PREDICTIVE PATHWAYS:AI FOR GRADUATE SUCCESS

WAJD ALHARBI
2211529

RETAL ALJAHDALI
2211685

RENAD ALSHAIK
2210940

TOLEEN ALHARBI
2210751

LISSA HARIRI
2211633

ABSTRACT

This research explores the application of machine learning techniques to classify students learning styles based on their responses to a VAK questionnaire. The study aims to develop a predictive model using the CatBoost algorithm to enhance personalized learning experiences. By analyzing a dataset of 1210 students, the research identifies patterns in learning preferences, which can inform educators in tailoring their teaching strategies. The anticipated outcome is a high-performing model that accurately predicts learning styles, ultimately improving educational outcomes through individualized instruction.

KEY WORDS

Keywords – Systematic review, Web scraper, Machine learning, Career recommendation, Higher education, Predictive modeling

CCS CONCEPTS

Computing methodologies → Machine Learning
Machine learning approaches → Supervised learning by classification
Unsupervised learning by clustering
Applied computing → Education → career recommendation system

1. INTRODUCTION

Many students face significant challenges when choosing their university major, which can lead to a mismatch between their qualifications and the job market after graduation. The problem arises because students often choose their majors based on factors that are not well-informed, without a clear understanding of available job opportunities or market demands. This creates a gap between what students learn at university and what the job market needs, making it difficult for them to find suitable employment after graduation. In light of these challenges, there is a growing need for innovative solutions that help students make more informed educational decisions that align with the evolving demands of the job market.

2. PROBLEM DESCRIPTION

Choosing the right university major and future career is one of the most significant and challenging decisions students face. These decisions are often made without sufficient personalized support, leaving students to rely on generic advice, peer influence, or limited academic indicators. As a result, many students end up pursuing academic paths that do not reflect their interests, strengths, or future career aspirations. This mismatch can lead to low academic performance, decreased motivation, and difficulty finding fulfilling job opportunities after graduation.

Traditional academic and career guidance systems primarily use academic scores, standardized tests, or predefined pathways to offer suggestions. While helpful to a degree, these systems often ignore critical aspects such as soft skills, learning preferences, extracurricular activities, and alignment with real-world job market demands. Students with strong interpersonal abilities or creative problem-solving skills may find themselves steered toward careers that don’t leverage those strengths, simply because such attributes are not factored into the decision-making process.

In recent years, some AI-driven systems have been developed to support educational and career decision-making. However, most existing solutions either focus on predicting academic performance using supervised models or use unsupervised learning to explore job-related clusters. Very few integrate both approaches in a seamless and meaningful way. This leads to fragmented insights that fail to connect students' academic profiles with future job opportunities in a practical, actionable manner.

Furthermore, the labor market is continuously evolving, with new job roles emerging that may not directly map to traditional academic majors. Without adaptive, AI-powered systems that can guide students through both educational and career planning while accounting for such changes, students risk being unprepared for future employment landscapes. There is an urgent need for a holistic, intelligent solution that bridges the gap between students' current abilities and the demands of the modern workforce.

3. LITERATURE REVIEW

Choosing an appropriate university major and career path is a crucial and intricate decision for many students. As highlighted in our problem statement, traditional guidance systems often rely primarily on academic performance, neglecting essential factors like interpersonal skills, personal aspirations, and the ever-evolving job market. This section examines previous studies addressing these concerns, evaluating their advantages, limitations, and the gaps that remain within existing research.

3.1. AI-Based Subject Recommendation Systems

The study presents an AI-powered career guidance system that leverages LSTM neural networks to recommend academic subjects to students. The recommendations are based on students' academic performance and their self-reported aspirations, aiming to provide a personalized pathway for early educational decisions. The model supports students in aligning their academic choices with long-term interests, offering tailored guidance rather than generic recommendations. The study highlights the potential of deep learning models, particularly LSTM, in enhancing the precision and relevance of subject suggestions in educational contexts.

3.2. Systematic Review of Machine Learning in Career Prediction

This study conducted a systematic literature review of 38 academic papers that explored the use of machine learning in educational and career prediction. The aim was to analyse how artificial intelligence is being applied to support personalized academic and career guidance. The review highlights a growing trend in leveraging AI techniques to tailor recommendations based on student profiles. It emphasizes the potential of machine learning to enhance decision-making processes in education by aligning students' characteristics with suitable academic and professional pathways.

3.3. Academic Major Prediction Using Early University Data

In their study titled "Predicting University Students' Academic Success and Major using Random Forests," Piollek and Rosenthal proposed the use of Random Forest classifiers to predict students' academic performance and likely majors based on their early university course selections. The model demonstrated strong predictive capability, highlighting the effectiveness of ensemble learning techniques in analyzing student trajectories and supporting early academic planning based on available academic data.

While these studies offer meaningful advancements, several critical gaps remain that limit the effectiveness of current AI-driven career guidance systems.

For instance, while Balkar et al. (2024) made strides by incorporating student aspirations into their LSTM-based recommendation system, their approach remains heavily weighted toward academic data. Broader attributes—such as soft skills, personal competencies, and alignment with labor market trends—were not considered.

Similarly, Trujillo et al. (2023), in their systematic review, underscored the increasing use of machine learning for educational guidance but also highlighted a persistent gap: most systems fail to connect academic planning with dynamic job market realities. The focus remains largely confined to academic indicators, neglecting to forecast actual career outcomes or respond to emerging job trends.

Piollek and Rosenthal's (2018) work further reflects this pattern. Although their Random Forest model effectively predicted student majors, it relied exclusively on academic records. Personal interests, interpersonal abilities, and extracurricular activities—elements that significantly shape career suitability—were left out of consideration.

Together, these studies point to three major limitations that define the current research landscape:

- Limited integration of supervised and unsupervised learning: Most models either classify or cluster student profiles but seldom combine both approaches to form a holistic view.
- Neglect of soft skills and individual context: Existing systems prioritize academic performance metrics while overlooking softer, yet equally critical, dimensions like learning styles, interests, and communication strengths.
- Weak adaptability to labor market shifts: Few models are designed to incorporate real-time job market data or predict emerging roles, reducing their relevance for future career planning.

These findings highlight an urgent need for a more comprehensive, AI-driven guidance solution—one that not only predicts academic success but also aligns educational choices with evolving career opportunities. Such a system should account for the full spectrum of a student's profile, bridging the gap between academic planning and real-world career readiness

4. DATA DESCRIPTION

Our project utilizes two datasets—Supervised and Unsupervised—both based on real-world trends in education and employment.

The Supervised Dataset contains 5,000 records focusing on the relationship between students' academic backgrounds and their career outcomes. It includes features such as high school and university GPA, standardized test scores, **field of study**, **certifications**, internships, **soft skills**, and networking scores. The target variables include job offers, starting salary, career satisfaction, promotion speed, and job level and more enabling predictive modeling to identify how education and skills influence career success.

The Unsupervised Dataset focuses on clustering graduates based on career outcomes without predefined labels. It includes **70,061** records and encompasses attributes such as field of study, **salary**, promotion rate, job satisfaction, years of experience, work environment, **certifications**, job level, remote work ratio, and performance scores and more. These features enable the identification of patterns in career development, such as grouping graduates with similar success trajectories or uncovering key factors that drive high performance and satisfaction across industries.

Together, these datasets provide a comprehensive foundation for analyzing the link between academic paths and real-world career outcomes, supporting applications in predictive analytics, career recommendation systems, and clustering-based career profiling.

5. Supervised Model SUPERVISED MODEL

5. UNSUPERVISED MODEL

5.1 METHODOLOGY

To ensure accuracy and relevance in predicting students' educational and career outcomes, a supervised learning methodology was implemented. This approach combines data preprocessing, feature selection, and model training using labeled outputs. Classification algorithms such as CatBoost, K-Nearest Neighbors (KNN), and Support Vector Machines (SVM) were applied to predict students' recommended majors based on key academic and behavioral features. CatBoost was ultimately selected due to its strong performance on categorical data and its ability to handle class imbalance effectively. The supervised pipeline was designed to generate reliable, interpretable predictions that align with real-world decision-making scenarios in academic guidance.

5.2 ALGORITHM EXPLANATION

In this research, the CatBoostClassifier was utilized as one of the supervised learning models. CatBoost is a gradient boosting algorithm that is particularly effective with categorical data and imbalanced classes. It was trained using the prepared dataset with default settings to predict students' fields of study. Among the evaluated models, CatBoost demonstrated superior performance in terms of accuracy and generalization, making it a strong candidate for academic and career-related predictions.

5.3 Preprocessing

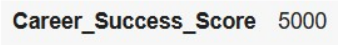
5.3.1 Dataset Description

The dataset used for supervised classification contains multiple academic, behavioral, and performance-related attributes. Each record includes a target variable (Field_of_Study) alongside features such as certifications, soft skills, standardized test scores, and GPA.

5.3.2 Handling Missing Values

Missing values were detected in the **Career_Success_Score** column and were handled by imputing the median value to maintain data balance and minimize the influence of outliers.

As shown in Figure 5.3.2, the number of missing entries in this column is illustrated prior to imputation.



```
Career_Success_Score    5000
```

Figure 1 5.3.2

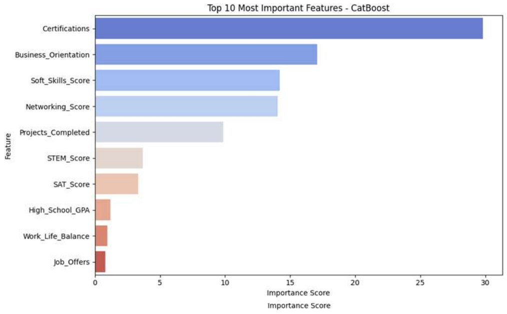
5.3.3 Label Encoding of the Target Variable

To prepare for supervised learning, the target column Field_of_Study was encoded using LabelEncoder. This numeric transformation allowed the target values to be used directly in classification models.

5.3.4 Feature Selection Using CatBoost

5.3.4 Label Encoding of the Target Variable

Feature importance was evaluated using the CatBoost algorithm. After dropping the non-informative identifier column (Student_ID), a CatBoostClassifier was trained with default hyperparameters. The model automatically handled categorical features and accounted for class imbalance via sample weights. Based on the model output, the top 10 features contributing to the prediction of a student's major were extracted.



5.4 Comparative Analysis of Clustering Models

5.3.5 Creating the Final Feature Set

The final selected features used for training and evaluating the supervised classification models were: Certifications, Soft Skills Score, Projects Completed, Networking Score, Work-Life Balance, Business Orientation, SAT Score, STEM Score, and High School GPA.

The feature "Job Offers" was excluded from the final set as it is not suitable to be framed as a question for high school students, given that they typically lack prior experience in the job market.

5.3 Preprocessing

Prior to training and evaluating the models, the dataset underwent several preprocessing steps to enhance data quality, reduce noise, and ensure compatibility with the algorithms. The process included data cleaning, imputation of missing values, feature encoding, scaling, and partitioning.

5.3.1 Data Cleaning and Initial Exploration

A preliminary exploration was conducted to understand the structure, distribution, and quality of the dataset. Key steps included:

- Reviewing data types to distinguish numerical and categorical features.
- Examining the distribution of the target variable (Field_of_Study) and identifying class imbalance.
- Removing duplicates and irrelevant identifiers such as Student_ID.

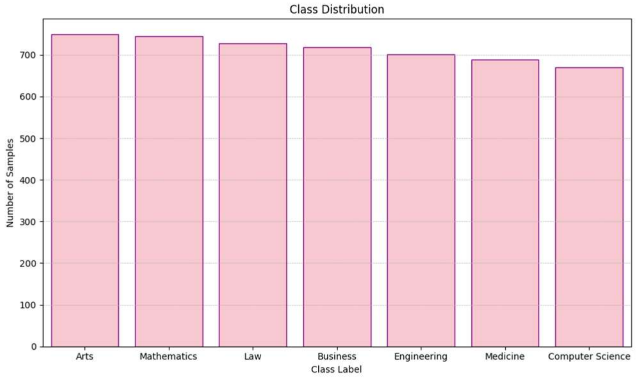


Figure 5.4 Class Distribution Before Preprocessing

5.3.2 Handling Missing Values.2 Handling Missing Values

Missing values were minimal and occurred only in the Career_Success_Score feature. These were imputed using the **median** value to preserve the central tendency without being skewed by outliers. No other features had missing values.

5.3.3 Encoding and Feature Selection

- The target variable Field_of_Study consisted of 7 unique classes (e.g., Arts, Law, Business, Engineering, Medicine, etc.).
- These categorical labels were transformed into numerical form using Label Encoding to facilitate multi-class classification.
- Other categorical or nominal features, if any, were handled through encoding pipelines where necessary.

4.2.4 Feature Scaling5..

5.3.4 Feature Scaling

All numerical features were standardized using StandardScaler, which transformed features to have zero mean and unit variance. This was critical for algorithms such as SVM and KNN, which are sensitive to feature magnitudes.

- Example: SAT_Score before scaling had a mean of 1272.41 and standard deviation of 205.38; after scaling, it was normalized to mean ≈ 0 and standard deviation ≈ 1.

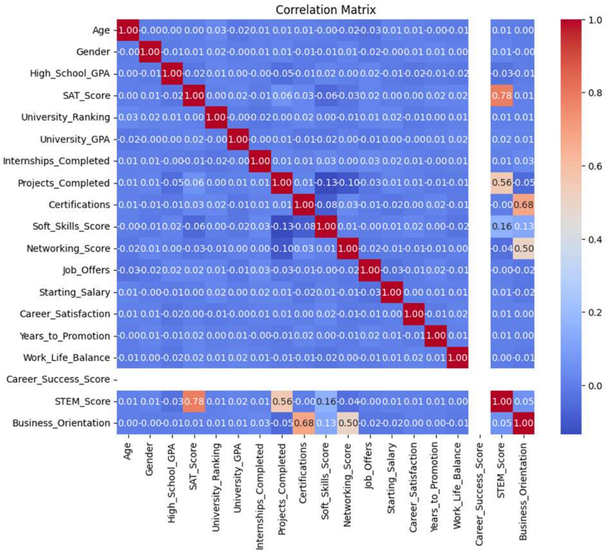


Figure 5.5 Feature Correlation Heatmap

5.3.5 Train-Test Splitting and Validation

- The dataset was split into 80% training and 20% testing subsets using stratified sampling to preserve the distribution of class labels.
- A 5-fold cross-validation strategy was used during model development, particularly for CatBoost, to tune hyperparameters and assess generalization performance.

6. UNSUPERVISED MODEL

6.1 METHODOLOGY

To ensure robustness and transparency in identifying patterns related to education and career success, a comprehensive unsupervised learning methodology was applied. The approach integrates data preprocessing, dimensionality reduction, and clustering. K-Means clustering was employed to identify natural groupings within the dataset based on career-related features, without relying on labeled outputs. K-Means is particularly well-suited for this analysis due to its simplicity, scalability, and ability to handle large datasets efficiently

6.2 ALGORITHM EXPLANATION

K-Means is a widely used unsupervised learning algorithm that partitions a dataset into a specified number of clusters (*k*) by minimizing intra-cluster variance. The algorithm follows an iterative approach:

- **Initialization**: Selects *k* initial centroids randomly or using K-Means++.
- **Assignment**: Assigns each data point to the nearest centroid using Euclidean distance.
- **Update**: Updates centroids by calculating the mean of the assigned points.

These steps are repeated until convergence is achieved—typically when centroids no longer move significantly.

6.3 Preprocessing

6.3.1 Dataset Description

The dataset contains 70,061 records and 15 features, including specialty, salary, and other career-related attributes. Proper preprocessing was essential to preserve data integrity and extract meaningful patterns.

6.3.2 Handling Missing Values

No missing values were found in the dataset. Therefore, the full dataset was used, ensuring reliability and completeness.

6.3.3 Handling Duplicate Records

Duplicate analysis confirmed no repeated entries, maintaining unbiased clustering outcomes.

6.3.4 Encoding Categorical Variables

Categorical features such as industry and specialty were converted into numerical format using Label Encoding, making them suitable for K-Means clustering while retaining categorical distinctions.

6.3.5 Dimensionality Reduction

Principal Component Analysis (PCA) was applied to reduce dimensionality while preserving the most significant variance. A scatterplot Figure 6.1 of the first two principal components showed an elliptical structure, validating the clustering approach.

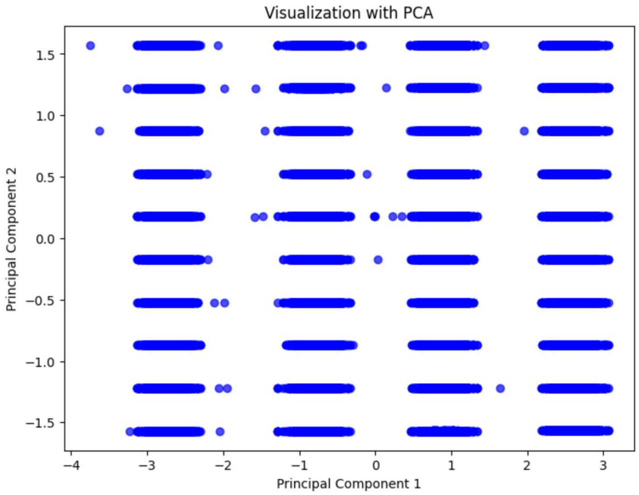


Figure 6.1 Scatterplot of PCA

6.4 Comparative Analysis of Clustering Models

A comparison of K-Means, GMM, Hierarchical Clustering, and DBSCAN was conducted using both internal and external validation metrics:

- **K-Means**: Highest Silhouette and Calinski-Harabasz scores (compact, distinct clusters).
- **DBSCAN**: Sensitivity to noise and density; lower internal scores.
- **GMM**: Best ARI and NMI due to probabilistic modelling.
- **Hierarchical**: Reasonable results on a subset (2,000 samples).

K-Means was chosen for its strong balance between performance and interpretability.
(Figures 4.2.1 to 4.2.4 illustrate model clustering results)

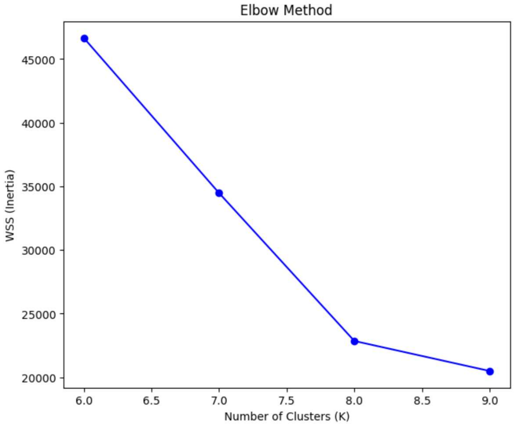


Figure 6.2 Elbow Curve

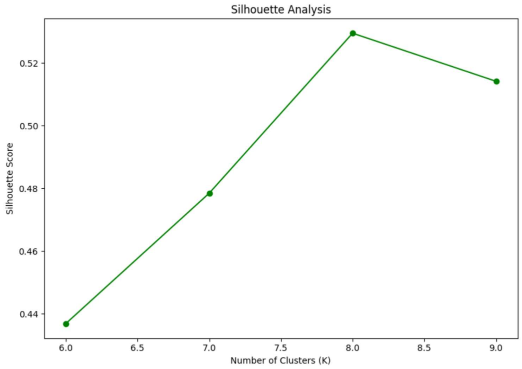


Figure 6.3 Silhouette Score

6.5 Cluster Selection and Interpretation

4.4.1 Determining Optimal Kn4.4.1 Determining Optimal K46.5

4.4.1 Determining Optimal Kn4.4.1 Determining Optimal K46.56

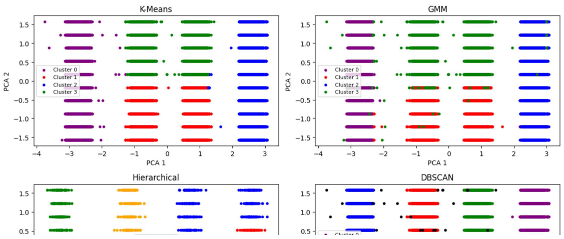
.5

.1 Determining Optimal K

6.5.1 Determining Optimal K

Two methods were used:

- **Elbow Method:** Shows diminishing returns after K = 7.
- **Elbow Method: Shows diminishing returns after K = 7 (Figure 4.3.1).ow Method: Shows diminishing returns after K = 7 (Figure 4.3.1).Elbow Method:** Shows diminishing returns after K = 7.



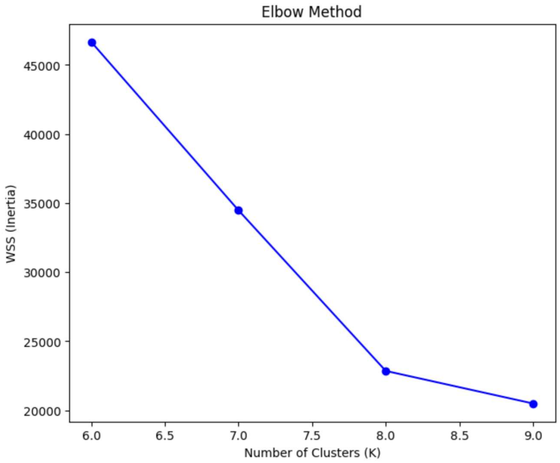


Figure 4.2 elbow curve

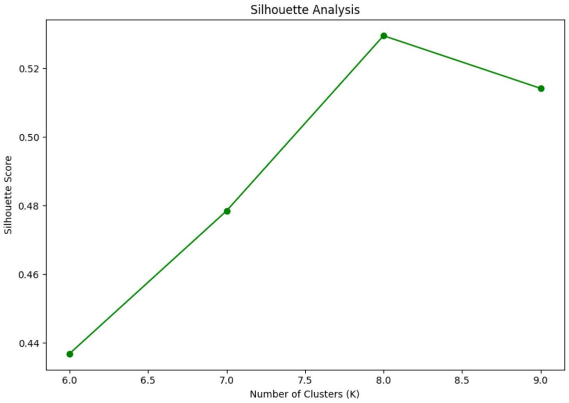


Figure 4.3 Silhouette Score Analysis

4.5 Cluster Selection and Interpretation

4.5.1 Determining Optimal K

Two methods were used:

- **Elbow Method**: Shows diminishing returns after K = 7 (Figure 4.3.1).

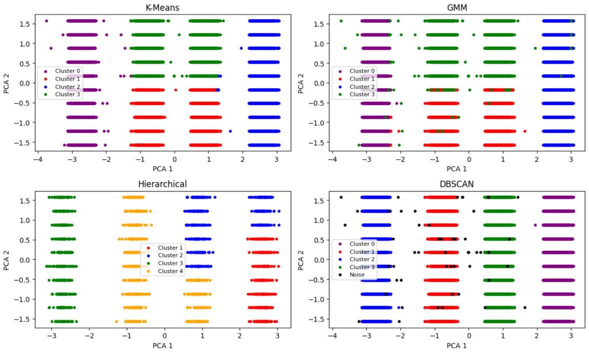


Figure 4.4 Elbow Method

- **Silhouette Score**: Peaks at K = 8 (Figure 4.3.2).

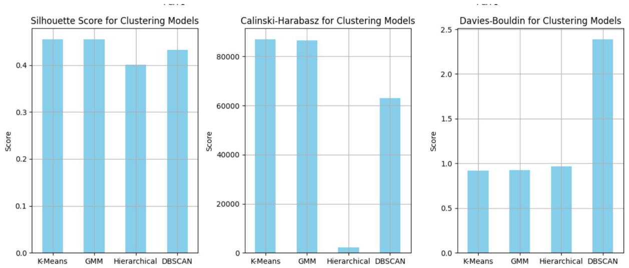


Figure 4.5 Silhouette Score

Thus, **K = 8** was selected for K-Means.

4.5.2 Cluster Analysis Process

The model was trained with K = 8 and analysed using:

- **Cluster Size Evaluation**: Reviewed for balance.
- **t-SNE and PCA Visualization**: Showed clear separation (Figures 4.3.3 & 4.3.4).
- **Centroid Analysis**: Revealed the central tendencies of each cluster.

4.5.3 Feature Interpretation

Boxplots and pairplots (Figures 4.6 & 4.7) revealed inter-cluster variation across key variables:

- Salary
- Job Satisfaction
- Promotion Rate

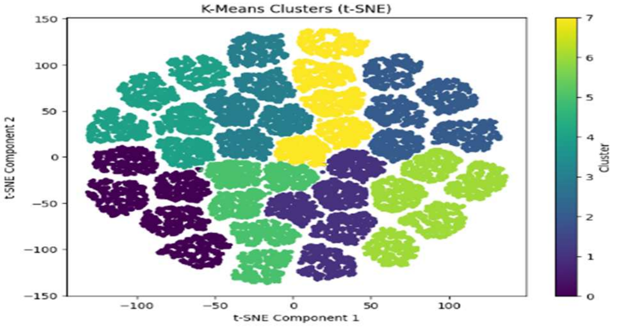


Figure 4.6

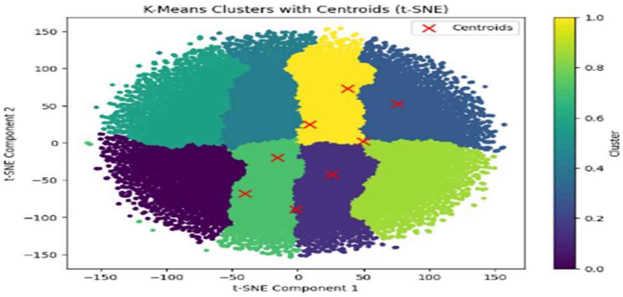


Figure 4.7

4.6 Model Validation

To validate clustering quality:

- **Silhouette Score**: 0.5295 – strong cohesion and separation.
- **Davies-Bouldin Index**: 0.6228 – minimal cluster overlap.

These scores confirm the robustness of the K = 8 configuration.

4.7 Cluster Characterization and Profile Analysis

Each cluster was profiled using:

- **Average values**: Salary, job satisfaction, and promotion rate.
- **Dominant Industry**: Most frequent industry in each cluster.
- **Top 3 Specialties**: Most common academic backgrounds.

Decoded categorical labels aided interpretability and facilitated analysis of graduate career trends.

7. MODEL INTEGRATION

To create a seamless connection between the supervised and unsupervised learning models, we implemented a two-phase system. The first phase involves a **Supervised Learning Model** designed to predict the most suitable university major for a user based on their skill set. Once the prediction is made, the output (predicted major) is passed to the second phase, which utilizes an **Unsupervised Learning Model**.

In the unsupervised phase, clustering techniques were applied to group job roles based on multiple career success factors such as salary, promotion opportunities, and job satisfaction. After generating clusters and analyzing their contents, we extracted insights regarding the distribution of majors within each cluster.
To bridge the two models, we developed a **custom JSON-based mapping structure**. This mapping aligns each predicted major with its most suitable cluster. The matching process was not arbitrary; it was guided by in-depth analysis and research that evaluated which clusters provided the best opportunities for each field of study. This integration enables the system to not only suggest a suitable major but also recommend potential job clusters that align with the user's future career success.

Moreover, a **dedicated Python integration script** was developed to bring all components together. This script loads the trained supervised model, uses it to predict the user's major, then accesses the appropriate unsupervised clustering results using the JSON mapping, and finally returns a complete and interpretable output to the user. It acts as the central pipeline coordinating between the models, data files, and deployment logic to ensure smooth end-to-end functionality.

This architecture ensures that users receive **both academic and career guidance**, enhancing the decision-making process with data-driven insights from both supervised predictions and unsupervised clustering.

8.DISCUSSION