CODE	COURSE NAME	CATEGORY	L	T	P	CREDIT
ITT306	DATA SCIENCE	PCC	3	1	0	4

Preamble: This course is designed to provide learners with working knowledge of the theoretical background of various aspects of Data Science and enable them to incorporate and apply the principles of statistics and machine learning to solve real-world problems for large-scale data analysis.

Prerequisites:

- MAT 101 Linear Algebra and Calculus
- MAT 208 Probability and Statistics and Advanced Graph Theory
- ITT 205 Problem Solving Using Python
- ITT 201 Data Structures
- ITT 206 Database Management Systems

Course Outcomes: After the completion of the course the student will be able to

CO No.	Course Outcome(CO)	Bloom's Category Level
CO 1	Explain the fundamental concepts and various aspects of data science	Level 2: Understand
CO 2	Choose data validationtechniques suitable for statistical analysis andpresent results using data visualization techniques.	Level 2: Understand
CO 3	Identify different statistical learning algorithm for solving a problem	Level 3: Apply
CO 4	Use statistical analysis to characterize and interpret data sets	Level 3: Apply
CO 5	Compare the pros/cons of various models and algorithms used for data analysis and data mining	Level 2: Understand
CO 6	Develop the ability to perform basic data analysis in Python and understand the fundamentals of deep learning.	Level 3: Apply

Mapping of course outcomes with program outcomes

POs / COs	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12
CO 1	1	1	1	1	2	-	-	-	-	-	-	-
CO 2	3	2	1	1	3	-	-	-	-	2	-	-
CO 3	3	2	1	1	3	-	-	-	1	2	-	-
CO 4	3	3	2	1	3	-	-	-	1	2	-	-

CO 5	2	3	1	1	3	-	-	-	1	2	-	-
CO 6	3	2	1	1	3	-	-	-	1	2	-	-

3/2/1: high/medium/low

Assessment Pattern

Bloom's Category	Continuous Assessment Tests		End Semester Examination Marks
	Test1 (Marks)	Test2 (Marks)	
Remember	10	10	20
Understand	25	25	50
Apply	15	15	30
Analyse			
Evaluate			
Create			

Marks distribution

Total Marks	CIE	ESE	ESE Duration
150	50	100	3 hours

Continuous Internal Evaluation Pattern

Attendance : 10 marks
Continuous Assessment Test (2 numbers) : 25 marks
Assignment/Quiz/Course project : 15 marks

End Semester Examination Pattern

There will be two parts; Part A and Part B. Part A contains 10 questions with 2 questions from each module, having 3 marks for each question. Students should answer all questions. Part B contains 2 questions from each module of which student should answer anyone. Each question can have maximum 2 sub-divisions and carry 14 marks.

Course Level Assessment Questions

Course Outcome 1 (CO 1):

1. What is data science? What are the different models for data science?
2. Explain data science process with a neat diagram.
3. Explain different types of Data Sets in Data Science

Course Outcome 2 (CO 2):

1. List any four two tools for data visualisation?

2. What is data visualization and what are the different techniques used for visualizing data?
3. Discuss methods of evaluating models in data science?

Course Outcome 3(CO 3):

1. Explain random forest ensemble method with an example.
2. What is data cleaning? What are the different operations in data cleaning?
3. Is regression a supervised learning technique? Justify your answer. Compare it with classification giving examples.
4. What are ensemble methods? Explain the bagging technique
5. Discuss Linear discriminant analysis.
6. What is decision tree? Explain the working of decision tree with information gain algorithm.

Course Outcome 4 (CO 4):

1. Differentiate between supervised and unsupervised learning techniques.
2. Classify different types of clustering. What are the practical issues in clustering?
3. Summarise different kernel tricks in SVM.
4. Illustrate with examples different Resampling methods.
5. Suppose that our task is to cluster data points into two clusters. Let the data points are {2, 4, 10, 12, 3, 20, 30, 11, 25}. Let 2 and 4 are initial cluster centroids. Apply Two rounds of k-means algorithm and find a set of clusters. Use Euclidean distance as the measure.

Course Outcome 5 (CO 5):

1. Compare Apriori and FP Growth algorithm. What are the advantages of FP Growth over Apriori algorithm?
2. How will you relate constraint-based mining with frequent pattern mining?

3. A database has five transactions. Let min_sup=60% and min_conf=80%. With the following transaction, list all the strong association rules.

T100 {M, O, N K, E, Y}

T200 {D, O, N, K, E, Y}

T300 {M, A, K, E}

T400 {M, U, C, K, Y}

T500 {C, O, O, K, I, E}

Course Outcome 6 (CO 6):

1. Write an example of multiplying three dimensional matrices in NumPy.
2. Identify the essential libraries in Python.
3. Is Jupyter notebook IDE?How can you relate IPython and Jupyter?
4. What are the ways to store text data in pandas?

Course Code: ITT306

Course Name: Data Science

Max.Marks:100 **Duration: 3 Hours**

Part A

Answer all questions. Each question carries 3 marks. (10 * 3 = 30 Marks)

1. What is data science? What are the different models for data science?
2. What is data visualization and what are the different techniques used for visualizing data?
3. Is regression a supervised learning technique? Justify your answer. Compare it with classification giving examples.
4. Explain random forest ensemble method with an example.
5. Explain different types of clustering. What are the practical issues in clustering?
6. What is Support Vector Machine? How classification is done using SVM?
7. Explain the concept of constraint-based mining.
8. Compare Apriori and FP Growth algorithm. What are the advantages of FP Growth over Apriori algorithm?
9. Briefly explain the essential libraries in Python.
10. What makes deep learning deep? What are the different deep learning techniques?

Part B

Each question set carries 14 marks (5 * 14 = 70 Marks)

11. Explain data science process with a neat diagram.
12. Describe data science classification with a neat diagram.

OR

13. What is data cleaning? What are the different operations in data cleaning?
14. Explain different types of Data Sets in Data Science.
15. What is decision tree? Explain the working of decision tree with information gain algorithm.
16. What are ensemble methods? Explain the bagging technique.

OR

17. Differentiate supervised and unsupervised learning techniques with examples.
18. Discuss Linear discriminant analysis.

19. Explain different types of Resampling methods.
20. What is SVM? Explain Different kernel tricks in SVM.

OR

21. Write a short note on Maximal Margin Hyperplanes. (MMH).
22. Suppose that our task is to cluster data points into two clusters. Let the data points are {2, 4, 10, 12, 3, 20, 30, 11, 25}. Let 2 and 4 are initial cluster centroids. Apply Two rounds of k-means algorithm and find a set of clusters. Use Euclidean distance as the measure.

23. Consider the transaction database given below. Set minimum support count as 2 and minimum confidence threshold as 70%.
    a) Find the frequent item-set using FP Growth Algorithm.
    b) Generate strong association rules.

Transaction ID	List of Item_Ids
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

24. Explain Multi-level and multi-dimensional pattern mining.

OR

25. What is data mining? Explain the process of Knowledge discovery from database.
26. A database has five transactions. Let min_sup=60% and min_conf=80%. With the following transaction, list all the strong association rules.

T100 {M, O, N K, E, Y}

T200 {D, O, N, K, E, Y}

T300 {M, A, K, E}

T400 {M, U, C, K, Y}

T500 {C, O, O, K, I, E}

27. What are the basic universal functions in Numpy?
28. What are the applications of deep learning?

OR

29. Write an example of multiplying three dimensional matrices in NumPy.

30. What are the ways to store text data in pandas?

Syllabus

Module 1: Foundations Data Science, process, and tools (9 Hours)
Introduction to data science, properties of data, asking interesting questions, classification of data science, data science process, collecting, cleaning and visualizing data, languages, and models for data science
Module 2: Statistical machine learning: introduction, regression, and classification, decision tress, random forests (11 Hours)
Introduction to statistical machine learning, parametric and non-parametric methods, supervised vs. unsupervised learning, regression and classification, linear discriminant analysis, decision trees, random forests, and bagging
Module 3: Unsupervised learning, support vector machines and resampling (9 Hours)
Principal Component Analysis, clustering algorithms, practical issues in clustering, support vector classifiers and support vector machines, resampling methods: cross-validation and bootstrapping
Module 4: Data mining, pattern mining and association rule mining (9 Hours)
Data and pattern mining, types, issues, mining frequent patterns and associations, apriori and FP growth algorithms, multi-level association mining, constraint-based mining, pruning pattern space and data space
Module 5: Python for Data Analysis, Deep learning (7Hours)
Using Python for data analysis, essential python libraries, IPython, Jupyter notebook, NumPy basics, working with pandas, deep learning methods.

Textbooks

1. Kotu, V., & Deshpande, B. (2019). Data science: Concepts and practice., Morgan Kaufmann.
2. Skiena, S. S. (2017). The data science design manual., Springer.
3. James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). An Introduction to Statistical Learning: with Applications in R., Springer.
4. Han, J., Kamber, M. & Pei, J. (2012). Data mining concepts and techniques, Morgan Kaufmann.

5. McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. Beijing: O'Reilly.

Reference Books

1. Montgomery, D. C., Runger, G. C. (2017). Applied Statistics and Probability for Engineers. John Wiley and Sons.

2. Provost, F., Fawcett, T. (2013). Data Science for Business. Beijing: O'Reilly

3. Igual, L., Seguí, S. (2017). Introduction to Data Science - A Python Approach to Concepts, Techniques and Applications. Springer.

Course Contents and Lecture Schedule

No	Topic	No. of Lectures
1	**Foundations Data Science, process, and tools**	**9 Hours**
1.1	What is Data Science, relation with AI and machine learning (1.1, 1.2 of Kotu, V., & Deshpande, B. (2019). *Data science: Concepts and practice*., Morgan Kaufmann.)	1 Hour
1.2	Case for Data Science, Data science classification (1.3, 1.4 of Kotu, V., & Deshpande, B. (2019). *Data science: Concepts and practice*., Morgan Kaufmann.)	1 Hour
1.3	Properties of data, asking interesting questions (1.1, 1.2 and 1.3 of Skiena, S. S. (2017). *The data science design manual*., Springer.)	1 Hour
1.4	Data Science process: preparation, modelling, and application (2.1, 2.2, 2.3 and 2.4 of Kotu, V., & Deshpande, B. (2019). *Data science: Concepts and practice*., Morgan Kaufmann.)	2Hours
1.5	Collecting and cleaning data (3.2 and 3.3 of Skiena, S. S. (2017). *The data science design manual*., Springer.)	1 Hour
1.6	Visualizing data (6.1, 6.2 and 6.3 of Skiena, S. S. (2017). *The data science design manual*., Springer.)	1Hour
1.7	Languages and models for Data Science, evaluating models (3.1, 7.2, 7.3 and 7.4 of Skiena, S. S. (2017). *The data science design manual*., Springer.)	2 Hours
2	**Statistical machine learning: introduction, regression, and classification, decision tress, random forests** **(Reference Textbook for all topics:** James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). *An Introduction to Statistical Learning: with Applications in R*., Springer.)	**11 Hours**
2.1	What is statistical learning, parametric and non-parametric methods (2.1)	1 Hour
2.2	Supervised vs. unsupervised learning, Classification vs. regression (2.1)	1 Hour
2.3	Simple linear regression, assessing model accuracy (3.1)	1Hour
2.4	Multiple linear regression, some important concerns (3.2)	1 Hour

2.5	Extensions of the linear model (3.3.2)	1 Hour
2.6	Classification (4.1)	1 Hour
2.7	Logistic regression: model, estimating coefficients, predicting (4.3.1, 4.3.2, 4.3.3)	2 Hours
2.8	Linear discriminant analysis, using Bayes' theorem for classification, case when p=1 (4.4.1, 4.4.2, 4.4.3)	1 Hour
2.9	Decision tress, regression and classification trees, tress vs. linear models, advantages, and disadvantages (8.1)	1 Hour
2.10	Bagging, random forests (8.2.1, 8.2.2)	1 Hour
3	**Unsupervised learning, support vector machines and resampling** **(Reference Textbook for all topics:** James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). *An Introduction to Statistical Learning: with Applications in R*., Springer.)	**9Hours**
3.1	Challenge of unsupervised learning, principal component analysis (10.1, 10.2.1)	1 Hour
3.2	Clustering techniques: k-means, hierarchical (10.3.1, 10.3.2)	1Hour
3.3	Practical issues in clustering (10.3.3)	1 Hour
3.4	Overview of the support vector classifier, hyperplane, maximal margin classifier (9.1.1, 9.1.2, 9.1.3)	2 Hours
3.5	Support vector classifiers: overview and details (9.2.1, 9.2.2)	1 Hour
3.6	Support vector machines: Classification with non-linear decision boundaries (9.3.1, 9.3.2)	1 Hour
3.7	Resampling: cross-validation and bootstrapping (5.1 and 5.2)	2 Hours
4	**Data mining, pattern mining and association rule mining(Reference Textbook for all topics:** Han, J., Kamber, M. & Pei, J. (2012). Data mining concepts and techniques, Morgan Kaufmann.)	**9 Hours**
4.1	Data mining, kinds of data that can be mined (1.2, 1.3,)	1 Hour
4.2	Pattern mining: class description, mining frequent patterns and associations, classification, and regression for predictive analysis (1.4.1, 1.4.2, 1.4.3)	1 Hour
4.3	Cluster analysis, outlier analysis (1.4.4, 1.4.5), measures of pattern interestingness (1.4.6), Issues in data mining (1.7)	1 Hour
4.4	Mining frequent patterns: market basket analysis, frequent and closed item sets, association rules (6.1.1, 6.1.2)	1 Hour
4.4	Apriori algorithm, generating rules, improving efficiency (6.2.1, 6.2.2, 6.2.3)	1 Hour
4.5	FP growth algorithm (6.2.4)	1 Hour
4.6	Multi-level and multi-dimensional pattern mining (7.2.1, 7.2.2)	1 Hour
4.7	Mining quantitative association rules (7.2.3) mining rare and negative patterns (7.2.4)	1 Hour
4.8	Constraint-based mining: meta-rule guided mining (7.3.1) pattern generation, pruning pattern space and data space (7.3.2)	1Hour

5	**Python for Data Analysis, Deep learning**	**7Hours**
5.1	Why Python for data analysis? Essential libraries (1.2 and 1.3 of McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. Beijing: O'Reilly.)	1 Hour
5.2	IPython basics and Jupyter notebook (2.2 of McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. Beijing: O'Reilly.), demo of appropriate examples	1 Hour
5.3	NumPy basics, universal functions, array-oriented programming, mathematical and statistical methods, file I/O, linear algebra (4.1. 4.2, 4,3, 4.4, 4.5 of McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. Beijing: O'Reilly.)	2 Hours
5.4	Pandas basics, essential functionality, summarizing and computing descriptive statistics (5.1, 5.2, 5.3 of McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. Beijing: O'Reilly.)	2 Hours
5.5	Deep learning: networks and depth, back propagation, word and graph embeddings (11.6.1, 11.6.2, 11.6.3 of Skiena, S. S. (2017). *The data science design manual*., Springer.)	1Hour