## **NOVA College-wide Course Content Summary**

## **ITD 195 – Introduction to Data Science/Analytics (3 cr.)**

**Course Description**

This course provides a broad survey of Big Data and data analytics, including demos and applications of widely used tools and methods. Topics include descriptive statistics, basic data analysis, common data extraction/translation/loading methods and tools from varied data sources and types, data visualizations, as well as machine learning (supervised and unsupervised). This course includes theory and practice, heavily emphasizing practical applications through case studies. Lecture 3 hours per week.

**General Course Purpose**

To prepare the student to derive meaningful and expressive information from a multitude of raw data sources, including the application of basic statistics, analysis tools and techniques, data extraction and cleaning, creation of visualizations, as well as the application of machine learning to analysis problems.

**Course Prerequisites/Corequisites**

Prerequisite: Math background equivalent to basic statistics, pre-algebra

Recommended: ITP 150 - Python Programming (or Python experience)

**Course Objectives**

1. Describe and use basic statistics on data.
2. Describe and work with datasets from a multitude of sources and formats.
3. Describe and manipulate datasets that are within the definition of “Big Data.”
4. Extract, transfer and clean up data from raw data sources, transforming them into usable forms.
5. Describe and generate various visualizations from raw and derived data.
6. Describe and use supervised and unsupervised machine learning.
7. Define and apply feature engineering techniques in the process of developing machine learning models.

**Major Topics to be Included**

1. Basic descriptive statistics
2. Statistical distributions
3. Data manipulation and cleaning/’wrangling’
4. Big Data theory/applications; extraction and manipulation tools
5. Data visualization
6. Machine learning, supervised and unsupervised
7. Feature engineering

**Student Learning Outcome**

### Explain the purpose of statistics and define:

### qualitative and quantitative variables

### continuous and discrete quantitative variables

### Define a data set

### Define distribution, Gaussian distribution (normal dist.)

### Measuring Central Tendency

### Define and calculate the mean, median and mode

### Measuring Dispersion

### Define and calculate range

### Define and assess skew

* 1. Define variability, outliers, variance (σ2)

### Define and calculate standard deviation (σ)

### Define correlations

### Define independent and dependent variables

### Explain the purpose of and create visualization plots

### Define a random variable and explain random variable distributions

1. Extract/Translate/Load (ETL); Data Wrangling; Basic Analysis
   1. Define and explain the process of extraction, translation and loading (ETL)
   2. Use a relational database and SQL to perform basic ETL
   3. Define and explain methods of data ‘wrangling’ and cleaning
   4. Apply basic tools to perform ETL, data wrangling/cleaning as well as analysis on ‘cleaned’ datasets
2. Identify, define and explain the purpose of machine learning (ML)
3. Define and apply basic feature engineering
   1. Define imputation and use basic imputation techniques
   2. Define nominal and ordinal attributes
4. Define and explain basic machine learning approaches, including regression, decision trees, clustering, etc.
5. Define supervised machine learning terminology
   1. Define and explain the purpose of supervised learning
   2. Demonstrate ability to identify supervised learning algorithms and identify appropriate applications
   3. Define regression as a supervised learning prediction task
   4. Define and explain the purpose of decision tree learning
   5. Define classification and demonstrate ability to identify appropriate applications of classification
   6. Define k-nearest neighbors method for classification prediction tasks
   7. Define bias within linear models
   8. Define and explain the purpose of deep learning
   9. Define bias, variance, overfitting, underfitting
   10. Define and explain the purpose of hyperparameters
   11. Explain the impact of hyperparameters on a model as well as complexity
   12. Define and explain the purpose of weights
   13. Apply supervised learning to analyze and solve real world problems through case studies
6. Define unsupervised machine learning terminology
   1. Define unsupervised learning
   2. Define clustering
   3. Define and explain the purpose of dimensionality reduction
   4. Demonstrate the ability to identify when dimensionality reduction is appropriate
   5. Define and explain the purpose of k-means clustering
      1. Define inter-cluster distance
      2. Define centroids
   6. Apply unsupervised learning to analyze and solve real world problems through case studies
7. Apply statistics, Python and GUI tools, as well as ML theory and applications to analyze real world problems through case studies

**Required Time Allocation per Topic**

To standardize the core topics of this course, the following student contact hours per topic are required. Each syllabus should be created to adhere as closely as possible to these allocations. Topics are not necessarily to be taught in the order shown.

There are normally 45 student contact-hours per semester for a three-credit course (14 weeks of instruction, excluding final exam week: 14\*3.2 = 45 hours). Sections of the course offered in alternative formats (i.e. not standard 15-week) still meet for the same number of contact hours. The final exam is not included in the timetable.

The quickly evolving nature of data analytics means that some content noted in this document may be superseded or made obsolete. As such, it is important to include such changes in individual syllabi. Additionally, time is allocated for additional and optional topics in order to provide instructors flexibility in tailoring the course to special needs or resources.

|  |  |  |
| --- | --- | --- |
| **Topics** | **Hours** | **Percentage** |
| Basic descriptive statistics | 5 | 11% |
| Statistical distributions | 3 | 7% |
| Data manipulation and cleaning/’wrangling’ | 6 | 13% |
| Big data theory/applications; extraction and manipulation tools | 6 | 13% |
| Data visualization | 5 | 11% |
| Machine learning, supervised and unsupervised | 6 | 13% |
| Feature engineering | 5 | 11% |
| Testing to include quizzes, tests and exams (excluding final exam) | 3 | 7% |
| Other optional topics | 6 | 13% |
| **Total** | **45** | **100%** |