

Problem Description

The goal is to hire the best intern from the given dataset. The dataset consists of 14 columns, and we need to evaluate the interns based on their scores in various categories and select the intern with the highest aggregate score.

Approach

The approach is to assign aggregates to different categories or features and calculate the aggregate score for each intern. The total aggregate score for all categories should sum up to 10.

The relevant columns considered for evaluation are:

- **Python** (Aggregate: 1.5)
- **Machine Learning** (Aggregate: 1.8)
- **Deep Learning** (Aggregate: 1.8)
- **Natural Language Processing (NLP)** (Aggregate: 1.5)
- **Other Skills** (Aggregate: 0.5)
- **Availability** (Aggregate: 0.5)
- **Degree** (Aggregate: 0.5)
- **Stream** (Aggregate: 1.4)
- **Year of Graduation** (Aggregate: 0.5)
- **Study** (Aggregate: 1)

The challenge in the dataset is handling missing values in the **Degree** and **Stream** columns. To address this:

- For the **Degree** column, replace the missing values with "Unknown" to retain valuable data for evaluation.
- For the **Stream** column, replace the missing values with "Unknown" to retain valuable data for evaluation.

Assigning weights, Values to each column in the dataset(specific values will be assigned based on the relevance to data science and data analytics internships):

- **Python, Machine Learning, Deep Learning, and Natural Language Processing (NLP)** columns are left as they are, on a scale of 0 to 3.
- For each extra skill mentioned in the **Other Skills** column, a value of 0.03 is assigned.
- For **Study Aggregate**:
 - For 10th grade, a value of 0.05 is assigned.
 - For 12th grade, a value of 0.1 is assigned.
 - For UG, a value of 0.3 is assigned.
 - For PG, a value of 0.05 is assigned.
 - The total value being 0.5, the associated weight should be 2 so that the aggregate is 1.
- For **Stream Aggregate**:
 - As studying in a relevant stream is important a aggregate of 1.4 is chosen.
 - stream is given priority based on AI /ML /DS > CSE > IT,IOT > ECE, Mechanical > Metallurgy, Pharmacy
 - based on keywords value and weight are chosen so that the total aggregate is 1.4
- For **** Degree Aggregate****:
 - Degree is given priority based on Masters > Bachelor's
 - based on keywords value and weight are chosen so that the aggregate is 0.5
- For **Python Aggregate**:
 - values were chosen directly from given data set (in scale of 0 to 3)
 - weight associated is 0.5
 - Total Aggregate is 1.5
- For **Machine Learning Aggregate**:
 - values were chosen directly from given data set (in scale of 0 to 3)
 - weight associated is 0.5
 - Total Aggregate is 1.8
- For **NLP Aggregate**:
 - values were chosen directly from given data set (in scale of 0 to 3)
 - weight associated is 0.5
 - Total Aggregate is 1.5
- For **Deep Learning Aggregate**:
 - values were chosen directly from given data set (in scale of 0 to 3)
 - weight associated is 0.5
 - Total Aggregate is 1.8
- For **Availability Aggregate**:
 - if the candidate is available for 3 months internship , immediately as value of 1 is assigned
 - weight associated is 0.5
 - Total Aggregate is 0.5
- For **Year of Graduation Aggregate**:
 - 2024 graduates are more preferred,so value of 3 - (abs(2024) - Current Year of Graduation) which returns a value in scale of (0-3)

- Thus weight of 0.166 is assigned
 - Total Aggregates is 0.5
- For **Other Skills Aggregate**:
 - Each Extra skill is assigned a value of

Scoring and Ranking Metrics

- Assign aggregates to each feature/category based on its importance and relevance to the hiring decision.
- Calculate the score for each intern by multiplying the feature values with their respective aggregates and summing them up.
- Rank the interns based on their aggregate scores.
- Select the intern with the highest aggregate score as the best candidate for hiring.

By following this approach, we can effectively evaluate the interns and identify the best candidate for the internship opportunity.

Please note that these aggregates and weights are adjustable and subjective based on the specific requirements and preferences of the evaluation criteria. Feel free to further customize them as per your specific needs. Please refer to [digi.ipynb](#) for implementation of above approach