Devi Ahilya University, Indore, India Institute of Engineering & Technology				III Year B.E. (Computer Engineering)(Full Time)			
Subject Code & Name	Instructions Hours per Week			Credits			
CER6E1 Data Warehousing & Mining Duration of Theory Paper:3 Hours	L	T	P	L	T	P	Total
	3	1	2	3	1	1	5

Learning Objectives:

1. Ability to understand the role of data mining in knowledge discovery process.
2. To familiarize with various data mining functionalities and how it can be applied to various real-world problems.
3. To learn about finding data characteristics and evaluating the outcome of data mining process.
4. To familiarize with various machine learning algorithms used in data mining.

Prerequisite(s):

The students are required to have some basic knowledge about Data structures and Databases.

COURSE CONTENTS

UNIT-I: **Introduction**

Data Mining: Overview, Type of data mined, Functionalities, Technology used, Target applications and challenges; Data Features: Attribute types, basic statistical description, measuring data similarity and dissimilarity.

UNIT-II: Data Preprocessing and Data Warehouse

Data Preprocessing: Overview, Data cleaning, Data integration, Data reduction, Data transformation and discretization; Data Warehouse: Basic concepts, Design and Usage, Implementation.

UNIT-III: Frequent Pattern and Association rule Mining

Basic concepts, Pattern Mining: Apriori algorithm, FP-growth Algorithm; Generating association rules, Pattern evaluation methods, Multi-level and multi-dimensional pattern mining.

UNIT-IV: Classification

Introduction, Decision tree induction, Bayes classification, Rule based classification, Advance classification methods: Bayesian belief networks, backpropogation etc.

UNIT-V: Cluster analysis and Advance Topics

Clustering: Introduction, Types of clustering; Partition-based clustering: K-Means, K-Medoids; Density based clustering: DBSCAN, Clustering evaluation.
Web Data Mining: Introduction, Types of Web mining, and Overview of web usage mining, web content mining and web structure mining.

Learning Outcomes:
Students who have completed this course should be able to:
1. Apply data mining functionalities on real world problems and datasets.
2. Have some knowledge about the couple of data mining tools and how they can be used for large data.
3. They would be able to find the characteristics of given data and may identify presence of outliers, if any.
4. The course would help them to pursue some advance course on data science and may help in subjects like Big Data, AI etc.

Assessment methods of all of the above: quizzes, exams, assignments, practicals

Books Recommended:
[1] Han, Kamber and Pi, Data Mining Concepts & Techniques, Morgan Kaufmann, 3rd Edition, India, 2012.
[2] Mohammed Zaki and Wagner Meira Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, 2014.
[3] Z. Markov, Daniel T. Larose Data Mining the Web, Jhon wiley & son, USA, 2007.
[4] Bing Liu, Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Springer, 2nd Edition, 2011.
[5] Sam Anahory and Dennis Murray, Data Warehousing in the Real World, Pearson Education Asia, 2000.
[6] W. H. Inmon, Building the Data Warehouse, 4th Ed Wiley India, 2005.

List of Experiments :
R-programming and WEKA tools need to be learn in order to complete the lab assignments.
- To know your data, an experiment to visualize summarized data using box-plot, scatter plot and quantile-quantile plot using R-Programming.
- To clean your data, an experiment to find outliers, remove noise and identify correlated data using R-Programming.

- To implement A-priori algorithm to find the frequent patterns in the given dataset. Students can use programming language of their choice to code.
- Use of WEKA tool to use various association mining algorithms on datasets and evaluate them based on pattern evaluation measures.
- Use of WEKA tool to use various classification algorithms on datasets and evaluate them on the basis of accuracy and other parameters.
- Use of WEKA tool to use various clustering algorithms on datasets and evaluate them based on cluster quality and other parameters.