CareerCompass: Simplifying Career Explorations Using Data Mining

Team Members:

- Hiral Rane (hirane@iu.edu)
- Vaishnavi Shastri (vshastri@iu.edu)
- Palavi Patil (palpatil@iu.edu)

Abstract

CareerCompass is a system designed to make career exploration easier for job seekers using data mining and machine learning. It helps address challenges like confusing job titles, difficulty matching skills to roles, and the time-consuming nature of job searches. Using a dataset of LinkedIn job postings, the system groups similar job titles into five categories: Management, Computer Science, Data Science, Core Engineering, and Finance. It then uses classification models to recommend career paths based on a person's skills and profile, with Logistic Regression achieving the best accuracy of 96.22%. By combining clustering and classification, CareerCompass provides clear, tailored guidance to help individuals find roles that fit their qualifications and aspirations.

Introduction

In today's fast-changing job market, job seekers face the daunting task of navigating a landscape marked by constant evolution and complexity. Finding roles that align with one's unique skills, qualifications, and aspirations often requires considerable effort, especially given the vast number of job titles and the varying requirements associated with positions across industries. These challenges are compounded by a lack of standardization in job titles and the overwhelming variety of skills demanded by different roles, leaving many individuals uncertain about their best career options.

To address these challenges, CareerCompass harnesses advanced data mining and machine learning techniques to guide job seekers toward fulfilling and relevant career paths. By analyzing user profiles - incorporating attributes such as job titles, skills, and geographic location - CareerCompass offers tailored recommendations designed to bridge the gap between individual capabilities and the opportunities available in the job market. This project centers on two core objectives: clustering job skills into standardized, industry-specific categories and building a predictive classification model to suggest suitable career paths. By streamlining the job search and providing actionable insights, CareerCompass empowers users to navigate their career journeys with clarity and confidence, making the process of career exploration more efficient and effective.

Motivation

The development of CareerCompass is inspired by the pressing challenges faced by job seekers in a highly competitive and dynamic job market. The motivation for this project stems from the following key issues:

1. **Diverse and Ambiguous Job Titles**
   The sheer variety of job titles across industries and organizations makes it difficult for job seekers to understand and compare roles effectively. Titles for similar positions often differ significantly, creating confusion and making it challenging to identify the most suitable opportunities.
2. **Skill-to-Role Alignment**
   For many job seekers, identifying roles that match their unique skill sets and career aspirations is a daunting task. Without a clear framework to link their skills to specific roles, individuals risk pursuing opportunities that do not fully leverage their potential or align with their goals.
3. **Time-Intensive Job Search**
   The process of searching for jobs is not only labor-intensive but also time-consuming. Sorting through hundreds of postings to find relevant roles is inefficient and can lead to frustration, particularly for those unsure of where to begin or what roles best suit their qualifications.

CareerCompass aims to overcome these hurdles through an innovative approach that includes:

- **Clustering Job Skills**: Organizing a wide range of job skills into cohesive, standardized groups based on industry norms, enabling easier comparison and understanding of roles.
- **Personalized Career Recommendations**: Utilizing machine learning classification models to analyze job seekers profiles and recommend tailored career paths that align with their skills, preferences, and location.

By addressing these challenges, CareerCompass seeks to transform the career exploration process, offering job seekers a structured, data-driven pathway to discover opportunities that match their unique profiles. This approach not only saves time but also ensures that individuals can focus their efforts on pursuing roles that align with their long-term professional aspirations. With CareerCompass, navigating the complexities of the modern job market becomes a more manageable and empowering experience.

Dataset and EDA:

The dataset (link) used for this project consists of over 1.3 million rows of LinkedIn job postings and skills, focusing on job postings in the United States. After preprocessing, the dataset was filtered to approximately 31,000 rows of job postings specifically in the fields of Science, Finance, and Management. The dataset includes the following key attributes:

- **job_title**: Titles of the job postings (e.g., "Data Analyst"), indicating the role being advertised.
- **company**: Companies offering the jobs, providing insights into hiring organizations.
- **job_location**: Location-based trends, reflecting the geographic distribution of job opportunities.
- **job_skills**: Required skills for the roles, offering valuable data for analyzing skill-role alignment.
- **job_level**: Seniority levels of the roles (e.g., "Entry Level"), helping to contextualize job postings.
- **job_type**: Employment type (e.g., "Full-time"), providing information on the nature of the job arrangement.

This refined dataset serves as the foundation for building the CareerCompass recommendation system, enabling the clustering of job titles and classification of career paths tailored to individual profiles.

Data Preprocessing and EDA:

The dataset underwent a comprehensive preprocessing and cleaning process to ensure consistency, reliability, and relevance for the CareerCompass recommendation system. The key steps performed during the EDA and preprocessing phases are as follows:

1. **Data Cleaning**
   - Handled missing and null values to maintain data completeness.
   - Removed duplicate rows to ensure data consistency and eliminate redundancy.

2. **Data Filtering**
   - Extracted rows corresponding to job postings located in the United States, reducing the dataset to approximately **1.1 million rows**.

- Filtered the dataset to focus on job postings from the **Top 15 cities** with the highest number of job listings, resulting in **~100,000 rows**.
- Further refined the dataset to specifically include job postings in the **Science, Finance, and Management** fields, narrowing it down to approximately **31,000 rows**.

3. **Skill Preprocessing**
   - Converted the **job_skills** column from object type to string format to facilitate preprocessing.
   - Removed special characters, extra spaces, and commas to convert the skills into a clean, unified string format.
   - Removed stop words and standardized the skills to **lowercase** to maintain uniformity.
   - Performed **skill mapping** to eliminate duplicates and ensure consistency in the skill list.

These preprocessing steps provided a clean, structured, and focused dataset that serves as the foundation for clustering job titles and building the classification model to recommend career paths effectively.

Methodology

The CareerCompass project methodology integrates data mining and machine learning techniques to streamline the career exploration process for job seekers. The approach focuses on two key processes: clustering job skills and classification to recommend career paths using defined clusters. These methods ensure an efficient alignment between user skills and relevant job roles, leveraging a comprehensive dataset and advanced algorithms.

Clustering Phase:

The primary objective of clustering was to group similar job skills into cohesive categories based on skill requirements, reducing the complexity of analyzing diverse job titles. Clustering was chosen as the initial step to manage the high variability in job titles and skill sets, ensuring that overlapping and ambiguous roles are categorized into standardized groups. By identifying commonalities in job skills, this process laid the foundation for the subsequent classification phase by simplifying the dataset and enhancing the relevance of recommendations.

1. **TF-IDF Vectorization:**
   ○ The cleaned_job_skills column was transformed into a numerical format using TF-IDF Vectorization. This technique captures the importance of

individual skills across the dataset, enabling meaningful comparisons between job postings.

2. **Dimensionality Reduction:**
   - The high-dimensional TF-IDF data was reduced to two dimensions using Principal Component Analysis (PCA). This step facilitated visualization and ensured that essential patterns in the data were preserved while simplifying complexity.

3. **K-Means Clustering:**
   - The K-Means algorithm grouped job titles into five clusters based on skill similarities. The optimal number of clusters (k=5) was determined using the Elbow Method, which minimizes within-cluster variance.
   - The identified clusters represented broad categories: Management, Computer Science, Data Science, Core Engineering, and Finance.

4. **Cluster Analysis and Visualization:**
   - The clustering results were visualized using a scatter plot, where each point represented a job skill, and colors indicated cluster membership. This visualization provided intuitive insights into how job skills group together.
   - For each cluster, the most common skills were identified, highlighting the core competencies associated with each category.

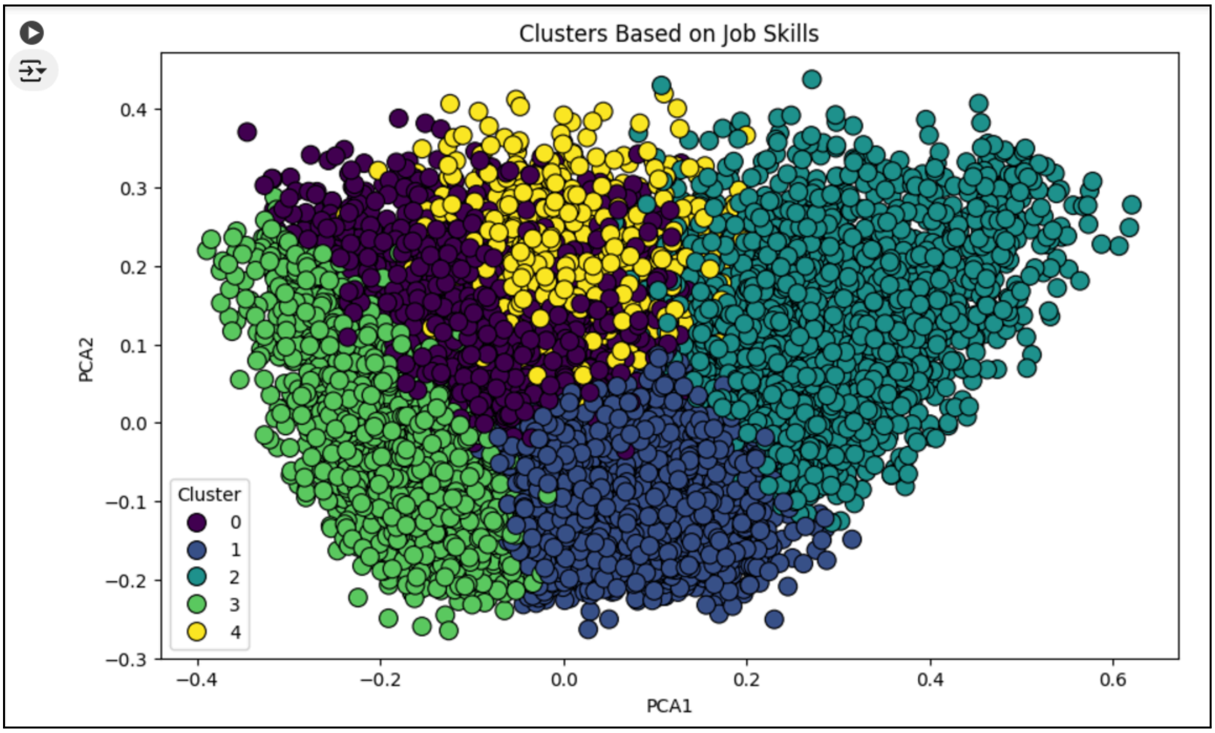


Fig. 1. Clustering based on Job Skills

5. Cluster Characteristics:

Analyzed the clusters to identify their dominant features:

Cluster 0: Management roles.

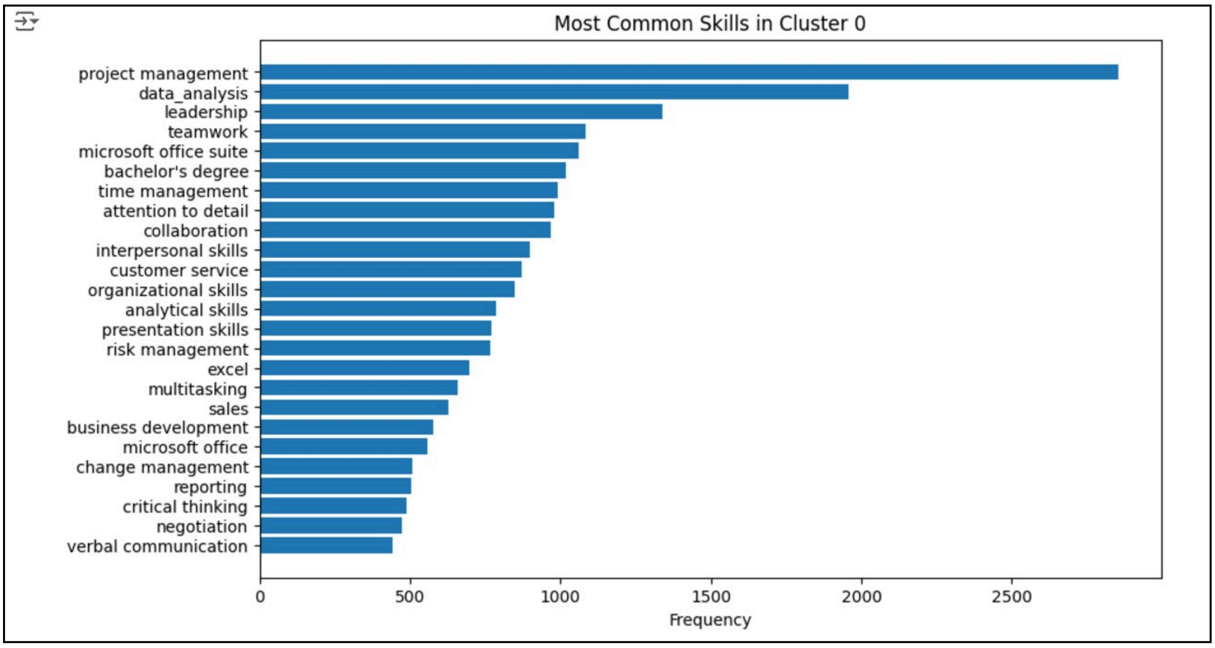


Fig. 2. Management Cluster

Cluster 1: Computer Science roles.

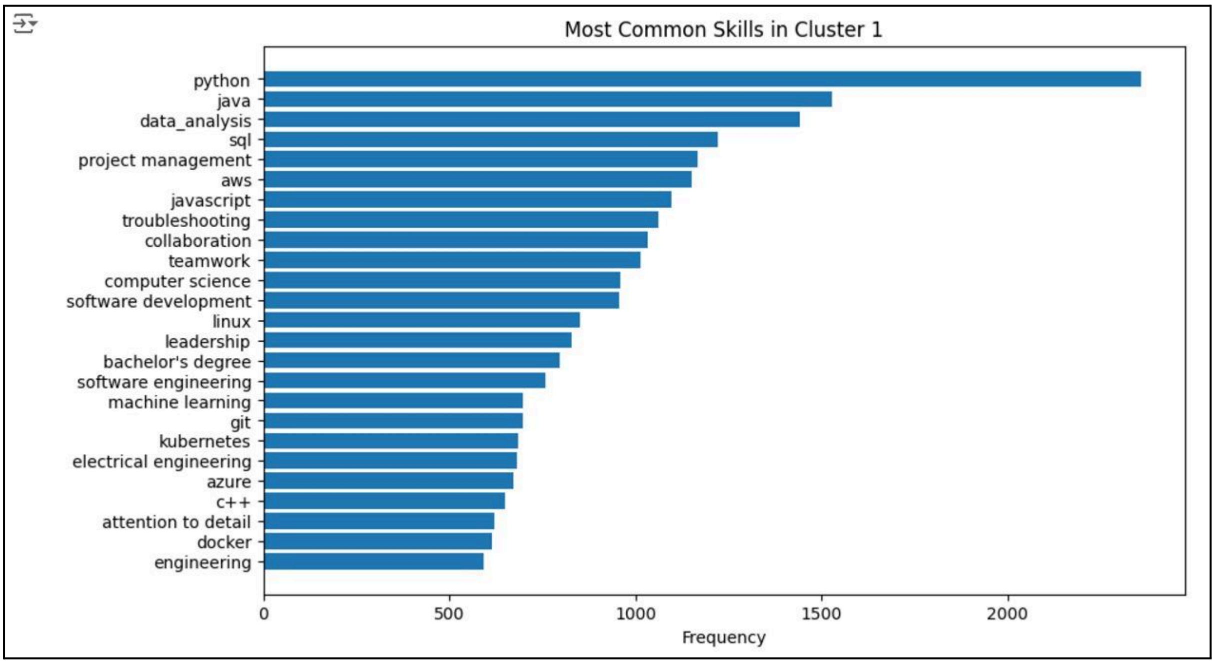


Fig. 3. Computer Science Cluster

Cluster 2: Data Science roles.

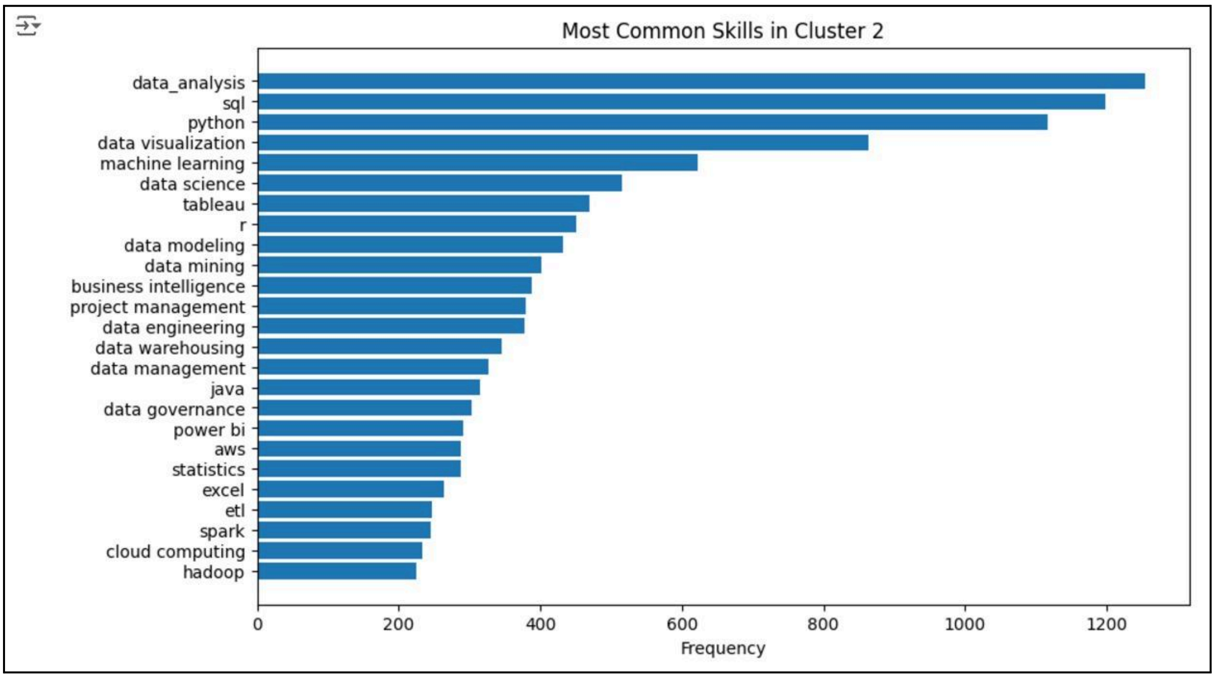


Fig. 4. Data Science Cluster

Cluster 3: Core Engineering roles.

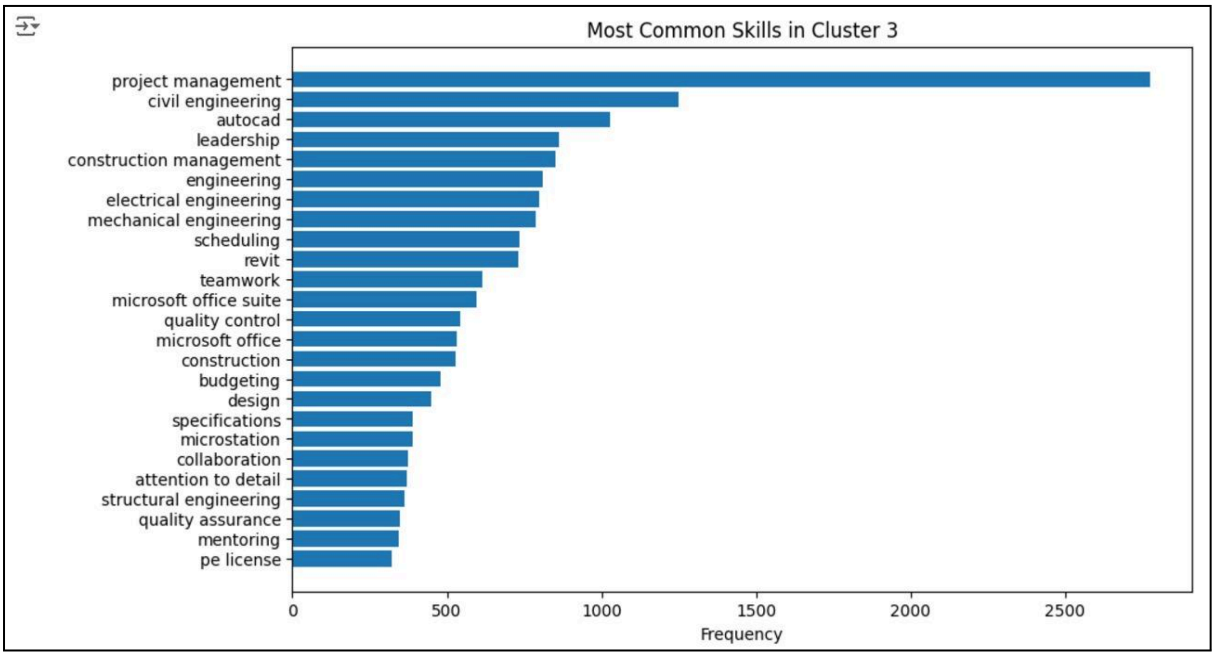


Fig. 5. Core Engineering Cluster

Cluster 4: Finance roles.

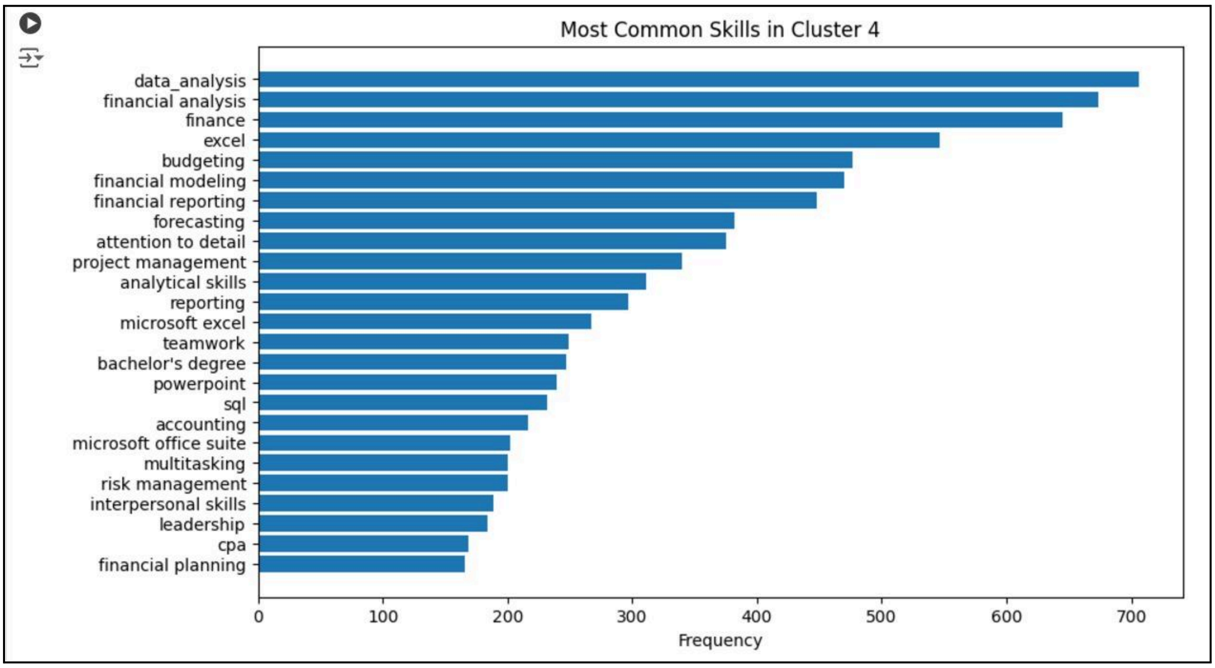


Fig. 6. Finance Cluster

Classification Phase:

Building on the clustering results, the classification phase aimed to predict suitable career paths for job seekers by assigning job skills to predefined categories based on their attributes.

Data Splitting:

Divided the dataset into training (80%) and testing (20%) subsets to evaluate model performance effectively.

1. **Classification Models:**
   - Three models were trained and compared to identify the most effective approach for predicting career paths:
     - **Logistic Regression**: Achieved the highest accuracy (96.22%) and proved effective in handling multiclass classification tasks.
     - **Random Forest**: Provided robustness but had slightly lower accuracy (89.82%).
     - **Decision Tree**: Offered interpretability but showed the least accuracy (83.64%).

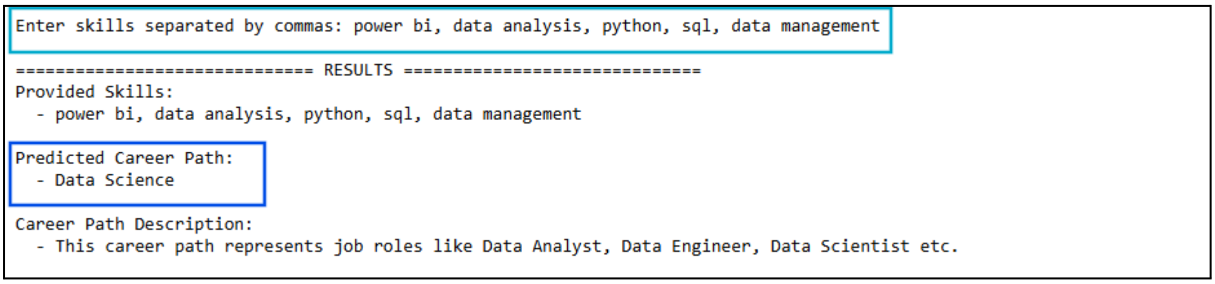
2. **Evaluation and Insights:**

   The models were evaluated using accuracy and detailed classification reports. Logistic Regression outperformed others, making it the primary choice for real-world deployment.

3. **Integration with Clustering:**

   Combined clustering and classification outputs to create a user-friendly recommendation system. Clustering grouped similar skills, while classification predicted the most relevant career path for a given user profile.

Results



```
Enter skills separated by commas: power bi, data analysis, python, sql, data management

============================== RESULTS ==============================
Provided Skills:
  - power bi, data analysis, python, sql, data management

Predicted Career Path:
  - Data Science

Career Path Description:
  - This career path represents job roles like Data Analyst, Data Engineer, Data Scientist etc.
```

- **Clustering Insights:** Visualization of clusters provided clear delineation of common skill sets across industries, aiding the classification process.
- **Model Accuracy:**
  - Logistic Regression: **96.22%**
  - Random Forest: **89.82%**
  - Decision Tree: **83.64%**

Logistic Regression was the most effective classification model for predicting career paths, highlighting its potential in real-world applications.

Conclusion

CareerCompass effectively utilizes data mining techniques to simplify and enhance the job search process for job seekers. By clustering job skills into meaningful categories and applying classification models, the project provides a structured approach to aligning users skills and profiles with relevant career paths.

This approach not only organizes job market data but also empowers job seekers to make informed decisions, reducing the complexity of career exploration.

In summary, CareerCompass demonstrates the value of data-driven solutions in addressing key challenges in the job search process, offering a reliable and practical tool for discovering tailored career opportunities.

Future Scope

1. **Expand Career Paths**: Extend the system to include fields such as Healthcare, Education, and the Public Sector to support a wider range of job seekers.
2. **Enhance Personalization**: Refine classification algorithms to deliver more tailored career recommendations by considering a broader range of user attributes like experience, and goals.
3. **Skill Gap Analysis**: Identify gaps in users skill sets and suggest targeted training or learning opportunities to improve career alignment.
4. **Geographic Insights**: Provide location-specific recommendations based on regional job demand and hiring trends to offer geographically relevant opportunities.

These enhancements will further strengthen CareerCompass, making it a comprehensive and adaptable tool for career guidance.

References

[1] Alsaif SA, Sassi Hidri M, Eleraky HA, Ferjani I, Amami R. Learning-Based Matched Representation System for Job Recommendation. *Computers*. 2022; 11(11):161. https://doi.org/10.3390/computers11110161

[2] Ghosh, A., & Woolf, B. (n.d.). Skill-based career path modeling and recommendation. https://people.umass.edu/~andrewlan/papers/20bigdata-mnss.pdf

[3] Dataset: https://www.kaggle.com/datasets/asaniczka/1-3m-linkedin-jobs-and-skills-2024/data