

# e-ITEC Course on BIG DATA ANALYTICS FOR POLICY PLANNERS

## Name of Course

**BIG DATA ANALYTICS FOR POLICY PLANNERS**

## Stream

Specialized Courses

## Country

Countries in Africa, the Middle East and the Far East, preferably those lying the time zones UTC+2 to UTC+7, like

Tanzania, Sudan, the Democratic Republic of the Congo, Kenya, Ethiopia, Zambia, Libya, Egypt, Mozambique, Malawi, Zimbabwe, Myanmar, Thailand, Cambodia, Malaysia, Indonesia, Laos, Vietnam, and so on.

## Start date

Thursday, March 11, 2021

## End date

Wednesday, April 07, 2021

## Schedule of classes

**FOUR WEEKS**, with three hours (16:00 IST to 19:00 IST) of lecture and/or interactive session 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 the participants*

- 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 proposed host agency from the country

In the prevailing COVID-19 pandemic scenario, it may not be necessary to identify a host agency, since it is proposed to stream live classes to participants' devices instead of delivering these in a classroom setting.

The proposing agency, namely, the **International Statistical Education Centre (ISEC)** is a sister institution of the Indian Statistical Institute and has been providing statistical education to international candidates for 70 years. The teaching support is provided entirely by the Indian Statistical Institute, which is acknowledged all over the world as a premier institute of statistical learning.

## Course Content

Topics	Syllabus	Lecture hours
<b>Introduction</b>	<ul style="list-style-type: none"> <li>• Introduction to statistical decision-making. Need for data driven decisions.</li> <li>• Concepts of supervised and unsupervised learning. Predictive Analytics</li> <li>• Big Data and their features. Capture, Screening and Storage of Big Data. Administrative Records as Big data.</li> </ul>	3 hours
<b>Dimension Reduction and Data Visualisation</b>	<ul style="list-style-type: none"> <li>• Principal Component Analysis</li> <li>• Cluster Analysis</li> <li>• Multidimensional Scaling</li> </ul> <b>Examples with R Package</b>	3 hours
<b>Linear Regression Model</b>	<p><b>Multiple linear regression:</b> Least squares estimation of the regression coefficients, Test of significance of regression coefficients, prediction of new observations etc. Inclusion of qualitative regressors.</p> <p><b>Variable selection and model building:</b> Stepwise regression methods (Forward and backward selection).</p> <p><b>Multicollinearity:</b> Effects of multicollinearity, Ridge regression and principal component regression.</p> <p><b>Model diagnostics, Model adequacy checking.</b></p> <p><b>Examples with R Package</b></p>	9 hours
<b>Generalized Linear Regression Model</b>	<p><b>Logistic regression:</b> Model fitting, interpretation of the coefficients in a logistic regression model, Odds ratio in logistic regression</p> <p><b>Poisson regression:</b> Model fitting and interpretation.</p> <p><b>Examples with R Package</b></p>	3 hours

<b>Classification</b>	An overview of classification, Linear and quadratic discriminant function, Classification for normal populations. K-nearest neighbour classifier, Naïve Bayes classifier, Classification using logistic regression.  <b>Examples with R Package</b>	9 hours
<b>Cross-validation</b>	The validation Set Approach, Leave One-Out Cross-Validation, k-fold Cross-validation  <b>Examples with R Package</b>	3 hours
<b>Tree based methods</b>	Background, Regression trees and classification trees  <b>Examples with R Package</b>	6 hours
<b>Support Vector Machine</b>	Overview of support vector classifier, Support vector machine, SVM for regression, Relationship to logistic regression  <b>Examples with R Package</b>	3 hours
<b>Time Series Modelling</b>	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.  <b>Examples with R Package</b>	9 hours
<b>Sampling and resampling methods</b>	Bootstrapping and 'bagging' of bootstrap estimates.	3 hours
<b>Predictive Analytics</b>	Based on Regression Analysis, Classification and on Time Series Data	3 hours
<b>Multi Criteria Decision Making (MCDM) Methods</b>	<ul style="list-style-type: none"> <li>• AHP (Analytical Hierarchy Process)</li> <li>• TOPSIS (Technique for Order Preference by the Similarity to Ideal Solution)</li> </ul> for prioritisation of Competing candidates (objects / subjects / projects etc.)	3 hours
<b>Applications</b>	<ul style="list-style-type: none"> <li>• Real-life Optimisation problems and meta-heuristics.</li> <li>• Uses of Big Data in Policy Planning Exercises.</li> </ul>	3 hours

## Mode of Evaluation of Performance of Participants

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

## Faculty

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. Smarajit Bose, Professor, Indian Statistical Institute, Kolkata

5. Md. Zafar Anis, Sr. Technical Officer, Indian Statistical Institute, Kolkata  
(Course In-charge)

### 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 Google Meet/ Webex, with a license, if necessary, that is appropriate for the number of participants and the duration of the course, for streaming live classes to the devices of the participants.

---