COURSE JOURNAL ON DATA MINING

Course Description: Data Mining studies algorithms and computational paradigms that allow computers to find patterns and regularities in databases, perform prediction and forecasting, and generally improve their performance through interaction with data. It is currently regarded as the key element of a more general process called Knowledge Discovery that deals with extracting useful knowledge from raw data. The knowledge discovery process includes data selection, cleaning, coding, using different statistical and machine learning techniques, and visualization of the generated structures. The course will cover all these issues and will illustrate the whole process by examples. Special emphasis will be give to the Machine Learning methods as they provide the real knowledge discovery tools. Important related technologies, as data warehousing and on-line analytical processing (OLAP) were also discussed.

Specific Goals of the course:

Data Mining, Knowledge Discovery from Data (KDD), Related Technologies, Applications, Major issues, Data Objects and Attribute types, Basic Statistical descriptions of data, Measuring data similarity and dissimilarity, Genetic algorithms.

Data Preprocessing: Data Cleaning, Integration and transformation, reduction (Haar wavelet transformation and PCA), discretization and Concept hierarchy generation. Frequent Pattern Mining and Association Rule MiningApriori and FP growth algorithm Correlation Analysis, Linear Regression Analysis Classification: General Approach, Decision tree induction-ID3 algorithm, Attribute selection measures, tree pruning, Bayes Classification, Rule based classification, k-NN classifier, Support Vector Machine, Bayesian belief networks, Back propogation, Metrics for evaluating classifier performance. Clustering: Partition based clustering: k-Means, k-Medoids algorithms, Heirarchical- Agglomerative and Divisive algorithms, Density based methods- DBSCAN algorithm, Grid based methods- STING algorithm. Evaluation of clustering Outlier Detection: Types of outliers, challenges in outlier detection, Statistical ApproachesParametric and NonParametric methods, Proximity based approaches-Distance based, Grid based, Density based outlier detection. Clustering based approaches, classification based approaches. Mining contextual and collective outliers.

Need of Data Mining:

1. Data mining is the procedure of capturing large sets of data in order to identify the insights and visions of that data. Nowadays, the demand of data industry is rapidly growing which has also increased the demands for Data analysts and Data scientists. 2. Since with this technique, we analyze the data and then convert that data into meaningful information. This helps the business to take accurate and better decisions in an organization. 3. Data mining helps to develop smart market decision, run accurate campaigns, predictions are taken and many more. 4. With the help of Data mining, we can analyze customer behaviors and their insights. This leads to great success and data-driven business.

References:

1. I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann. 2000.
2. J. Han and M. Kamber. Data Mining: Concepts and Techniques, 2nd Ed. Morgan Kaufman. 2006.
3. M. H. Dunham. Data Mining: Introductory and Advanced Topics. Pearson Education. 2001.
4. D. Hand, H. Mannila and P. Smyth. Principles of Data Mining. Prentice-Hall. 2001.
5. Pang-Ning Tan, Michael Steinbach, Vipin Kumar. Introduction to Data Mining. Addison-Wesley Longman Publishing Co.
6. Introduction to Data Mining-Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Addison- Wesley Longman Publishing Co.
7. https://onlinecourses.nptel.ac.in- NPTEL (National Programme on Technology Enhanced Learning) is a joint initiative of the IITs and IISc
8. Data Mining: Introductory and Advanced Topics-M. H. Dunham, Pearson Education. 2001.