e-ITEC Course

offered by

International Statistical Education Centre (ISEC)

Kolkata, INDIA

Name of Course

BIG DATA ANALYTICS FOR POLICY PLANNERS

Stream

Specialized Courses

Country

Countries in Africa and the Middle East falling in the time zones UTC+2 and UTC+3 ONLY, like

> Tanzania, Sudan, the Democratic Republic of the Congo, Kenya, Ethiopia, Zambia, Libya, Egypt, Mozambique, Malawi, Zimbabwe, and so on.

Start date

Monday, AUGUST 30, 2021

End date

Friday, SEPTEMBER 24, 2021

Schedule of classes

FOUR WEEKS, with **2½ hours (during the period 13:00 IST to 16:00 IST)** of lectures and/or interactive sessions with instructor per weekday (Monday to Friday). For each lecture there will be assignments/self-study material of roughly two hours.

Participant Profile

Officers in government, industry and academia involved in policy-planning activities.

Eligibility conditions of Applicants

The lectures will be conducted on working days during a period which is very likely to overlap with normal working hours of the target group. In view of this, it is very important that interested persons should have their applications forwarded by their immediate superiors at work or obtain a NO OBJECTION CERTIFICATE from them. This will enable them to participate freely and fully in the programme by attending all the lectures.

Those who cannot ensure regular attendance in the lectures due to the reason cited above are not encouraged to apply.

An applicant for this course must have

- an undergraduate degree in Science (preferably in the areas of Mathematics/ Statistics/ Economics), with adequate knowledge of the following topics:

  Probability, Standard Probability distributions, Estimation, Testing of Hypotheses, Analysis of Variance (ANOVA), Simple Linear Regression

- knowledge of the R software, at least at the preliminary level

- **proficiency in the English language (in which this course will be conducted entirely)**

SYNOPSIS

Administrative data that policy planners normally need to work with, easily qualifies as Big Data, since it is generally too large or complex to be dealt with by traditional data-processing methodologies. This course intends to provide policy planners with an overview of statistical methodologies that can be useful for decision-making with Big Data. Emphasis will be on explaining the motivation and fundamental ideas behind the techniques, their applicability and illustration of their use with the R software.

Justifications/rationale

In the present-day scenario, data is not limited to what is collected through surveys and censuses. Neither is it limited to quantitative or qualitative data alone. Complex forms of data like text, images, etc. are increasingly being gathered in voluminous amounts by cheap and numerous information-sensing devices such as mobile devices, remote sensing devices, cameras, microphones, radio-frequency identification (RFID) readers, wireless sensor networks, and so on. Digital data storage devices have progressively become cheaper as they have gained in capacity. These modern-day mountains of data are rich in useful information that can and should be exploited for making more informed decisions. Effective mining of this so-called Big Data for gems of information is as rewarding as it is challenging. Policy planners in various realms of human activity, including Government, can formulate more effective policies if they are made aware of the tools and techniques available and are confident enough to apply them.

Objectives of the course

To introduce policy planners working in various areas, like Government and industry, to methodologies that are useful for analysis of Big Data, with the ultimate objective of being able to make more informed decisions leading to better policy-planning.

Expected outcome of the course

At the end of the programme, it is expected that participants will become acquainted with the fundamental statistical techniques that are extremely useful for analyzing Big Data, and also become familiar with the computational aspects, including interpretation of results, through the use of the R software.

Details about the host agency (ISEC)

The **International Statistical Education Centre** (ISEC) is an associate institution of the Indian Statistical Institute and has been providing education and training in statistics for the past 70 years to sponsored students mainly from the countries of the Middle East, South and South-east Asia, the Far East and the Commonwealth Countries of Africa. The Centre also offers various short-term courses in Statistics. The teaching support is provided largely by the Indian Statistical Institute, which is acknowledged all over the world as a premier institute of statistical learning.

Course Content

TOPIC	SYLLABUS	LECTURE HOURS
General Introduction	• Introduction to statistical decision-making. Need for data-driven decisions. • Concepts of supervised and unsupervised learning. Predictive Analytics	2½ hours
Introduction to Big Data Analytics	• Big Data and their features. Capture, Screening and Storage of Big Data. Administrative Records as Big Data. • Analytics for Big Data **Examples with R Package**	2½ hours
Dimension Reduction and Data Visualisation	• Principal Component Analysis • Cluster Analysis • Multidimensional Scaling **Examples with R Package**	2½ hours
Linear Regression Models	**Multiple linear regression**: Least squares estimation of the regression coefficients, Test of significance of regression coefficients, prediction of new observations etc. Inclusion of qualitative regressors.	7½ hours

	Variable selection and model building: Stepwise regression methods (Forward and backward selection). **Multicollinearity**: Effects of multicollinearity, Ridge regression and principal component regression. **Model diagnostics, Model adequacy checking.** **Examples with R Package**	
Generalized Linear Regression Models	**Logistic regression**: Model fitting, interpretation of the coefficients in a logistic regression model, Odds ratio in logistic regression **Poisson regression:** Model fitting and interpretation. **Examples with R Package**	2½ hours
Classification	An overview of classification, Linear and quadratic discriminant function, Classification for normal populations. K-nearest neighbour classifier, Naïve Bayes classifier, Classification using logistic regression. **Examples with R Package**	7½ hours
Cross-validation	The validation Set Approach, Leave One-Out Cross-Validation, k-fold Cross-validation **Examples with R Package**	2½ hours
Tree based methods	Background, Regression tress and classification trees **Examples with R Package**	2½ hours
Support Vector Machine	Overview of support vector classifier, Support vector machine, SVM for regression, Relationship to logistic regression **Examples with R Package**	2½ hours
Time Series Modelling	Introduction to time series data with examples. Components of a time series (trend, seasonality etc.). Stationary time series - Autocorrelation and partial autocorrelation functions. Forecasting. AR, MA, ARMA and ARIMA models. Forecasting by using AR, MA and ARMA models. Moving average and Exponential smoothing for forecasting. Measure of forecasting accuracy. **Examples with R Package**	7½ hours
Sampling and Resampling Methods	Bootstrapping and 'bagging' of bootstrap estimates. Resampling methods for Big Data	2½ hours
Predictive Analytics	Based on Regression Analysis, Classification and on Time Series Data	2½ hours
Multi Criteria Decision Making (MCDM) Methods	• AHP (Analytical Hierarchy Process) • TOPSIS (Technique for Order Preference by the Similarity to Ideal Solution) for prioritisation of Competing candidates (objects / subjects / projects etc.)	2½ hours
Real-life Applications: Project work	• Real-life Big Data sets identified by participants from their respective areas of work/ professional interest • Identification of appropriate data analytics problems on these data sets with the help of faculty members	2½ hours

Mode of Evaluation of Performance of Participants

Continuous evaluation by faculty in the form of assignments/projects.

Appropriate certificates will be awarded to the participants at the end of the course as per criteria specified below:

CRITERIA	AWARD
Satisfactory Performance* in at least 50% of the topics AND attendance in at least 50% of the sessions	Certificate of Proficiency
Satisfactory Performance* in less than 50% of the topics AND attendance in at least 50% of the sessions	Certificate of Attendance
Satisfactory Performance* in at least 50% of the topics AND attendance in less than 50% of the sessions	Certificate of Participation
All other cases	NO Certificate

*Through assessment based on assignments/test/project

Faculty

The list of resource persons for this course is likely to include

1. Ayanendranath Basu, Professor (HAG), Indian Statistical Institute, Kolkata
2. Mausumi Bose, Professor (HAG), Indian Statistical Institute, Kolkata
3. Amita Pal, Professor, Indian Statistical Institute, Kolkata
4. Rita SahaRay, Professor, Indian Statistical Institute, Kolkata
5. Diganta Mukherjee, Professor, Indian Statistical Institute, Kolkata
6. Utpal Garain, Professor, Indian Statistical Institute, Kolkata

Seats

Minimum: 15

Maximum: 30

Technical requirement at far-end/participants' end

A personal laptop or desktop for exclusive use of the participant with functional webcam, microphone and speakers, having the specified minimal system requirements of the online platform to be used.

Video conferencing/online platform to be used by the institute

The latest version of Zoom, with a license, if necessary, that is appropriate for the number of participants and the duration of the course, for live streaming of lectures to the devices of the participants.