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)

**e-Learning Course**

# Machine Learning for Official Statistics and SDG Indicators

**7 December 2026 – 22 January 2027**

## I. Background

The growing availability of data, together with rapid advances in Artificial Intelligence (AI), is transforming the landscape of Official Statistics. For National Statistical Offices (NSOs), these developments bring new opportunities to improve statistical production, make better use of emerging data sources, and respond to the increasing demand for timely, relevant and high-quality statistics.

At the heart of many of these developments is Machine Learning (ML), which provides a set of powerful methods for analyzing large and complex datasets, identifying patterns, making predictions and automating tasks such as classification. These methods can support different stages of statistical production and complement traditional statistical approaches.

For NSOs, unlocking the value of Machine Learning and AI requires not only access to new technologies, but also a workforce complementing their strong foundations in statistical theory and methodology with new competencies. For NSOs, investing in the development of these skills is therefore becoming a strategic priority to remain relevant and effective and to ensure that Official Statistics continue to meet the highest standards of quality, transparency, and public trust.

## II. About the Course

This course introduces Machine Learning as a set of methods for making effective use of both traditional data sources, such as surveys and microdata, and non-traditional sources, including Big Data, to support the production of high-quality estimates and predictions for Official Statistics and Sustainable Development Goal (SDG) indicators.

Participants will first explore the main concepts and techniques of Machine Learning and their connections with traditional statistical methods. The course then focuses on a broad selection of ML's current methods and their applications in Official Statistics, combining accessible methodological explanations with practical case studies, interactive learning resources and exercises.

The course is not based, nor does it require, a particular software. However, reproducible examples on either simulated or real data sets are provided using the R/RStudio environment. Python procedures and packages are also provided.

The course has been developed as an interactive training composed of 6 + 1 modules. Each module is composed of several mandatory activities, following a pedagogical structure. Activities include videos, interactive videos, interactive web-based apps, chats, document reading, exercises, polls and quizzes as well as live lectures and webinars. A preliminary module, M0, serves as a reminder and proposes a summary of statistical notions, terminology and basic concepts used by data analysts. These notions will be used throughout the course.

The course is hosted on the SIAP's e-learning platform (LMS) which contains a forum for general questions and interactions with the SIAP's lecturers and e-learning platform administrators. Mandatory weekly webinars, in the form of presentations or Q&A sessions will be proposed using the Microsoft Teams platform. The participants are expected to interact during live sessions and through forums embedded in each module.

**Instructions, conversations, pedagogical resources and webinars are in English.**

## III. Target Audience

This course is intended for professionals working in the field of statistics, particularly those involved in the analysis of SDG indicators and related data. It is specifically targeted at data scientists from National Statistical Offices (NSOs) who have a solid experience in both statistical modelling (e.g., regression analysis, prediction, classification) and in programming or algorithmic thinking.

Prior exposure to programming in at least one scientific or statistical language (e.g., R, Python, Stata,..), is recommended, even though active coding will not be required during the course.

The course materials include reproducible R code in Markdown notebooks, along with supplementary Python resources. To ensure a smooth learning experience, participants are encouraged to complete a preparatory tutorial[1] before the course begins.

## IV. 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 and main steps of a ML project
- Differentiate Supervised *vs* Unsupervised ML problems
- Manipulate classification algorithms
- Perform Regressors selection using ML techniques

[1] This tutorial will be quite easy to complete for any participant having some experience in analysing statistical data using any modern programming language (Stata, Python or R).

- 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 hyperparameters
- Identify the limitations of ML for official Statistics and SDGs monitoring
- Interpret ML outcomes and predictions and recognize their biases
- Identify Ethical issues when using Machine Learning

## V. Course Design, Workload and Content

The course is divided in 6 (+1) modules and will last for 7 weeks. The pedagogical materials (slides, code and references) used for each activity will be made available for download in the e-learning platform in the form of pdf or html documents. Optional documents include R-Markdown documents with reproducible code. Python users will find some code and functions similar to ones used in R in the Markdown documents.

**The workload** of this course is of **a minimum of 3h/week.** It is thus important for participants to organize themselves and allocate a sufficient amount of learning time during the course.

| Module | Coverage |
|---|---|
| M0: Statistical Learning: "You've seen this before" | • Linear and non-linear regression<br>• Supervised *vs* unsupervised learning<br>• k-Nearest Neighbours<br>• Statistical Learning *vs* Machine Learning<br>• Cross validation |
| M1: Machine Learning for Official Statistics and the SDGs | • Introduction to the course<br>• Using Machine Leaning for Official Statistics<br>• Tools for applied Machine Learning<br>• Limitations of Machine Learning predictions |
| M2: Machine Learning in Practice: Classification | • How classification works<br>• Supervised *vs* unsupervised classification<br>• Examples of classifiers<br>• Measures of fit<br>• Logit as a classifier<br>• How to choose the best model? |
| M3: Machine Learning in practice: Regression | • Linear Regression and all his friends<br>• Selection of regressors<br>• Penalization Methods<br>• How to choose the best model? |

| | |
|---|---|
| M4: Machine Learning in Practice: Decision Trees | • Decisions Trees<br>• Visualizing a decision tree<br>• Bagging and Boosting<br>• Random Forest<br>• Random Forest and Imputation |
| M5: Machine Learning in Practice: Advanced Methods | • Support Vector Machines<br>• Unsupervised Learning (k-Means)<br>• Introduction to Neural Networks |
| M6: "Real-World Machine Learning":<br>Part 1: Estimating Poverty From Satellite Images (Asian Development Bank)<br><br>Part 2: Ethical Considerations in Machine Learning | • Using satellite imagery to estimate poverty<br>- Training a Neural Network<br>- From images to data<br>- Predicting poverty and visualization<br>• Ethical considerations in the use of Machine Learning<br>- Which Ethical and Legal Constrains for Machine Learning projects? |

## VI. Evaluation

Assessment will focus on participants' ability to apply the methods and concepts introduced throughout the course, as well as their understanding of the key learning outcomes of each module. The final grade will be based on the results of six mandatory assessments, one to be completed at the end of each module.

An individual data-based project will be proposed as an optional activity, and will provide some **bonus points** if completed before the end of the course.

**Participation to the various chats and Q&A will also provide bonus points.**

**To succeed** in the course, participants must comply with the following conditions:

1) Have a **final** grade **greater than 70**/100**,**
2) Participate in **all** weekly webinars.
3) Complete all **mandatory** activities in due time

**To receive a certificate,** participants **must complete** the end-of-course evaluation.