PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100 Feet Ring Road, BSK III Stage, Bengaluru-560 085
Department of Computer Science and Engineering
Session : Aug-Dec 2019

UE18CS203:Introduction to Data Science

Course Evaluation Scheme

Sl. No.	Details	Marks	Reduced to	Final Marks
1	ISA/CBT-1	40	**15**	
2	ISA/CBT-2	40	**15**	
3	**Assignment-IDS Project**	40	**10**	
	In Semester Assessment (ISA)			**40**
4	ESA – pen and paper	100	**60**	
	End Semester Assessment (ESA)			**60**
	Total			**100**

As part of the course Assignment, students are expected to do a project.
The details of how this project must be done and evaluated is described below.

Assignment Objective: To gather and analyze a data set. Perform Exploratory Data Analysis.

What is Exploratory Data Analysis (EDA)?

- How to ensure you are ready to use machine learning algorithms in a project?
- How to choose the most suitable algorithms for your data set?
- How to define the feature variables that can potentially be used for machine learning?

Exploratory Data Analysis (EDA) helps to answer all these questions, ensuring the best outcomes for the project. It is an approach for summarizing, visualizing, and becoming intimately familiar with the important characteristics of a data set.

Exploratory Data Analysis (EDA) is the first step in your data analysis process. Here, you make sense of the data you have and then figure out what questions you want to ask and how to frame them, as well as how best to manipulate your available data sources to get the answers you need.

You do this by taking a broad look at patterns, trends, outliers, unexpected results and so on in your existing data, using visual and quantitative methods to get a sense of the story this tells. You're looking for clues that suggest your logical next steps, questions or areas of research.

Developed by John Tukey in the 1970s, exploratory analysis is often described as a philosophy, and there are no hard-and-fast rules for how you approach it. EDA is used to tackle specific tasks such as:

- Spotting mistakes and missing data;
- Mapping out the underlying structure of the data;
- Identifying the most important variables;
- Listing anomalies and outliers;
- Testing a hypotheses / checking assumptions related to a specific model;
- Establishing a parsimonious model (one that can be used to explain the data with minimal predictor variables);
- Estimating parameters and figuring out the associated confidence intervals or margins of error.

Value of Exploratory Data Analysis

Exploratory Data Analysis is valuable to data science projects since it allows to get closer to the certainty that the future results will be valid, correctly interpreted, and applicable to the desired business contexts. Such level of certainty can be achieved only after raw data is validated and checked for anomalies, ensuring that the data set was collected without errors. EDA also helps to find insights that were not evident or worth investigating to business stakeholders and data scientists but can be very informative about a particular business.

EDA is performed in order to define and refine the selection of feature variables that will be used for machine learning. Once data scientists become familiar with the data set, they often have to return to feature engineering step, since the initial features may turn out not to be serving their intended purpose. Once the EDA stage is complete, data scientists get a firm feature set they need for supervised and unsupervised machine learning.

Tools and Techniques

Among the most important statistical programming packages used to conduct exploratory data analysis are S-Plus , R and Python libraries for data analysis.
(**NumPy**,**SciPy**,**Matplotlib**,**Pandas**,**ScikitLearn**,**Statsmodels**,**Seaborn**,**Bokeh**,**Blaze**,**Scrapy**,**Requests**,**BeautifulSoup**)

Sample Exploratory Data Analysis Case Studies

https://www.kaggle.com/c/house-prices-advanced-regression-techniques

http://ucanalytics.com/blogs/exploratory-data-analysis-retail-case-study-example-part-3/

Project Teams

The project will be done by a *group of 3 students (3 and no more or no less)*. One or two teams in a

class may be an exception (one/two teams of two members each) with prior approval of class teacher. *All teams must be among students belonging to the same section. (No teams are allowed to span multiple sections)*

Data Sets :

Look for data sets online like

https://www.kaggle.com/datasets

https://www.tableau.com/learn/articles/free-public-data-sets

Detailed Schedule/Milestones

Sl. No.	Task	Deliverable	Week No.
1	Team Formation	Team Forms in Google+Printed Copy to Class teacher.	2
2	Dataset selection	See the Guidelines+Approval from Class teacher	4
3	**Final Evaluation**	**Presentatiom, Report, Viva (to the class Teacher)**	14

Note:Check the documents

1. Guidelines for the IDS project assignment
2. Evaluation Scheme for the IDS project assignment

-