UNITED NATIONS

ECONOMIC AND SOCIAL COMMISSION FOR ASIA AND THE PACIFIC (ESCAP)

STATISTICAL INSTITUTE FOR ASIA AND THE PACIFIC (SIAP)
In partnership with
THE ASIAN DEVELOPMENT BANK (ADB)[1]

Principles of Machine Learning for Official Statistics and Sustainable Development Goals Indicators

Self-paced course

I. About the Course

This self-paced course introduces the principles of Machine Learning (ML) for using either traditional or non-traditional data sources (big data) to produce high quality predictions for official statistics or Sustainable Development Goals (SDGs) indicators.

The course provides an opportunity for participants to explore and comprehend the techniques of machine learning in comparison with more traditional statistical methods. It aims at providing an overview of the current methods and applications of Machine Learning, through theoretical concepts, pedagogical case studies and interactive resources. The course is not based on, nor does it require, a particular software. However, reproducible examples on either simulated or real data are provided using the R/RStudio environment. Some Python procedures and packages are also provided.

This self-paced e-course has been developed as an interactive training composed of 6 modules following a detailed introduction. Each module is composed of several pedagogical activities, some being mandatory (marked with a *****), following a logical structure. Activities include interactive lessons, practitioners interviews and case studies, interactive web-based apps, optional tutorials, articles and quizzes. When a module is completed, a new piece of text appears with congratulations and the main elements to remember from the module.

[1] *The contributions of ADB staff and consultants for this initiative were supported by the Japan Fund for Prosperous and Resilient Asia and the Pacific financed by the Government of Japan through the ADB (TA 6721-REG: Using Frontier Technology and Big Data Analytics for Smart Infrastructure Facility Planning and Monitoring and TA 6856-REG: Development of New Statistical Resources and Building Capacity in New Data Sources and Technologies*

II. Target Audience

The course is designed for personnel working in the field of statistics, whose main responsibilities include data analysis of SDG indicators and related statistics with a specific target on data scientists from NSOs with an experience in both statistical modelling (regression analysis, prediction, classification, ...) and with programming or algorithmic skills. The participants must have a good practice in the manipulation of data as well as a good understanding of statistical methods. Although no programming will be required to follow the course, some optional pedagogical materials include R code, in the form of reproducible markdown notebooks, as well as some Python resources and code.

III. Learning Objectives

At the end of the course, the learners should be able to:

- Evaluate the potential use of ML for official statistics and SDGs
- Compare classical statistical and ML methods
- Describe the features of ML techniques
- Apply cross-validation method to estimate the quality of ML methods
- Describe the main steps of a ML project
- Recognize the ethical challenges and potential issues in the use of ML techniques
- Differentiate Supervised *vs* Unsupervised ML problems
- Manipulate classification algorithms
- Perform regressors selection using ML techniques
- Produce predictions using ML algorithms
- Evaluate the quality of ML predictions
- Apply data visualization techniques to assess the quality of prediction
- Select the right criteria/ visualization for determining some ML hyperparameters
- Identify the limitations of ML for official Statistics and SDGs monitoring
- Interpret ML outcomes and predictions

IV. Course Design and Content

The course is divided in 6 modules, plus an introduction. Each module requires approximately three hours of focused work. Motivated participants can expect to spend much more time to replicate the analysis proposed optionally in each module using the notebooks, code and data provided. All elements (slides, code and references) used for each activity are available for download in the LMS in the form of pdf or html documents.

Module	Coverage
Introduction: Machine Learning for Official Statistics	• ML for Official Statistics and SDGs • Case Study: Using Satellite Images and ML to estimate poverty (ADB)
M1: Statistical vs machine learning	• Statistical learning • Machine Learning • K-Nearest Neighbours
M2: Classification	• How classification works • Supervised vs unsupervised classification • Measures of fit • ML with Logit as a classifier • Case Study: ML for Automatic Classification (Danish Statistical Authority)
M3: Regression	• ML with Linear Regression • Selection of regressors • Penalization Methods
M4: Decision Trees	• Decisions Trees • Tuning and pruning trees • Visualizing a decision tree
M5: Random Forest	• Bagging and Boosting • Random Forest • Random Forest and Imputation • Case Study: Machine learning for mobile phone data
M6: Advanced Methods	• Ethical considerations in ML • Origins of Support Vector Machine • Support Vector Machines • Unsupervised Learning k-Means

V. Certificates

A certificate of completion is available for participants:

- who have completed **all** the mandatory activities (marked with a ***** in each module) and
- who have a grade equal or greater than **70%** in the final exam, and
- who have completed the **feedback** evaluation form.