**HITEC UNIVERSITY**

Department of Computer Science

**BS Computer Science Program**

**CS-429: Introduction to Data Science 3 (2+1)**

|  |  |
| --- | --- |
| **Course Code** | **CS –429** |
| **Course Title** | Introduction to Data Science |
| **Credit Hours** | 3 (2+1) |
| **Contact Hours** | 5 (2+3) |
| **Semester** | BSCS 7th Semester — Fall 2024 |
| **Prerequisite** | Some familiarity with Probability and Statistics, Databases |
| **Course Instructors** | Mr. Mubashir Iqbal |
| **Designation** | Lecturer |
| **Email** | [mubashir.iqbal@hitecuni.edu.pk](mailto:mubashir.iqbal@hitecuni.edu.pk) |

**Course Description**

Data Science is the study of the generalizable extraction of knowledge from data. Being a data scientist requires an integrated skill set spanning mathematics, statistics, machine learning, databases and other branches of computer science along with a good understanding of the craft of problem formulation to engineer effective solutions. This course will introduce students to this rapidly growing field and equip them with some of its basic principles and tools as well as its general mindset. Students will learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration, exploratory data analysis, predictive modeling, descriptive modeling, data product creation, evaluation, and effective communication. The focus in the treatment of these topics will be on breadth, rather than depth, and emphasis will be placed on integration and synthesis of concepts and their application to solving problems. To make the learning contextual, real datasets from a variety of disciplines will be used.

**Course Objective:**

Data Science is an emerging field and involves the study of the generalizable extraction of knowledge from data. This is a merger of various disciplines including mathematics, statistics, machine learning, databases and other branches of computer science along with a good understanding of the craft of problem formulation to engineer effective solutions. It has overlapping boundaries with already established areas like Data Mining and Machine Learning. The major concepts that will be covered during course are probability, statistical inference, visualization, exploratory data analysis (EDA), linear and logistic regression, model evaluation and various machine learning algorithms such as random forests, k-means clustering, and naïve Bayes. On the programming aspect, Python will be used as an implementation tool.

**Pre-Requisite**

Students are expected to have

* Some familiarity with Probability and Statistics, Databases
* Any programming language

### Teaching-Learning Methodology

Lectures will be delivered using whiteboard (most of the time), sometimes we may use slides/multimedia. Class participation will be encouraged. Ask questions immediately when it comes to your mind. Any difficulty in the lecture must be pointed out in the coming lecture (immediately after the lecture). It is strongly advised to concentrate and participate in the discussions during class hours, because you may not find the topics in a book, the same way as discussed in the class. Be on time, as there may be a quiz in the start of the class.

**Course Outline:**

Introduction: What is Data Science? Big Data and Data Science hype, Datafication, Current landscape of perspectives, Skill sets needed; Statistical Inference: Populations and samples, Statistical modeling, probability distributions, fitting a model, Intro to Python; Exploratory Data Analysis and the Data Science Process; Basic Machine Learning Algorithms: Linear Regression, k-Nearest Neighbors (k-NN), k-means, Naive Bayes; Feature Generation and Feature Selection; Dimensionality Reduction: Singular Value Decomposition, Principal Component Analysis; Mining Social-Network Graphs: Social networks as graphs, Clustering of graphs, Direct discovery of communities in graphs, Partitioning of graphs, Neighborhood properties in graphs; Data Visualization: Basic principles, ideas and tools for data visualization; Data Science and Ethical Issues: Discussions on privacy, security, ethics, Next-generation data scientists

**TEXTBOOK/ REFERENCES MATERIAL:**

1. Doing Data Science, Straight Talk from the Frontline, Cathy O'Neil and Rachel Schutt, O'Reilly. 2014 (**Main**).
2. Steven S. Skiena, *“*The Data Science Design Manual”, Springer, 2017 (Main)
3. Foundations of data science, Blum, A., Hopcroft, J., & Kannan, R., Vorabversion eines Lehrbuchs, 2016.
4. An Introduction to Data Science, Jeffrey S. Saltz, Jeffrey M. Stanton, SAGE Publications, 2017.
5. Python for everybody: Exploring data using Python 3, Severance, C.R., CreateSpace Independent Pub Platform. 2016.
6. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, EMC Education Services, John Wiley & Sons, 2015.

**Program Learning Outcomes:**

SO 1: Academic Education:

SO 2: Knowledge for solving computing problems:

SO 3: Problem Analysis:

SO 4: Design/Development of Solutions:

**Course Learning Outcomes:**

**CO 1:** **Describe** basic concepts of exploratory data analysis, machine learning, visualization techniques, and their applications [C2, Understand].

**CO 2:** **Apply** various predictive modeling and data analysis techniques to solve real-life problems [C3, Apply]

**CO 3: Explore and Analyze** various data science algorithms for designing optimal solutions for real-life case studies [C4, Analyze]

**CO 4: Compare and Evaluate** supervised/unsupervised,statistical, visualization modelson real-life scenarios [C5, Evaluate]

**Table 1:**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Course** | **COs** | **SO 1** | **SO 2** | **SO 3** | **SO 4** | **SO 5** | **SO 6** | **SO 7** | **SO 8** | **SO 9** | **Learning Levels** |
| Essentials of Artificial Intelligence | CO 1 | ✔ |  |  |  |  |  |  |  |  | C2 |
| CO 2 |  | ✔ |  |  |  |  |  |  |  | C3 |
| CO 3 |  |  | ✔ |  |  |  |  |  |  | C4 |
| CO 4 |  |  |  | ✔ |  |  |  |  |  | C5 |

**Mapping Guide:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Program Learning Outcomes** | **Learning Levels** | | |
| **Cognitive Domain** | **Affective Domain** | **Psychomotor Domain** |
| SO1: Engineering Knowledge  SO2: Problem Analysis  SO3: Design/Development of Solutions  SO4: Investigation  SO5: Modern Tool Usage  SO6: The Engineer and Society  SO7: Environment and Sustainability  SO8: Ethics  SO9: Individual and Team Work  SO10: Communication  SO11: Project Management  SO12: Lifelong Learning | C1- Knowledge  C2- Comprehension  C3- Application  C4- Analysis  C5- Synthesis  C6- Evaluation | A1- Receiving  A2- Responding  A3- Valuing  A4- Organizing or conceptualizing values  A5- Characterizing or internalizing values | P1- Observe  P2- Model  P3- Recognize Standards  P4- Correct P5- Apply P6- Coach |

**Week-wise breakdown:**

|  |  |  |
| --- | --- | --- |
| **Week No** | **Lecture Breakdown** | **Activity** |
| 1 | **Introduction (Chapter 1 [from 1, 2])**  What is data science, Big data, and data science hype, properties of data, data science functionalities, what is a data scientist in academia and industry | — |
| 2 | **Statistical Analysis (Chapter 2 [1], 5[2])**  Statistical Distributions (Binomial Distribution, Normal Distribution) | — |
| 3 | **Statistical Analysis (Chapter 2 [1], 5[2])**  Statistical Distributions (Poisson Distribution), populations and samples, populations and samples of big data | — |
| 4 | **Statistical Analysis (Chapter 2 [1], 5[2])**  Exploratory Data Analysis, Data Science process, Correlation Analysis, Pearson Correlation Coefficient, Spearman Rank Correlation Coefficient | — |
| 5 | **ML Algorithms and Spam Filters (Chapter 3, 4 [1])**  What is ML, three basic algorithms, linear regression (LR), Fitting model, Example with complete steps, Evaluation metrics, k-nearest neighbors (k-NN), | — |
| 6 | **ML Algorithms and Spam Filters (Chapter 3, 4 [1])**  Similarity Metrics, Pick an Evaluation Metric, confusion matrix, evaluation metrics, Examples, **Clustering,** k-means, k-means steps/algorithm, Examples,Issues | — |
| 7 | **ML Algorithms and Spam Filters (Chapter 3, 4 [1])**  Spam filters, clear signs of spam, suggestions, why LR won't work for spam filtering, Example, A spam filter for individual word, | — |
| 8 | **ML Algorithms and Spam Filters (Chapter 3, 4 [1])**  Spam filter for combined words, Laplace smoothing, comparing Naïve Bayes and k-NN | — |
| **Mid Term** | | |
| 9 | **Recommendation Engines (Chapter 8[1], 10[2])**  Feature Generation and Feature Selection; Dimensionality Reduction: Singular Value Decomposition | — |
| 10 | **Recommendation Engines (Chapter 8[1], 10[2])**  Dimensionality Reduction: Principal Component Analysis | — |
| 11 | **Recommendation Engines (Chapter 8[1], 10[2])**  Measuring distance, graphs, networks, and distances, PageRank Algorithm | — |
| 12 | **Social Networks**  Social-Network Graphs: Social networks as graphs, Clustering of graphs | — |
| 13 | **Social Networks**  Direct discovery of communities in graphs, Partitioning of graphs, Neighborhood properties in graphs | — |
| 14 | **Visualizing Data (Chapter 6[2])**  Why Visualizing Data?, EDA (Confronting a New Data Set, Anscombe’s Quartet, Visualization Tools), developing a visualization aesthetic (Tufte’s Visualization Aesthetic), | — |
| 15 | **Visualizing Data (Chapter 6[2])**  Chart types (Tabular data, Improvement, Line and dot plots, Line plots, scatter plots, heat maps, bubble charts, Bar plots, and pie charts, Histograms, data maps) | — |
| 16 | **Data Science and Ethical Issues**  Discussions on privacy, security, ethics, Next-generation data scientists | — |
| **Final Term** | | |