Project Statement:

Task 1: Data Cleaning with Azure Databricks

Dataset: https://raw.githubusercontent.com/justmarkham/DAT8/master/data/u.user

Tasks: Clean the dataset and perform the following using Databricks:

Import the dataset.

Assign it to a variable called users and use the 'user\_id' as index

See the first 25 entries

See the last 10 entries

What is the number of observations in the dataset?

What is the number of columns in the dataset?

Print the name of all the columns.

How is the dataset indexed?

What is the data type of each column?

Print only the occupation column

How many different occupations are in this dataset?

What is the most frequent occupation?

Summarize the DataFrame.

Summarize all the columns

Summarize only the occupation column

What is the mean age of users?

What is the age with least occurrence?

Task 2: Machine Learning with Azure Databricks

Create a Machine Learning model to find the diagnosis of Breast Cancer Patients based on the features in the dataset below.

Dataset: https://raw.githubusercontent.com/deepanshuMeteor/ComputerVision/refs/heads/main/data.csv