Title: Report on Resume Filtering using Unsupervised Learning and Classification Models

1. Introduction

- The purpose of this report is to present the findings and methodology of a project aimed at resume filtering using unsupervised learning and classification models.
- The project involves analyzing a dataset containing candidate details such as grades, skills (Python, machine learning, deep learning), and grades of undergraduate studies (10th and 12th). The target label is unavailable, necessitating the use of unsupervised learning techniques.

2. Data Preprocessing

- A significant challenge in the dataset was the presence of more than 70% missing values in the grades column. To address this issue, a missing value indicator column was created.
- For each entry in the original column, if a NaN value was present, a corresponding 0 was assigned to the indicator column. Conversely, if a value was present, weights were assigned based on specific criteria to handle the missing values effectively.

3. Feature Selection

- To identify the most relevant features for the resume filtering task, Principal Component Analysis (PCA) and Iterative Feature Ascent (IFA) were employed. PCA helped identify features with high variance and reduced dimensionality,
- while IFA focused on handling outliers and selecting optimal features. The aim was to identify the best combination of features that would contribute to the overall accuracy of the models.
- The best features are: “NLP”,” Python”,’’ UG_grade_Score’’

4. Unsupervised Techniques or Cluster Formation

- Two separate models were trained to classify candidates into four distinct clusters: moderate fit, good fit, bad fit, and best fit.
- These clusters were created based on the properties of the candidates derived from their features, allowing for a more nuanced evaluation of their suitability for a given position.

- The first model focused on candidates whose data contained complete information, excluding the features with missing values. This model aimed to capture the patterns and characteristics of candidates who had comprehensive data available, thus providing insights into the best-fit candidates without the potential bias introduced by missing values.
- The trained model classified candidates into the moderate fit, good fit, bad fit, and best-fit clusters based on their features.

- The second model aimed to incorporate the remaining candidates whose data contained missing values. As previously mentioned, a missing value indicator column was created to handle these missing values.
- By considering this additional information, the model sought to provide a comprehensive evaluation of candidates, taking into account the missing data. The trained model classified these candidates into the same four clusters: moderate fit, good fit, bad fit, and best fit.

- It is important to note that the assumption underlying this approach is that the missing value indicator column adequately represents the missing values in the original features.

- While this assumption allowed for the inclusion of candidates with missing data, it should be acknowledged that a model trained on all

features, including the indicator column, may still be subject to some degree of bias.

- By employing two separate models and considering the missing value indicator column, the cluster formation process aimed to reduce bias and provide a more comprehensive understanding of candidate fit based on their properties.
- This approach facilitated a more nuanced analysis and enabled better decision-making in the resume-filtering task.

5. Labeling Data(Probabilistic Approach)
   - With the clusters formed, a labeling process was conducted to assign labels to the data. Each candidate's data was labeled according to the cluster they belonged to.
   - This step transformed the dataset into labeled data, enabling the application of classification models for further analysis.

6. Classification Model Training
   - Using the labeled data, a classification model was trained to predict the fit of a candidate based on their resume details.

   - Despite the challenges posed by the dataset's quality and the high percentage of missing values, the classification model achieved an accuracy of 90%. This accomplishment highlights the effectiveness of the approach in addressing the data limitations.

7. Conclusion

- In conclusion, this project utilized unsupervised learning techniques, such as feature selection, cluster formation, and labeling, in combination with a classification model to perform resume filtering.
- The application of the missing value indicator column, PCA, and IFA enabled the effective handling of missing values and the identification of relevant features.
- The introduction of clusters mitigated potential bias, while the classification model achieved a commendable accuracy of 90% despite the challenges presented by the dataset.