GUJARAT TECHNOLOGICAL UNIVERSITY

Data Science
SUBJECT CODE: 3710219

Type of course: Elective

Prerequisite: Data Structures, Basics of Probability and Statistics

Rationale: Data Science is a blend of many fields, including many sub domains of mathematics, computer science, computational science, statistics, and information science. In contrast to "pure" mathematicians, statisticians, or computer and information scientists, a data scientist has a breadth of experience across all of these fields, but may not have as much knowledge as a specialist in any particular field. This subject will help students to efficiently conduct computational analysis with their own knowledge domain.

Teaching and Examination Scheme:

Teaching Scheme			Credits	Examination Marks				Total Marks
L	T	P	C	Theory Marks		Practical Marks		
				ESE(E)	PA (M)	PA (V)	PA (I)	
3	0	2	4	70	30	30	20	150

Content:

Sr. No.	Content	Total Hrs	% Weightage
1	**An Introduction to core concepts & technologies:** Introduction, Terminology, data science process, data science toolkit, Types of data, Example applications.	6	10%
2	**Data collection and management:** Introduction, Sources of data, Data collection and APIs, Exploring and fixing data, Data storage and management, Using multiple data sources	7	15%
3	**Data analysis:** Introduction, Terminology and concepts, Introduction to statistics, Central tendencies and distributions, Variance, Distribution properties and arithmetic, Samples/CLT, Basic machine learning algorithms, Linear regression, SVM, Naive Bayes.	10	25%
4	**Data visualisation:** Introduction, Types of data visualisation, Data for visualisation: Data types, Data encodings, Retinal variables, Mapping variables to encodings, Visual encodings.	11	25%
5	Applications of Data Science, Technologies for visualisation, Bokeh (Python)	7	15%
6	Recent trends in various data collection and analysis techniques, various visualization techniques, application development methods of used in data science.	7	10%

Reference Books:

1. Doing Data Science, Cathy O'Neil and Rachel Schutt, Straight Talk From The Frontline. O'Reilly.
2. Introduction to Data Science, Davy Cielen, Arno D B Meysman and Mohamed Ali, Manning, dreamtech press
3. Practical Data Science, Nina Zumwl and John Mount, Manning, dreamtech press
4. The Data Science Handbook, Field Cady, Wiley
5. Getting Started with Data Science, Murtaza, Haider, Pearson
6. Data Science and Big Data Analytics, EMC Education Services, Wiley
7. Data Science, John D Kellehar, MIT Press
8. Mining of Massive Datasets. v2.1, Jure Leskovek, AnandRajaraman and Jeffrey Ullman, Cambridge University Press

Course Outcome:

After learning the course the students should be able to:

- Explain how data is collected, managed and stored for data science;
- Understand the key concepts in data science, including their real-world applications and the toolkit used by data scientists;
- Implement data collection and management scripts using MongoDB

List of Experiments

- Minimum 10 experiments based on the contents.
- Mini Project in a group of max. 3 students
- Writing a research paper on selected topic from content with latest research issues in that topic

Major Equipments:

- Modern System with related software

List of Open Source Software/learning website:

https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/

https://www.rstudio.com/online-learning/

Additional Resources:

Books for Unit 1 and 2:

1. Data Mining Concepts & Techniques, J Han, M Kamber, J Pei ((chapter 2 & 3)
2. Data science process flowchart from "Doing Data Science", Cathy O'Neil and Rachel Schutt, 2013 (chapter 2)
3. Data Science and Big Data Analytics, EMC Education Services, Wiley

Unit 1: An introduction to core concepts and technologies

https://www.edureka.co/blog/what-is-data-science/

https://intellipaat.com/blog/what-is-data-science/

Data types:

https://www.youtube.com/watch?v=hZxnzfnt5v8
https://www.youtube.com/watch?v=zHcQPKP6NpM&t=247s
https://www.youtube.com/watch?v=zHcQPKP6NpM&t=247s
http://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/

Tools:
https://www.ngdata.com/top-tools-for-data-scientists/

8 open Source Big Data Tools to use in 2018:

https://towardsdatascience.com/8-open-source-big-data-tools-to-use-in-2018-e35cab47ca1d

Basic Libraries for Data Science:
https://www.upwork.com/hiring/data/15-python-libraries-data-science/

Unit 2: Data collection and management

Data collection:
http://bigdata-madesimple.com/3-effective-methods-of-data-collection-for-market-research/
Data Wrangling with example:
https://towardsdatascience.com/intro-to-data-science-part-2-data-wrangling-75835b9129b4
https://medium.fr7eecodecamp.org/discovering-the-secrets-of-baseball-with-data-56f793852de0
Data Analysis with example:
https://medium.com/@williamkoehrsen/data-analysis-with-python-19434f5d6324
Data cleaning with example:
https://www.kdnuggets.com/2016/03/doing-data-science-kaggle-walkthrough-cleaning-data.html
5 APIS a data scientist must know:
https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-apis-application-programming-interfaces-5-apis-a-data-scientist-must-know/

Data storage

https://searchstorage.techtarget.com/definition/big-data-storage

http://www.enterprisestorageforum.com/storage-management/storage-trends/top-10-trends-for-data-storage-with-big-data-analytics.html

https://www.computerweekly.com/tip/Big-data-storage-management-challenges-and-how-to-deal-with-them

Multiple data sources:

https://www.allerin.com/blog/top-5-sources-of-big-data

http://tdan.com/combining-data-from-multiple-sources-join-integrate-blend/19877

https://www.techrepublic.com/blog/big-data-analytics/use-normalization-and-etl-to-get-the-big-data-results-you-want/

https://www.youtube.com/watch?v=f0nMfV1GvOg

Books for Units 3 to 6

Book1: Data Mining Concepts and Techniques by Jiawei Han, MichelineKamber and Jian Pei

Book2: Statistics and Data Analysis by A. Abebe (available online in .pdf format

Unit 3 Data Analysis

- For Introduction, Terminology and Concepts
  - Chapter 3 of Book1 for Data analysis process
  - https://www.tutorialspoint.com/excel_data_analysis/data_analysis_overview.htm
  - https://www.tutorialspoint.com/excel_data_analysis/data_analysis_process.htm
- Introduction to statistics, central tendencies and distributions, Variance, distribution properties and arithmetic
  - Section 2.2 (Basic Statistical Descriptions of Data) of book
  - http://statistics.wikidot.com/ch3
  - https://www.listendata.com/2014/04/descriptive-statistics.html
- Central Limit Theorem (CLT)
  - Chap. 6 from Book2
  - https://web.stanford.edu/class/archive/cs/cs109/cs109.1178/lectureHandouts/190-central-limit-theorem.pdf
  - https://towardsdatascience.com/understanding-the-central-limit-theorem-642473c63ad8
  - https://www.tutorialspoint.com/statistics/central_limit_theorem.htm
- Basic Machine Learning Algorithms, Linear regression, SVM, Naïve Bayes
  - **Machine Learning**
    - https://www.geeksforgeeks.org/machine-learning/(What is machine learning, applications of machine learning, classification of machine learning methods)
  - **Naïve Bayes**
    - Section 8.3 from Book1
    - https://www.geeksforgeeks.org/naive-bayes-classifiers/
  - **SVM**
    - Section 9.3 from Book1
    - https://machinelearningmastery.com/support-vector-machines-for-machine-learning/
    - https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/
  - **Linear Regression**
    - http://cs229.stanford.edu/notes/cs229-notes1.pdf
    - https://www.geeksforgeeks.org/linear-regression-python-implementation/
    - http://ufldl.stanford.edu/tutorial/supervised/LinearRegression/

Unit 4 Data Visualization

- Introduction
  - Section 2.3 from Book1
- Types of data visualization
  - https://info.datalabsagency.com/blog/data-visualization-news/15-most-common-types-of-data-visualisation , https://datavizcatalogue.com/
- Data for visualization
  - Data types (already covered in Unit 1)
  - Data Encoding

        - https://www.oreilly.com/library/view/designing-data-visualizations/9781449314774/ch04.html
        - http://paldhous.github.io/ucb/2016/dataviz/week2.html
        - http://www.faculty.jacobs-university.de/llinsen/teaching/340131/Lecture03.pdf
- Retinal variables, mapping variables to encoding, visual encoding
    - https://www.targetprocess.com/articles/visual-encoding/
    - http://vda.univie.ac.at/Teaching/Vis/13s/LectureNotes/05_visual_encodings.pdf
    - https://www.cs.sfu.ca/~torsten/Teaching/Cmpt467/LectureNotes/05_visual_mappings.pdf

Unit 5

- Applications of Data Science – Applications in healthcare, finance, ecommerece, education, and agriculture can be covered.
    https://www.analyticsvidhya.com/blog/2015/09/applications-data-science/
    - Healthcare:
        - https://medium.com/activewizards-machine-learning-company/top-7-data-science-use-cases-in-healthcare-cddfa82fd9e3
        - https://www.datapine.com/blog/big-data-examples-in-healthcare/
        - http://article.sciencepublishinggroup.com/pdf/10.11648.j.ajtab.20180402.14.pdf
    - Finance
        - https://www.mastersindatascience.org/industry/finance/
        - https://www.techemergence.com/predictive-analytics-in-finance/
    - E-commerce:
        - https://towardsdatascience.com/5-data-science-project-every-e-commerce-company-should-do-8746c5ab4604
        - https://www.datascience.com/blog/data-science-for-ecommece-businesses-predictive-modeling
        - https://dataconomy.com/2017/07/6-ways-use-big-data-ecommerce/
    - Education:
        - https://www.expresscomputer.in/magazine/data-analytics-in-education-sector-to-see-high-growth/14468/
        - https://www.analyticsindiamag.com/top-6-ways-make-education-institutions-smarter-data-analytics/
        - https://www.allerin.com/blog/4-ways-big-data-is-transforming-the-education-sector
    - Agriculture:
        - https://www.analyticsvidhya.com/blog/2018/05/data-analytics-in-the-indian-agriculture-industry/
        - https://www.wur.nl/upload_mm/6/0/4/307c3061-35ea-4339-a33b-d21f047d2d38_Wolfert%20et%20al%20Big%20Data%20in%20Smart%20Farming.pdf
        - https://www.sciencedirect.com/science/article/pii/S0308521X16303754
- Technologies for visualization
    - https://tdwi.org/articles/2011/11/09/research-excerpt-data-visualization-technology.aspx
- Bokeh (Python)
    - https://bokeh.pydata.org/en/latest/
    - https://www.analyticsvidhya.com/blog/2015/08/interactive-data-visualization-library-python-bokeh/

Unit 6

- Recent trends in various data collection techniques - https://www.tutorialspoint.com/statistics/data_collection.htm https://avaresearch.com.au/different-types-of-data-collection-methodologies/
- Various visualization techniques – already covered in Unit 4
- Application development methods used in data science
    - Python Programming
    - R Programming

    **Students must be able to implement concepts learned in data science (concepts learned in previous units) using Python and R programming