NATIONAL UNIVERSITY
of Computer & Emerging Sciences, Lahore

Department of Computer Science

DS500 – Introduction to Data science: Tools and Techniques

SPRING 2019

Instructor Name: Dr. Irfan Younas **TA Name (if any):**
Email address: irfan.younas@nu.edu.pk **Email address:**
Office Location/Number: C-146 **Office Location/Number:**
Office Hours: Monday 3:30-4:30 PM, Thursday 3:30 to 4:30pm

Course Information

Program: MS **Credit Hours:** 3 **Type:** Elective
Pre-requisites (if any): Programming competence, Discrete Maths, Linear Algebra, Probabilty& Statistics

Course Website (if any) : piazza.com/fast_lahore/spring2019/cs500, **access code:** cs500
Class Meeting Time: Monday, Wed 12:30 – 1:50 PM
Class Venue: CS-01

Course Description/Objectives/Goals:

- To introduce students to data science and the Data Science program being offered
- To give overview of the data, questions, and tools that data analysts and data scientists work with
- To understand foundational principles such as using data to get information about an unknown quantity of interest, calculating and using data similarity, data summarization, data visualization etc.
- To build the data-analytics mentality such that should learn to understand a phenomenon better and especially to make better-informed decisions
- To give overview of general approach to solve the data science problems and how to manage the data science cycle

including getting data, cleaning data, data preprocessing, data visualization, selection of right algorithm or methodology to solve the problem, fitting models to data, evaluation and model analytics

- To understand practical issues in statistical computing which includes programming in R and python

Course Learning Outcomes (CLOs):		
At the end of the course students will be able to:	**Domain**	**BT* Level**
understand the concepts of Data Science,		
prepare and wrangle the data for analysis,		
perform exploratory data analysis to investigate data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations,		
understand and apply machine learning algorithms to gain insight from the data		

* BT= Bloom's Taxonomy, C=Cognitive domain, P=Psychomotor domain, A= Affective domain.

Bloom's taxonomy Levels: 1. Knowledge, 2. Comprehension, 3. Application, 4. Analysis, 5. Synthesis, 6. Evaluation

Textbook(s) /Supplementary Readings:

There is no standard one "textbook" for this course. The following book will be used as a primary text to guide some of the discussions, but it will be heavily supplemented with lecture notes and reading assignments from other sources.

Cathy O'Neil and Rachel Schutt. Doing Data Science, Straight Talk From The Frontline. O'Reilly. 2014. ISBN 978-1-449-35865-5.

Additional references and books related to the course:

Jure Leskovek, Anand Rajaraman and Jeffrey Ullman. Mining of Massive Datasets. v2.1, Cambridge University Press. 2014. (Free online.)

Jiawei Han, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques, Third Edition. Morgan Kaufmann Publishers. 2012. ISBN 978-0-12-381479-1.

Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. MIT Press. 2013. ISBN 0262018020. (Online info available here.)

Foster Provost and Tom Fawcett. Data Science for Business: What You Need to Know about Data Mining and Data-analytic Thinking. O'Reilly 2013. ISBN 978-1-449-36132-7.

Contents:

- The life cycle of a data science project, understanding uncertainty in its outcome, types of data and datasets, Measurement and data acquisition and storage issues.

- Introduction: What is Data Science?

- Statistical Inference and (Python): Populations and samples, Statistical modeling, probability distributions, fitting a model, Intro to Python

- Data pre-processing stages which includes Data wrangling, Sampling, Dimensionality Reduction, Feature subset selection etc.

- Machine Learning Algorithms: Linear Regression, Gradient Descent, Logistic Regression, Regularization, Advice for apply ML and ML System Design, Support Vector Machines, Dimensionality Reduction (Principal Component Analysis), outliers detection, Clustering Algorithms

- Exploratory Data Analysis

- Data Visualization: Basic principles, ideas and tools for data visualization, Examples of inspiring projects

- Python (Anaconda) Installation, python packages, pip, conda and other related commands, Python data types and structures (variables, array, list, dictionary), sequences, Basic control structures (do while for if else) functions, exception handling, stack trace, Indexing, loading and querying Data frames, Merge DataFrames, generate summary tables, grouping data, packages (Numpy, Pandas)

- Basic descriptive and exploratory analysis, market basket analysis and the importance of correlation in gaining business insights, Big Data Introoduction.

Grading

Assignments/Quiz(s)/Homeworks 20 - 25 %

Project (Implementation & Presentation) / Research Paper	*10* %
Midterms	22 - 23 %
Final Exam	40 - 45%
Total:	**100 %**