Skill-Based Job Role Predictor

Author

Nargiza Mukhametrakhimova

Course: Artificial Intelligence Final Project

Defense Date: 12 May 2025

1. Problem Statement

The goal of this project is to develop a Machine Learning system that can predict the most suitable job role based on a given list of skills and qualifications.

2. Dataset Description

We used a combined dataset consisting of:

- Kaggle dataset (~300,000 rows): 'Job Descriptions and Required Skills'
- Manually created dataset with 10 entries to satisfy the dataset creation requirement.

Examples from the custom dataset:

? 'Python, Flask, REST API' ? Backend Developer

? 'TensorFlow, Python, Pandas' ? ML Engineer

? 'Project Management, Scrum, Jira' ? Project Manager

? 'SQL, ETL, Data Warehousing' ? Data Engineer

Total merged rows: ~300,010

3. AI Technique Used

- Vectorization: TfidfVectorizer (max_features=1000)
- Model: RandomForestClassifier (100 trees)
- Train/Test Split: 80% / 20%

4. Implementation Details

- Language: Python

Skill-Based Job Role Predictor

- Libraries: pandas, scikit-learn, joblib
- Scripts: job_role_predictor.py, merge_datasets.py
- Output: job_predictor_model.joblib, vectorizer.joblib

5. Evaluation & Results

Classification Report:

Accuracy: 1.00

Precision: 1.00

Recall: 1.00

F1 Score: 1.00

Note: High scores indicate strong feature-label correlation and possibly low noise.

6. Challenges & Limitations

- Some job titles are overly specific (e.g., 'Senior Backend Engineer')
- Skill terms are not standardized
- Dataset is skewed toward technical/IT roles

7. Future Improvements

- Use contextual embeddings (e.g., BERT)
- Add skill normalization and ontology
- Provide a web interface for real-time prediction

8. Deliverables

- Code: job_role_predictor.py
- Dataset: combined_dataset.csv (includes 10 custom rows)
- Model: job_predictor_model.joblib
- Poster: included separately
- Video Demo: [insert YouTube or public link here]
- Website: [if applicable, include URL in README]