

Institute of Business Administration

CSE 306: Foundations to Data Science

(Tentative Course Outline and Syllabus)



School of Mathematics and Computer Science

Spring 2024

CSE 306 BSEM : Foundations to Data Science

“Learn continually-there's always “one more thing” to learn” – Steve Jobs

1. Logistics

Course:	CSE 306 BSEM: Foundations to Data Science
Class Number:	71274
Class timings:	2:30PM – 3:45 PM (Monday & Wednesday)
Classroom:	Auditorium (Main Campus)
Instructor(s):	Usman Ali
Email:	uali@iba.edu.pk
Phone:	0336 2090484
Office:	BDA Lab (Main Campus)
Office hours:	12:30 PM – 02:30 PM (Monday & Wednesday)
LMS:	https://lms.iba.edu.pk/portal/site/b6ac26f4-3d61-4b6a-bac7-8fa5510648f4
TA:	Naeemullah Baig
TA email:	n.baig.25284@khi.iba.edu.pk

2. Course Description/Objectives

Foundations to Data Science is a 3-credit hour non-specialization core course from Computer Science domain for BSEM discipline. The objective of this course is to teach the fundamentals of data science, particularly cleaning, data pre-processing, and data analytics with respect to statistical data comprehension, predictive modeling (supervised learning) and data profiling (unsupervised learning). The specific goal is to impart the skill of cleaning the data from all types of noise and errors, understanding the data through statistical analysis and then applying mathematical /machine learning modeling on data to extract actionable insights in a smart and informed way through different data science techniques. For this, the course is designed to be a hands-on course (mainly in Python) . The hands-on activities will be conducted with core representative case studies/datasets related to the local and international industries and markets. This course would aid:

- To excite students about the potential that resides in data and the value that data analytics can add to business processes.
- To impart skills related to data cleaning/wrangling, data transformation/preprocessing and data comprehension through statistical analysis.
- To impart skills related to analytical (mathematical / machine learning) data modeling.

3. Program Learning Outcomes/Graduate Attributes

PLO-X. TBD (To be decided)

PLO-Y. TBD

PLO-Z. TBD

4. Course Learning Outcomes

Course Learning Outcome	
CLO	Description
CLO-1	Understand the data science domain , its role and impact in solving business problems
CLO-2	Get thorough understanding of data quality problems and develop skills in data cleaning , preprocessing and feature selection to make the data fit for any downstream analytical task through hands-on activities and case studies
CLO-3	Theoretical / Mathematical knowledge about a standard subset of supervised , unsupervised and ARM related machine learning models [and /or timeseries models] along with implementation.

4.1 CLOs to PLOs Mapping

	PLO-X	PLO-Y	PLO-Z
CLO-1			
CLO-2			
CLO-3			

5. Format and Procedures

- The LMS site will be used to share the syllabus, give out assignments, and to share other course resources.
- The University's standard policies on attendance, inclusivity, office hours, and academic integrity apply in this course. These are described in later sections below.

6. Course Requirements

- Class participation policy: Background reading for next session and active participation in class discussions.
- **Textbooks:**
 1. **Data Mining: Concepts and Techniques**, Third Edition , Jiawei Han, Micheline Kamber, and Jian Pei, (The Morgan Kaufmann Series in Data Management Systems)

2. **Data Science Concepts and Practice** , Second edition , Vijay Kotu and Bala Deshpande

- **Reference Books**

1. **Building Machine Learning Systems with Python** , Second Edition , Luis Pedro Coelho Willi Richert
2. **The Elements of Statistical Learning ,Data Mining, Inference, and Prediction**, Second Edition , Trevor Hastie, Robert Tibshirani , Jerome Friedman

7. Grading Procedures

Grades will be computed as follows (Tentative)

Homework	15%
Quizzes	15%
Project	15%
CP	05%
Midterm	20%
Final	30%

- **Instructions regarding homework , quizzes and exams.**

- Quizzes are mostly unannounced.
- Homework assignments will be posted and accepted through LMS only .
- **10 % marks deduction on late submission on LMS.** No submission will be accepted through other means (like drop box) if LMS late submission expires.
- **No email submissions** will be accepted.

- **Instructions regarding project.**

- Students are expected to work on a real-world dataset in a group of two or three. A student can take the project individually as well but working in a group is preferred. No more than three members will be allowed.
- Project requirements include business problem analysis , data analysis (after cleaning and pre-processing) and solution design / modelling.
- Deliverables include a project report , code /application and presentation.

8. Makeup Policy

There will be no retake of quizzes and homework assignments. In case a student is unable to appear in mid-term or final exam due to unavoidable circumstances , IBA policy applies. Any such requests should be made to the program office keeping the examinations and instructor in cc with required documents of proof.

9. Attendance Policy

IBA attendance policy applies.

10. Academic Integrity

Each student in this course is expected to abide by the IBA Code of Conduct. Scholastic dishonesty shall be considered a serious violation of these rules and regulations and is subject to strict disciplinary action as prescribed by IBA regulations and policies. Scholastic dishonesty includes, but is not limited to, cheating on exams, plagiarism on assignments, and collusion. Kindly refer to <https://examination.iba.edu.pk/CheatingPlagiarism.php> for more details.

- **PLAGIARISM:** Plagiarism is the act of taking the work created by another person or entity and presenting it as one's own for the purpose of personal gain or of obtaining academic credit. Plagiarism includes the submission of or incorporation of the work of others without acknowledging its provenance or giving due credit according to established academic practices. This includes the submission of material that has been appropriated, bought, received as a gift, downloaded, or obtained by any other means. Students must not, unless they have been granted permission from all faculty members concerned, submit the same assignment or project for academic credit for different courses.
- **CHEATING:** The term cheating shall refer to the use of or obtaining of unauthorized information to obtain personal benefit or academic credit.
- **COLLUSION:** Collusion is the act of providing unauthorized assistance to one or more person or of not taking the appropriate precautions against doing so.

Any student violating academic integrity a second time in this course will receive a failing grade for the course, and additional disciplinary sanctions may be administered.

- **SHARING CREDENTIALS:** It has been observed that some students share their credentials (login ids and passwords) of LMS, portal, email, etc., with other students. These credentials are private and confidential and not to be shared with anyone. Any violation will be considered as aiding in plagiarism/collusion/cheating and appropriate action might be taken against such students.

11. Office Hours

10:00 AM-11:30 AM , Monday & Wednesday (BDA Lab , Ground Floor, Tabba Block).

12. Weekly Breakdown of Classes

Week	Topic	Mapping to Course Learning Outcome
1	Introduction to Data Science Data Science , Machine Learning and Artificial Intelligence Relationship between Data Science , Machine Learning and AI Data Science Process Machine Learning , Types of Machine Learning Role and Impact of Data Science in Data Driven Decision Making Anaconda Navigator , Jupyter Notebook Installations and Environment Setup. KNIME/RapidMiner Introduction	CLO1

2	Handling Data Quality Problems What is data quality ? Handling Missing Values Handling Incorrect Data Types	CLO2
3	Data Preprocessing Attribute Types Data Discretization , Normalization, Feature Encoding , Feature Engineering , Train-Test Split	CLO2
4 & 5	Feature Selection & Statistical Analysis Feature Selection Techniques Filter Methods (Pearson's Correlation , Covariance, Chi-Square Test , ANOVA) Wrapper Methods & Embedded Methods	CLO2
6	Linear Regression – Predicting Numerical Response Variable Simple Linear Regression , Multiple Linear Regression , Assumptions for Linear Regression Model, Evaluating Regression Model, F-Value and R^2 , Beta Coefficients, Evaluation Metrics for Linear Regression, Applications of Linear Regression	CLO3
7	Logistic Regression – Predicting Categorical Response Variable Regression vs Classification , Hypothesis Representation, Activation Function, Cost Function, Gradient Descent and Maximum Likelihood Estimation, Scikit Learn implementation of Logistic Regression, Types of Logistic Regression, Use Cases	CLO3
8	Decision Tree Classifier Decision Tree Induction , Attribute selection Measures(Gini , entropy, misclassification error) , Tree Pruning , Scikit Learn Implementation , Use Cases	CLO3
9	Model Evaluation & Selection Metrics for evaluating classifier's performance (Accuracy, Precision, Recall, F-Measure) , Holdout Method , Random Subsampling, Cross Validation , Bootstrap , Generalization , Overfitting	CLO3
10	Ensemble Methods Bagging , Boosting, Random Forest, Adaboost, XGBoost	CLO3
11 & 12	Naïve Bayes , KNN , SVM , Artificial Neural Networks	CLO3
13 & 14	Unsupervised Learning - Clustering Cluster Analysis , Partitioning Methods , Hierarchical Methods, Density Based Methods, Evaluation of Clustering	CLO3
	Project Presentations	