Data Collection Report 6-sept -2025

Assignment 2 Name: Abdi fatah Mahad Isse (Tamer)

Title:

Unemployment Status Dataset

1. Collection Method

This dataset was collected from my friends to study the employment status of 50

individuals. The data was gathered through surveys and in terviews, including details such

as age, gender, education level, skills, work experience, and participation in training

programs. The employment status of each person was determined based on these factors,

with some natural variation. To reflect real -worl d data, there may be issues such as missing

values, typos, or outliers.

2. Features & Label

Features (X):

 person\_id – Unique identifier of individual.

 Age – Age of the individual.

 Gender – Male/Female.

 education\_level – Education level (Primary, Secondary, University).

 Skills – Skill level (Low, Medium, High).

 job\_experien ce\_years – Years of work experience.

 training\_program – Participation in training program (Yes/No).

Label (y):

 employment\_status – Employment condition (Employed, Unemployed).

3. Dataset Structure

Rows: 50

Columns: 8 (7 features + 1 label)

Format: CSV

Samp le (5 rows):

4. Data Quality Issues

 Missing values : Some individuals may have missing skills or training data.

 Typos : Possible errors in employment status labels.

 Duplicates : Some records may be repeated.

 Outliers : Example, a 20 -year -old with 15 years of experience.

5. Use Case

 Classification : Predict whether a person is employed or unemployed.

 Policy Making : Support governments in creating employment programs.

 Clustering : Group individuals based on education, skills, and employment status. person\_id age gender education\_level skills job\_experience\_years training\_program employment\_status

1 25 Male University Medium 4 Yes Employed

2 24 Male University Medium 6 Yes Employed

3 31 Male University High 13 Yes Employed

4 44 Male Secondary High 22 Yes Employed

5 45 Female Secondary Low 24 No Employed

6. Conclusion

This dataset provides a foundation for practicing data preprocessing, cleaning, and applying

machine learning models. It mirrors real -world imperfections and supports classification

and clustering tasks for analyzing unemployment factors.