README

Female Labor Force Participation and Gendered Social Norms: A Machine Learning Analysis of Egypt

Author: Natalia Pecorari and Jose Cuesta

1. Overview

This reproducibility package contains the data, code, and outputs required to replicate the empirical analysis of female labor force participation and gendered social norms in Egypt using both descriptive statistics and machine learning (decision trees). A third party can run the package by updating file paths in the scripts to their local environment and ensuring all required Python packages are installed. No additional modifications are required.

The package is organized into folders corresponding to the main components of the analysis:

- Descriptive statistics
- Decision tree models
- Data cleaning

2. Folder Structure

The repository is structured as follows:

Data cleaning/

- **Code/**
    - Egypt AB.do: Do file to clean and prepare the original survey data and create variables for the analysis.
    - Graphs.do: Do file that reproduces the data in the excel file Graphs.xlsx (these are simply tabulations from the original dataset).
    - main.do: main script that runs the two do-files above by changing the file path in line 7.
- **Data/**
    - AB7_ENG_Release_Version6.dta: Original Arab Barometer round 7 dataset.
- **Output/**
    - AB7_EGY_clean_with_missing.dta: Clean dataset including missing values used for descriptive analysis.
    - AB7_EGY_clean.dta: Clean dataset used for machine learning analysis.

Descriptives table/

- **Code/**
    - Descriptives_table.py: Script to compute weighted descriptive statistics by gender.
- **Data/**
    - AB7_EGY_clean_with_missing.dta: Dataset including missing values used for descriptive analysis.
- **Output/**
    - Excel file with weighted descriptive statistics (generated by the script).

Decision Trees/

- **Code/**
    - decision tree.py: Script to estimate and visualize decision tree models.
- **Data/**
    - AB7_EGY_clean.dta: Clean dataset used for machine learning analysis.

Additional file

- Graphs.xlsx: Contains figures generated for the analysis.

3. Data Availability Statement

The data used in this project come from a cleaned version of the AB7 Egypt survey dataset. The original dataset can be downloaded here: https://www.arabbarometer.org/surveys/arab-barometer-wave-vii/

- The datasets included in this package are:
    - AB7_EGY_clean_with_missing.dta
    - AB7_EGY_clean.dta
- The data are provided solely for replication purposes within this package.
- If access to the original raw data is required, users should refer to the original data provider.

4. Instructions for Replicators

To replicate the results, proceed as follows:

Step 1: Set up environment

Ensure all required Python packages are installed (see Section 6).

Step 2: Update file paths

Both scripts contain local file paths.
Before running, update the paths to match your local directory structure.

Step 3: Run descriptive statistics

Navigate to:

Descriptives table/Code/

Run:

Descriptives_table.py

This will generate a weighted descriptive statistics table (overall, men, women) and export it as an Excel file.

Step 4: Run decision tree analysis

Navigate to:

Decision Trees/Code/

Run:

decision tree.py

This will:

- Perform hyperparameter tuning via grid search
- Train decision tree models
- Generate visualizations of trees
- Output feature importance plots

Step 5: Review outputs

- Descriptive tables are saved in the **Outputs/** folder
- Graphical outputs and results are displayed and/or saved during execution
- Additional figures are available in Graphs.xlsx

5. List of Exhibits

The package reproduces the following outputs:

1. **Weighted descriptive statistics table** (overall, men, women)
2. **Decision tree visualizations** (branch structure)
3. **Feature importance plots** (not used in paper)
4. Supporting graphs compiled in Graphs.xlsx

6. Computational Requirements

Software

- Stata
- Python 3.x

Required Python packages

- pandas
- numpy
- statsmodels
- scikit-learn
- matplotlib
- seaborn
- graphviz

Notes

- Graphviz must be installed separately for tree visualization.
- The code was developed and tested in a standard Python environment (e.g., Anaconda or similar).

7. Description of Code

Descriptives_table.py

Computes weighted means, standard deviations, and sample sizes for selected variables using survey weights, disaggregated by gender (all, men, women), and exports the results to Excel.

decision tree.py

Implements decision tree classification models to predict employment status, including hyperparameter tuning, model training, visualization of decision trees, and feature importance analysis.

Egypt AB.do

Keeps Egypt and prepares the variables for analysis. Also deals with missing observations and generates two clean datasets, one with missing values and the other without.

Graphs.do

Includes the tabulations for Figure 1, 2, and 3 in the paper. These are later used in Graphs.xlsx to prepare the final figures.

main.do

main do-file that runs Egypt AB.do and Graphs.do.

8. Notes

- All results are fully reproducible conditional on correct path configuration.
- The analysis uses survey weights (WT) throughout.
- Minor differences may arise depending on the computational environment (e.g., package versions).

9. Exhibits

This table links each exhibit in the paper to the corresponding output file and code used to generate it.

Exhibit in Paper	Description	Output File / Source	Code Used
Appendix 1	Descriptive statistics (overall, men, women)	Excel file generated in Descriptives table/Outputs/	Descriptives_table.py
Figure 1	Perceived barriers to workforce entry (overall and by gender)	Graphs.xlsx	Excel-based processing
Figure 2	Reported barriers to female workforce entry (both men and women)	Graphs.xlsx	Excel-based processing
Figure 3	Reported barriers to female workforce entry (only women)	Graphs.xlsx	Excel-based processing
Figure 4	Decision tree predicting female labor force participation	Generated during script execution (Graphviz visualization)	decision tree.py
Figure 5	Decision tree predicting male labor force participation	Generated during script execution (Graphviz visualization)	decision tree.py

Figures 4 and 5 are generated dynamically during execution of the decision tree script and displayed using Graphviz. These figures are not saved automatically but can be exported manually by the user if needed.

10. Data Availability Statement

This project uses survey data from the Arab Barometer, specifically Wave VII.

Dataset

- **File name:** ._AB7_ENG_Release_Version6
- **Brief description:** Original Arab Barometer Wave VII dataset (English version), which contains nationally representative survey data on socio-economic conditions, demographics, and attitudes across countries in the Middle East and North Africa. The analysis in this project uses the Egypt sample extracted from this dataset.
- **Citation:** Arab Barometer (2022). *Arab Barometer Wave VII Dataset*.
- **URL / access:** https://www.arabbarometer.org/surveys/arab-barometer-wave-vii/
- **Access / download date:** March 2025
- **Download steps:**
    1. Go to the Arab Barometer website
    2. Navigate to "Survey Data" → "Data Downloads"
    3. Register or log in
    4. Download the Wave VII dataset (English version)
- **License / access restrictions:**
  The data are publicly available upon free registration. Users must agree to the Arab Barometer terms of use. Redistribution of raw data may be subject to restrictions imposed by the data provider.