (cutoff=0.55) against BLS data, producing incorrect mappings such as "Security Guards" → Technology and "Shoe Machine Operators" → Technology. The fix uses the first 2 digits of the O*NET SOC code as an authoritative, deterministic family key. This is aligned with the US Bureau of Labor Statistics Standard Occupational Classification system and eliminates all fuzzy-match noise.

SOC Prefix	Career Family	SOC Prefix	Career Family
11	Management	25	Education
13	Business And Finance	27	Arts And Media
15	Technology	29	Healthcare
17	Engineering	33	Law And Public Service
19	Science	41-53	Transport / Construction / Production
21	Social Services	45-47	Agriculture / Construction

Unified Course Catalogue

edX, Udemy, and Coursera were merged into a single 8,050-row unified_courses.parquet with 14 standardised columns: course_title, platform, institution, subject, level, skills_covered, effort, price, url, rating, num_reviews, rating_norm, popularity_norm, quality_score. Course levels were standardised using a keyword map into three tiers: Foundation (5,555 courses, 69%), Intermediate (2,205 courses, 27%), and Advanced (290 courses, 4%).

CBC Pathway Processing

22 Kenya CBC pathways were mapped to implicit O*NET skill scores (boost weight 0.6) and 28 KCSE subjects were similarly encoded (boost weight 0.55). This enrichment ensures that a student who identifies their CBC track (e.g., Pure Science, Drama & Theatre, Home Science, or Agriculture) receives implicit skill boosts that the semantic embedding can leverage — improving recommendation relevance for users who cannot explicitly name formal skills.

Phase 4: Modelling

4.1 Model Architecture Overview

The system implements a four-stage sequential pipeline, where each stage enriches the output of the previous stage. This architecture is described in Notebook 07 (Full System Integration).

Stage	Notebook	Component	Algorithm
Stage 1	NB 03	Career Recommendation Engine	Sentence Transformers (all-MiniLM-L6-v2) + NearestNeighbors (cosine) + GradientBoostingClassifier
Stage 2	NB 05	AI Risk Scoring Engine	Skill-weighted risk formula + GradientBoostingRegressor + BLS blending
Stage 3	NB 04	Skills Gap Analyser	Semantic skill vector matching (cosine similarity) + MinMaxScaler
Stage 4	NB 06	Course Recommender	TF-IDF Vectorizer (15,000 features) + cosine similarity search

4.2 Stage 1: Career Recommendation Engine (Notebook 03)

Two-Stage Semantic Retrieval-Ranking Pipeline

The recommendation pipeline uses a two-stage approach that separates broad recall from precise ranking:

- Stage 1a — Semantic Retrieval: User skills and education background are encoded with the all-MiniLM-L6-v2 Sentence Transformer model (384-dimensional embeddings). A pre-built NearestNeighbors index over all 894 occupation embeddings retrieves the top-50 candidate careers using cosine distance.
- Stage 1b — GBM Ranking: A GradientBoostingClassifier re-ranks the 50 candidates using weighted features: skill match score, BLS/composite demand, AI risk score (inverted), job zone fit, LinkedIn posting demand, and career goal keyword boosts. The top-5 careers are returned.

Skill Enrichment Pipeline

- A 250+ entry skill synonym map normalises free-text user inputs to O*NET canonical terms (e.g., "coding" → "programming", "excel" → "spreadsheet software").
- 22 CBC pathway implicit skills and 28 KCSE subject skill boosts are appended to user queries before embedding, providing the transformer with richer context than raw free-text alone.
- 150+ career goal boost keywords across 15 career families (including Kenyan-specific terms) steer ranking toward user-stated career interests.

4.3 Stage 2: AI Risk Scoring Engine (Notebook 05)

Blended Risk Model

The AI replacement risk model blends two complementary signals:

- BLS Automation Risk: Direct probability from BLS projections data (available for 102 occupations).
- Skill-Based Risk: Derived from O*NET skill profiles using a domain expert taxonomy of high-risk (automatable) and low-risk (human-centric) skill dimensions:
  - High-Risk Skills: Operation & Control, Equipment Maintenance, Operations Monitoring, Quality Control, Mathematics, Equipment Selection — these correlate with routine, automatable task profiles.
  - Low-Risk Skills: Social Perceptiveness, Judgment & Decision Making, Instructing, Negotiation, Complex Problem Solving, Originality, Speaking, Coordination, Service Orientation — correlate with human-judgment and social interaction.

Risk Score Distribution

Risk Category	Threshold	Count	Percentage	Interpretation
Low	0.00 – 0.35	859	96.1%	Career is relatively safe from AI displacement
Medium	0.35 – 0.55	25	2.8%	Some tasks automatable; upskilling recommended
High	0.55 – 0.72	10	1.1%	Significant automation exposure; career pivot advised
Very High	0.72 – 1.00	0	0.0%	Near-full automation risk (none in current dataset)

Note on Risk Distribution

The predominance of Low risk (96.1%) reflects that the O*NET dataset skews toward professional and skilled occupations, which inherently have higher human-judgment content. The risk model's range (0.122 – 0.684) indicates meaningful discrimination despite the skewed distribution. Future versions should validate against empirical automation studies (e.g., Frey & Osborne, McKinsey) to calibrate thresholds.

Future-Proof Score

A composite headline metric (range 29.0 – 76.7, mean 57.0) is computed as: 40% demand signal + 40% AI risk shield (1 - blended_risk) + 20% normalised median wage. This single number is the primary user-facing signal for career desirability.

4.4 Stage 3: Skills Gap Engine (Notebook 04)

The gap engine performs a per-career analysis of the user's skill vector against O*NET occupation requirements, producing three gap categories:

- Strong Skills: User score exceeds occupation requirement — showcased as transferable assets.
- Moderate Gaps: User is close to requirement — targeted micro-courses recommended.
- Critical Gaps: User is significantly below requirement — foundational courses recommended.

Skill matching uses cosine similarity between 384-dimensional Sentence Transformer embeddings of user skill descriptions and O*NET canonical skill dimension descriptions, enabling language-agnostic matching across English and Swahili inputs.

4.5 Stage 4: Course Recommender (Notebook 06)

The course recommender builds a TF-IDF index (15,000-feature vocabulary) over a rich course text field that concatenates: course title (triple-weighted), skills_covered, subject, and standardised level label. Gap-driven search queries each skill gap separately, ensuring course recommendations directly address the user's deficit rather than broadly matching the career title.

Phase 5: Evaluation

5.1 Model Evaluation Approach

Formal hold-out evaluation metrics (precision@k, NDCG, AUC) are not yet reported in the notebooks, as the system does not yet have a labelled evaluation dataset of user-career outcome pairs. The evaluation evidence available from the notebooks is primarily qualitative and structural.

5.2 Qualitative Evaluation Evidence

Career Recommendation Coherence

- SOC-code-based career family fix eliminated semantic misclassification (e.g., Security Guards → Technology, Shoe Machine Operators → Technology). Post-fix, all 894 occupations map to their correct SOC major group.
- The skill synonym map (250+ entries) and semantic embeddings handle colloquial and cross-linguistic inputs — including Swahili skill descriptions and Kenyan-specific terms like "crop rotation" and "litigation."
- Education zone filtering ensures recommendations respect the user's qualification level — a Form 4 leaver (Zone 2) will not receive recommendations requiring postgraduate study (Zone 5).

Risk Score Validation

- Blended risk scores for 894 occupations range 0.122 – 0.684; the mean of 0.162 (Low risk) is consistent with the O*NET occupation profile skewing toward skilled roles.
- Cross-validation against BLS automation probabilities for the 102 BLS-covered occupations provides a ground-truth anchor for the skill-based risk estimates.
- Top future-proof careers (Wind Energy Engineers: 76.7; Animal Scientists: 73.8; Conservation Scientists: ~73) reflect plausible combinations of growing demand and low automation susceptibility.

Course Coverage Assessment

- All 8,050 course records have 100% field completion for the seven key fields (course_title, subject, skills_covered, platform, level, quality_score, url).
- The TF-IDF vocabulary covers 15,000 terms across the unified catalogue, providing broad skill-to-course matching capability.
- Platform diversity ensures multiple price-point options: Udemy (typically paid, broad), edX (institutional, often free to audit), Coursera (quality-scored via reviews).

5.3 Identified Limitations & Risks

Limitation	Impact	Mitigation Status
No labelled evaluation dataset	Cannot report quantitative recommendation quality metrics	Planned — requires user study or expert annotation
skills_abr missing from LinkedIn	Cannot link specific skills to job demand directly	Compensated with title-level posting_demand_norm
BLS covers only 102 occupations	Risk scores for 792 occupations are estimated, not empirical	Skill-based model provides reasonable proxy; BLS expansion needed
No course duration data	Cannot estimate total learning time per path	Future: scrape duration from course URLs
Course outliers retained (Udemy: 1,529)	May surface low-quality courses in recommendations	quality_score filter (min 0.3) applied in course search
Risk distribution skewed (96.1% Low)	Threshold calibration may need adjustment	Recommend validation against Frey & Osborne taxonomy
CBC implicit skills are heuristic	May not perfectly reflect individual student profiles	Synonym map + semantic embedding provides partial compensation

Phase 6: Deployment

6.1 System Integration Architecture

Notebook 07 implements the production-ready integration pipeline, assembling all components into a single recommend_careers() API function that processes a user input dictionary and returns a structured JSON response containing career cards, gap reports, learning paths, and dashboard data.

6.2 Input/Output Specification

User Input Schema

The system accepts a Python dictionary with the following key fields:

- user_type: One of "cbc", "8-4-4", "graduate", "postgraduate", "diploma", "professional".
- skills: Comma-separated free-text skills in any language (e.g., "Python, data analysis, critical thinking").
- soft_skills: Optional interpersonal skill descriptions.
- cbc_pathway: Optional CBC track (e.g., "Pure Science", "Agriculture & Nutrition").
- kcse_subjects: Optional list of studied KCSE subjects for implicit skill enrichment.
- career_goal: Optional free-text career goal (e.g., "I want to work in healthcare technology").

System Output Schema

- Top-5 career recommendation cards with: occupation name, career family, job zone, future-proof score, risk category, skill alignment score, demand level.
- Per-career skills gap report: strong skills, moderate gaps, critical gaps, each with skill name and gap magnitude.
- Per-career learning path: Foundation → Intermediate → Advanced course sequence with platform, quality score, and URL.
- Dashboard data: sector risk summary, demand chart data, transferable skills summary.

6.3 Deployment Considerations

Recommendations

Key Findings

1. The SOC-code-based career family fix is a critical correctness improvement — fuzzy title matching produced semantically wrong career family labels for a non-trivial fraction of occupations.

2. The skill synonym map + Sentence Transformer combination enables robust multi-lingual, multi-vocabulary skill matching, which is essential for a Kenyan context where users may describe skills in Swahili or colloquial terms.
3. The AI risk distribution (96.1% Low) suggests the current O*NET-based skill taxonomy may underestimate automation risk for some categories. Calibration against independent automation research is recommended.
4. BLS data coverage (102/894 occupations = 11.4%) is the single largest data gap, requiring the skill-based risk model to estimate risk for 88.6% of the occupational space.
5. The absence of a formal labelled evaluation set is the highest-priority gap for demonstrating model quality to stakeholders. A user study or expert annotation exercise is recommended as the next milestone.

Recommendations

- Establish a ground-truth evaluation dataset by recruiting 50–100 Kenyan users from each user type to rate the relevance of their top-5 career recommendations.
- Expand BLS risk coverage by integrating the Frey & Osborne (2013) automation probability dataset or the McKinsey automation feasibility scores for broader occupational coverage.
- Implement real-time LinkedIn demand signals via the LinkedIn Job Search API to replace the static batch-processed postings file.
- Add Kenyan-specific job board data (e.g., Brightermonday, MyJobMag) to provide locally-relevant demand signals alongside global O*NET/BLS data.
- Containerise the system (Docker + FastAPI) and deploy to a cloud provider (GCP Cloud Run or AWS Lambda) for production serving.
- Introduce model monitoring dashboards tracking recommendation distribution drift, skill gap coverage rates, and course click-through rates.

End of CRISP-DM Data Analysis Report

Career and skills recommendation system.