Title: A Guided Tour of Practical Data Science with R

Prerequisites:

- Keen interest and passion for ideas of data science and sincere desire to find patterns in the data.
- Basic college algebra and familiarity with the computer environment.
- No prior knowledge of statistical computing required, but again familiarity with computers essential

Description: Upon completing this 10 hour course, you should be able to claim a hands-on awareness of the following:

• [**Preparation and stat laboratory setup**] Foundational working knowledge of R and R Studio as environments for exploring Applied Statistical Machine Learning and Practical Data Science. Basic introduction to the R Graphical User Interface (GUI) environment with R Commander and Rattle along with the Installation of some key R packages and most common CRAN Task Views. (**1 hour**)

• [**Basic statistics and statistical computing**] Elements of Exploratory Data Analysis (EDA) featuring most commonly used plots, graphs and statistical summaries along with distributional assessments, These two hours are dedicated in part to helping the audience get familiar with some key statistical concepts that ubiquitously permeate Statistical Machine Learning, Data Science and Artificial Intelligence (**2 hours**)

• [**Regression Learning**] Discovering practical regression with hands-on examples in R, plus fundamentals of interpreting regression outputs from basic concepts. This segment is used to introduce some of the most powerful machine learning concepts like Cross Validation, Bootstrap, Information Criteria and Bias-Variance trade-off. Motivating examples and exercises are drawn from a wide variety of fields. Linear regression is explored extensively, but a few other regression learning methods are touched upon, all in a hands-on manner driven by datasets. (**2 hours**)

• [**Classification learning**] Encountering the power of Pattern of Recognition and Classification with a touch on the key concepts. Key concepts are presented within the first 30 minutes, and the rest of the 2 hours of this segment are dedicated to a tour of the most commonly used classification method, applied on several different datasets and predictively compared, using both comparative boxplots and ROC Curves (kNearest Neighbors, Classification Tress, Logistic Regression, Support Vector Machines, Boosting and Random Forest are applied and compared various datasets) (**2 hours**)

• [**Unsupervised Learning**] Clustering, Dimensionality Reduction, Feature Extraction and Novelty Detection are explored here. Finding groups and patterns in various types of data via cluster analysis and other unsupervised learning, with compelling R examples. KMeans clustering is visited, but also Hierarchical Clustering and PAM + a bit of Recommender Systems and Nonnegative Matrix Factorization (**2 hours**)

• [**Time Series**] Here we touch on the ARIMA family of models for analyzing time series but we also touch on emerging nonparametric methods for time series analysis. A taste of predictive modelling and elements of forecasting (**1 hour**)

Instructor's suggestion: To make the most of this workshop, it is ideal for all attendees to be active participants diligently exploring along with the instructors. Exercises are given and all are expected to explore them.

Important datasets
CiFAR 10
read.csv('gifted.csv')
attitude
longley
bodyfat
soccer
NFL and third down conversion
golf

prostate
software engineering datasets
astronomy and astrophysics datasets
music data sets from qiuyi
epileptic seizure datasets
Turkiye student evaluation data sets
biology dataset from evolutionary molecular biology
datasets from the psych packages
ruspini datasets
religion datasets from Preeti
NIPS datasets
New York Times datasets
Gutenberg datasets
Electric grid BPA datasets
Datasets recommended in Rwanda (see gmail trail)
Niranjana datasets for High Performance computing
ORL faces dataset
Yale faces dataset
MNIST Data from USPS
Fashion MNIST
Audio dataset from RIT
Time Series Data Sets
Body gesture datasets from STAT 747
Amanda datasets for DNA dataset in textual (nominal) form
DNA MicroArray gene expression datasets
Spam
E-coli
reuters
Pima.Indian
European Jobs dataset
Vote.repub
AirPassengers
Sunspots Vowel Phoneme
SAHeart
Ozone

Important R Packages

```
library(ctv)
install.packages(psych)
install.views(MachineLearning)
install.views(HighPerformanceComputing)
Install.packages(`ggplot2`)
install.packages('rstan')
install.views('Bayesian')
 install.views('Cluster')
install.views('Robust')
install.views('Survival')
install.views('TimeSeries')
install.views('Psychometrics')
library(MASS)
library(RandomForest)
library(adabag)
library(ipred)
library(class)
library(e1071)
library(ElemStatLearn)
library(neuralnet)
library(glmnet)
library(elasticnet)
library(lasso2)
library(keras)
```

Explore the fascinating world of CRAN Views

https://cran.r-project.org/web/views/

Image Processing and Computer Vision with R
Text Analytics and Natural Language in R
Psychology and sociology with R
Marketing and Market Segmentation with R
Biology with R Medical Statistics and Medical Diagnosis with R
Astrostatistics with R