GUJARAT TECHNOLOGICAL UNIVERSITY

Syllabus for Master of Computer Applications, 4th Semester
Subject Name: Software Project -2 (Data Science)
Subject Code: 4649304

With effective from academic year 2018-19

1. Learning Objectives:

- To be able to develop Data Science Project using open source technologies
- To learn Data Processing, Visualization and Analytical techniques on data set

2. Prerequisites: Knowledge of Programming Language, Python and Database concepts

3. General Guidelines for Data Science Project using Open source Technologies

1) Group: 2-3 Person.
2) The project should be free from plagiarism of any kind.
3) It is mandatory that the project should be developed using Python on Linux or Windows Platform. Linux is the preferred platform
4) Project must have proper documentation
5) This may not be a live project
6) Use of a data set is mandatory
7) Use of Data Visualization and Analytical methods is mandatory
8) Use of any popular Libraries/Framework based on Python is not prohibited, in which case the framework internals should also be known.

4. Knowledge about the following is expected to be demonstrated.

1) Proper knowledge about the purpose of the application
2) Use of Large Data set
3) Use of Data pre-processing (Cleaning, Dimension reduction etc.)
4) Use of Data Visualization
5) Use of Data Analytical algorithm

5. Expected Outcome

1) The objective of the Data Science Project Development is to make students aware about the industry based process and workings. As a result, Project must meet with the industry standards.
2) There will not be any compulsion to prepare a project report for the students but an application and supportive documents should be self-explanatory, so that evaluator may get the detail about the Project developed and can evaluate the students as per the evaluation criteria.
- Group size: 2-3 Persons.

3) Power Point Presentation Content (30 Slides Max.):

Business Objective	Introduction, Problem Statement
Plagiarism Report	Similarity should be less more than 30%
Understand Data	About Data Source Understand Data: Basic Questions Understand Data: Data Wrangling Understand Data: Exploratory Analysis

Methodology	Extract Features & Model, Methodology
Data Visualization	Implementation of Data visualization Techniques
Present Results	Present Results
Conclusion	Conclusion
References	References
YouTube Link (*)	

Note (*): It is preferable to have project Presentation uploaded on YouTube.

6. Suggested

PS: Below list (a & b) are suggestive one. You may select any other relevant topics/Data Sets.

a) Project Definitions

1) A Study on Employee Attrition Prediction and Analysis
2) A Study on Student Dropout Prediction and Analysis
3) A Study on Student Result Prediction and Analysis
4) A Study on Heights and Weights Data
5) A Study on Loan Prediction and Analysis
6) A Study on Housing Data
7) A Study on Weather Data
8) A Study on Movie Lens (https://movielens.org)
9) A Study on Trip Data
10) A Study on Census and Income Data
11) A Study on Songs Data
12) A Study on Sales Data
13) A Study on Online Shopping Data
14) A Study on Cyber Crime Data
15) A Study on Airline Safety
16) A Study on Spam emails /Get rid of Spam emails ()
17) A Study on Pictures / Working with Pictures
18) Working with Handwritten Information
19) Analyzing Reviews (e.g. amazon.com)

b) Suggested Data Sets

1) Kaggle Data Sets (https://www.kaggle.com/datasets)
2) Data.gov Data Sets (https://data.gov.in/https://www.data.gov/)
3) kdnuggets Data Sets (https://www.kdnuggets.com/datasets/index.html)
4) Titanic Data Set. (https://www.rdocumentation.org/packages/titanic/versions/0.1.0)
5) Boston Housing Data Set. (https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html)
6) Walmart Sales Forecasting Data Set. (https://data.world/just4jcgeorge/walmart-sales-datahttps://relational.fit.cvut.cz/dataset/Walmart)
7) Hubway Data Visualization Challenge. (**https://anitagraser.com/projects/hubway-data-visualization-challenge/)**

8) Text Mining Data Set. (https://archive.ics.uci.edu/ml/datasets.htmlhttp://www.rdatamining.com/resources/datahttps://searchbusinessanalytics.techtarget.com/feature/Simple-data-mining-examples-and-datasets)
9) Census Income Data Set. (https://www2.1010data.com/documentationcenter/prod/Tutorials/MachineLearningExamples/CensusIncomeDataSet.html)
10) Movie Lens Data Set. (http://files.grouplens.org/datasets/movielens/ml-latest-small-README.htmlhttp://files.grouplens.org/datasets/movielens/ml-latest-README.html)
11) Yelp Data Set. (https://www.yelp.com/datasethttps://catalog.data.gov/dataset/yelp-datahttps://github.com/vc1492a/Yelp-Challenge-Dataset)
12) AWS Public Data Sets (https://aws.amazon.com/opendata/https://registry.opendata.aws/)
13) Google Public Data Sets (https://cloud.google.com/public-datasets/https://www.google.com/publicdata/directoryhttps://ai.google/tools/datasets/)
14) Wikipedia Data Sets (https://en.wikipedia.org/wiki/Wikipedia:Database_downloadhttps://meta.wikimedia.org/wiki/Datasetshttps://wiki.dbpedia.org/data-set-34https://piktochart.com/blog/8-useful-databases-to-dig-for-data/https://perso.telecom-paristech.fr/eagan/class/igr204/datasets)
15) UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php)
16) World Bank Data Sets (https://data.worldbank.org/indicatorhttp://wdi.worldbank.org/tableshttps://openknowledge.worldbank.org/http://www.worldbank.org/en/news/feature/2012/03/29/world-bank-more-open-accessible-and-searchable-data-support-vietnams-developmenthttps://twitter.com/worldbankdata?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor
17) Academic Torrents Data Sets (http://academictorrents.com/browse.php?cat=6http://academictorrents.com/collection/nasa-datasetshttps://www.techleer.com/articles/545-academic-torrents-a-distributed-system-for-sharing-enormous-datasets/)
18) Twitter Data Set (https://twitter.github.io/typeahead.js/examples/https://github.com/twitter/typeahead.js/issues/1004https://github.com/guyz/twitter-sentiment-datasethttps://data.world/datasets/twitterhttps://old.datahub.io/dataset/twitter-sentiment-analysis
19) https://vincentarelbundock.github.io/Rdatasets/datasets.html
20) https://www.dataquest.io/blog/free-datasets-for-projects/

7. Evaluation

Sr. No	Particulars	Weightage
1	Topic & Selection of Algorithm	10%
2	Data Pre-processing (Cleaning, Reducing Dimensionality)	30%
3	Data Visualization	20%
4	Data Analysis / Algorithm	30%
5	Result	10%

Recommended Book(s):

1) Field Cady, 'The Data Science Handbook ', Wiley Publication ISBN-13: 978-1119092940
2) Jake VanderPlas, 'Python Data Science Handbook ESSENTIAL TOOLS FOR WORKING WITH DATA', O'REILLY ISBN:978-1-491-91205-8
3) Rachel Schutt and Cathy O'Neil, Doing Data Science, O'REILLY
4) Wes McKinney,Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython, 2nd Edition , O'REILLY
5) Anand Rajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", Cambridge University Press, 2012
6) John W. Foreman (Author), Data Smart: Using Data Science to Transform Information into Insight, WILEY
7) John Paul Mueller, Luca Massaron, Python for Data Science For Dummies , WILEY

Note: Some of the practicals form the above practical list may have seemingly similar definitions. For better learning and good practice, it is advised that students do maximum number of practicals. In the practical examination, the definition asked need not have the same wordings as given in the practical list. However, the definitions asked in the exams will be similar to the ones given in the practical list.