**TÊN HỌC PHẦN (COURSE NAME): INTRODUCTION TO DATA SCIENCE**

**1.Thông tin về học phần (General Information)**

**Tên học phần (Course name)**: Introduction to Data Science

**Mã học phần (Course code)**: INT\_E14120

**Số tín chỉ (Number of credits)**: 3

**Loại học phần (Course type)**: Compulsory

**Học phần tiên quyết (Prerequisites)**:

- Probability and Statistics (BAS1226)

**Học phần trước (Previous courses)**:

**Học phần song hành (Parallel courses)**:

**Các yêu cầu đối với học phần (Course requirements)**:

- Lecture room: Projector, microphone and speaker, black board or white board.

- Laboratory:

**Giờ tín chỉ đối với các hoạt động (Teaching and Learning hours)**:

- Lý thuyết (Lectures): 32h

- Bài tập (Exercises): 8h

- Bài tập lớn (Projects): 4h

- Thực hành (Labs): 0h

- Tự học (Individual reading): 1h

**Địa chỉ Khoa/Bộ môn phụ trách học phần (Address of the Faculty/Department in charge of the course):**

- Address: Faculty of Information Technology 1 - Posts and Telecommunications Institute of Technology, Km10, Nguyen Trai Street, Ha Dong District, Hanoi.

- Phone number: (024) 33510432

**2. Mục tiêu học phần (Objectives)**

**Về kiến thức (Knowledge):**

The aim of this course is to provide learners with basic knowledge about data science and methods for data processing and analytics, including:

* data processing (data gathering, data cleaning, data scaling and normalization, data reduction and transformation, SQL)
* data visualization
* methods for data analytics (prediction/classification/clustering/recommendation)

**Kỹ năng (Skills):**

The aim of this course is to equip learners with skills in:

* applying the learned knowledge to process and analyze some types of data
* evaluation of data analytic models.

**Thái độ, Chuyên cần (Attitude):**

Students are required to attend the classes and complete exercises and assignments.

**3. Tóm tắt nội dung học phần (Description)**

This course introduces learners to basic knowledge about data science (data, data types, hypothesis, and statistical hypothesis testing) and methods for data processing and analytics, including data preparation (data gathering, data cleaning, data scaling and normalization, data reduction and transformation), data visualization, and data prediction/classification/clustering/recommendation models. The course also teaches students how to evaluate data analytic models.

**4. Nội dung chi tiết học phần (Outlines)**

**Chapter 1: Background**

1.1. Introduction

1.2. Linear algebra review

1.2.1. Vectors

1.2.2. Matrices

1.2.3. Determinants

1.3. Probability review

1.3.1. Sample Spaces and Events

1.3.2. Formulas for calculating probability

1.3.2. Bayes’ theorem

1.3.3. Random variables

1.3.4. Continuous distributions

1.3.5. The Normal distribution

1.4. Statistical review

1.4.1. Confidence Interval

1.4.2. Hypothesis Testing

1.4.3. Examples

**Chapter 2: Data Preparation**

2.1. Data collection

2.1.1 Data Collection Methods

2.1.2 Data Collection Tool

2.1.3 Web Data Collection

2.2. Data type portability

2.3. Data cleaning

2.3.1. Handling missing entries

2.3.2. Handling incorrect and inconsistent entries

2.4. Scaling and normalization

2.5. Data reduction and transformation

2.5.1. Sampling

2.5.2. Feature subset selection

2.5.3. Dimensionality reduction

2.5. Projects

**Chapter 3: Descriptive statistics**

3.1 Frequency distribution

3.2 Measures of central tendency

3.3 Measures of variability

3.4 Data Visualization

3.4.1. Simple line plots

3.4.2. Simple scatter plots

3.4.3. Visualizing errors

3.4.4. Density and contour plots

3.4.5. Histograms, bins, and density

3.4.6. Multiple subplots

3.4.7. Text and annotation

3.4.8. Three-dimensional plotting

3.4.9. Geographic data

**Chapter 4: Machine Learning**

4.1. Basic concepts

4.1.1. Learning and inference

4.1.2. Model evaluation

4.1.3. Overfitting and underfitting

4.1.4. Bias and variance

4.2. Feature engineering

4.2.1. Categorical features

4.2.2. Text features

4.2.3. Image features

4.2.4. Derived features

4.3 Main types of machine learning

4.3.1. Supervised learning

4.3.2. Unsupervised learning

4.3.3. Semi-supervised learning

4.3.4. Ensemble learning

4.4. Projects

**Chapter 5: Databases and SQL**

5.1 Introduction to Databases and SQL

5.2 Fundamental SQL

5.1.1. CREATE TABLE and INSERT

5.1.2. UPDATE

5.1.3. DELETE

5.1.4. SELECT

5.1.5. GROUP BY

5.1.6. ORDER BY

5.1.7. JOIN

5.3. Advanced SQL

5.2.1. Subqueries

5.2.2. Query Optimization

5.2.3. NoSQL

**Chapter 6: Recommender Systems**

6.1. Introduction to recommender systems

6.2. Content-based filtering

6.3. Collaborative filtering

6.3.1. User-based collaborative filtering

6.3.2. Item-based collaborative filtering

6.3.3. Matrix factorization

6.4. Hybrid recommender systems

6.5. Context-aware recommender systems

6.6. Session-based recommendation

**5. Học liệu (Textbooks)**

**5.1. Học liệu bắt buộc (Required Textbooks)**

[1]. *Joel Grus. Data Science from Scratch: First Principles with Python. O’Reilly Media, 2nd edition, 2019.*

[2]. Jake VanderPlas. *Python Data Science Handbook: Essential Tools for Working with Data*. *O’Reilly Media, 2017.*

**5.2. Học liệu tham khảo (Optional Textbooks)**

[3]. Charu C. Aggarwal. *Data Mining: The Textbook*. Springer International Publishing Switzerland, 2015.

**6. Phương pháp, hình thức kiểm tra – đánh giá kết quả học tập học phần (Grading Policy)**

| **Grading method** | **Percentage** | **Group/Individual** |
| --- | --- | --- |
| - Attendance | 10% | Individual |
| - Exercises | 10% | Individual |
| - Mid-term projects/exam | 20% | Group or individual |
| - Final examination | 60% | Individual |

| **Trưởng Bộ môn**  **(Head of Department)**  **Ngô Xuân Bách** |  | **Giảng viên biên soạn**  **(Lecturer)**  **Ngô Xuân Bách** |
| --- | --- | --- |