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]

e-Learning Course

Machine Learning for Official Statistics and SDG Indicators

27 November 2023 – 19 January 2024

I. About the Course

This course introduces machine learning as a tool for using either traditional (surveys, micro data, …) 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 and their links with traditional statistical methods. It aims at providing an overview of the current methods and applications of Machine Learning, through simplified theoretical concepts, pedagogical case studies and interactive resources. The course is not based, 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.

The e-course has been developed as an interactive training composed of 6 + 1 modules. Each module is composed of several mandatory pedagogical activities, following a logical structure. Activities include videos, interactive videos, interactive web-based apps, chats, live lectures and webinars, document reading, exercises, polls and quizzes. 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 Learning Management System (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.**

[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. They must have some experience in programming in any scientific or statistical language (R, Python, Stata, SAS, Matlab, ...). Although no programming will be required to follow and succeed in the course, the 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
- Describe the main steps of a ML project
- 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 hyperparameters
- Identify the limitations of ML for official Statistics and SDGs monitoring
- Interpret ML outcomes and predictions
- Recognize the drawbacks and biases of ML
- Identify Ethical issues when using Machine Learning

IV. Course Design, Workload and Content

The course is divided in 6 (+1) modules and will last for 8 weeks. The elements (slides, code and references) used for each activity will be made available for download in the LMS 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 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 • Supervised *vs* unsupervised learning • k-Nearest Neighbours • Statistical Learning *vs* Machine Learning • Cross validation
M1: Machine Learning for Official Statistics and the SDGs	• Introduction to the course • Using Machine Leaning for Official Statistics • Tools for applied Machine Learning • Limitations of Machine Learning predictions
M2: Machine Learning in Practice: Classification	• How classification works • Supervised *vs* unsupervised classification • Examples of classifiers • Measures of fit • Logit as a classifier • How to choose the best model?
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	• Decisions Trees • Visualizing a decision tree • Bagging and Boosting • Random Forest • Random Forest 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	• Using satellite imagery to estimate poverty - Training a Neural Network - From images to data - Predicting poverty and visualization • Ethical considerations in the use of Machine Learning - Which Ethical and Legal Constrains for Machine Learning projects?

V. Evaluation

The evaluation will be based on the learner's ability to apply the methods and training materials described during the course, as well as on their capacity to recall the main outcome of each module. The final grade will be based on the result to mandatory assessments exercises, in the form of Multiple Choice Questions (MCQ)-like tests to complete 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 have at least a final score of 70/100**, and** comply with the following conditions:

- Follow all of the mandatory activities in due time
- Participate in **all** weekly webinars
- Complete all the weekly assessments (tests) in due time

Participants must complete the end-of-course evaluation to receive a certificate.