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 offers powerful methods for analyzing large, 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, realizing the potential of ML and AI requires more than access to new technologies. It also requires strengthening the competencies of statistical professionals, building on their strong foundations in statistical theory and methodology. Investing in these skills is therefore increasingly important for modernizing statistical production and use while maintaining the quality, transparency, and public trust essential to official statistics.

Investing in these skills is therefore increasingly important for modernizing statistical production, enabling statistical professionals to identify and apply appropriate machine learning methods to address practical statistical challenges, while maintaining the quality, transparency, and public trust essential to official statistics.

II. About the Course

This course introduces machine learning as a tool for using either traditional data sources, such as surveys and microdata, or non-traditional data sources, including big data, to produce high-quality predictions that support the production of 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 introduces a range of ML methods and their applications in official statistics, combining methodological explanations with practical case studies, interactive learning resources, and exercises.

The emphasis is on helping participants assess when and how machine learning methods can be applied to practical statistical problems, including their potential benefits, limitations, and implications for the quality of official statistics.

The course is not based, nor does it require, a particular software. However, reproducible examples on using simulated and real data sets will be provided through R/RStudio environment. Python procedures and packages are also provided. Participants will be able to explore the methods and their applications without being required to undertake active coding.

The course consists of six core modules and one preparatory (M0) module. Each module follows a structured learning sequence comprising mandatory activities, including video lectures, interactive learning applications, interactive web-based apps, chats, readings, exercises, polls, and quizzes. The preparatory module provides a review of key statistical concepts, terminology, and methods used throughout the course.

The course will be hosted on SIAP's e-learning platform (Learning Management System (LMS)), which includes discussion forums for interaction among participants, SIAP lecturers, and platform administrators. Mandatory weekly webinars will be conducted through Microsoft Teams, providing opportunities for presentations, questions and answers, and discussions of the course materials.

The course will be conducted entirely in English, including all learning materials, communication, exercises, and webinars.

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 solid experience in both statistical modelling (e.g., regression analysis, prediction, classification) and programming or algorithmic thinking.

Participants should have solid experience in data manipulation, a sound understanding of statistical methods, and familiarity with at least one scientific or statistical language (e.g., R, Python, Stata, SAS, or MATLAB). However, active coding will not be required during the course.

Course materials include reproducible R code in Markdown notebooks, supplemented by Python resources. To help participants prepare, they are encouraged to complete a preparatory tutorial[1] before

[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).

the course begins. The tutorial is designed to be accessible to participants with prior experience in statistical data analysis using languages such as R, Python, or Stata.

IV. Learning Objectives

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

- Assess the potential applications of ML in the production of official statistics and SDG indicators
- Compare classical statistical methods and ML approaches
- Describe the key features and main steps of an ML project
- Distinguish between supervised and unsupervised ML problems
- Describe and apply common classification algorithms
- Select explanatory variables using ML techniques
- Produce predictions using ML algorithms
- Evaluate the quality and performance of ML predictions
- Apply data visualization techniques to assess prediction quality
- Select appropriate criteria and visualization techniques for model selection and determining hyperparameters
- Identify the limitations of ML in the production of official statistics and SDGs monitoring
- Interpret ML results and predictions, including recognizing their biases
- Identify ethical considerations associated with the use of ML

V. Course Design, Workload and Content

The course will be delivered over seven weeks and consists of six core modules and one preparatory module (M0). Learning materials, including presentations, code examples, and references, will be made available through SIAP's e-learning platform in PDF or HTML format. Reproducible R-Markdown notebooks and supplementary Python resources will also be provided.

Participants are expected to devote at least 3 hours per week to self-paced learning, in addition to participating in a mandatory weekly webinar. The participants should ensure that they can allocate sufficient time to meet these requirements throughout the course.

The modules cover the following topics:

Module	Coverage
M0: Statistical Learning: "You've seen this before" (Preparatory module)	• Linear and non-linear regression • Supervised and unsupervised learning • k-Nearest Neighbours • From statistical learning to machine learning • Cross-validation
M1: Machine Learning for Official Statistics and the SDGs	• Introduction to the course

Module	Coverage
	• Applications of machine learning in official statistics • Tools for applied machine learning • Limitations of machine learning predictions
M2: Machine Learning in Practice: Classification	• How classification works • Supervised and unsupervised classification • Examples of classifiers • Measures of model fit and performance • Logistic regression as a classifier • Model selection
M3: Machine Learning in practice – Regression	• Linear Regression and all his friends • Selection of regressors • Penalization Methods • How to choose the best model?
M4: Machine Learning in Practice – Decision Trees	• Decision trees • Visualizing a decision tree • Bagging and boosting • Random forests • Random forests and imputation
M5: Machine Learning in Practice – Advanced Methods	• Support vector machines • Unsupervised learning (k-means) • Introduction to neural networks
M6: Real-World Machine Learning Part 1: Estimating Poverty From Satellite Images (Asian Development Bank) Part 2: Ethical Considerations in Machine Learning	• Training a neural network • Transforming images into data • Poverty prediction and visualization • Ethical and legal constraints in 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 six mandatory assessments, one at the end of each core module. An optional individual data-based project will also be offered, with **bonus points** awarded for completion before the end of the course. Participation in discussion forums and question-and-answer sessions will also contribute to bonus points.

To successfully complete the course, participants must:

1) Achieve a **final** grade **more than 70** out of 100.
2) Participate in **all** mandatory weekly webinars.
3) Complete all **mandatory** activities within the prescribed deadlines.

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

Participants are also encouraged to share with SIAP how they have **applied the knowledge and skills gained** from the course in their respective institutions, including any **practical applications of machine learning in statistical production or SDG monitoring**. Such feedback will help SIAP better understand the results of its capacity development activities and inform future training programmes. Participants may share their experiences and feedback by emailing **escap-siaptraining@un.org**.