# Research Project
# CSF415 Data Mining

## Introduction

Data mining is the process of working with large data sets to identify patterns and establish relationships to solve problems through data analysis. As a part of the research project, you will be learning to design and implement the complete DM processing pipeline from dataset selection and understanding to preprocessing to analysis to drawing insights from real world datasets. Using the knowledge discovery process, you would be required to work with large amounts of data to find the hidden patterns and draw useful insights, while working towards the goal of familiarisation with essential techniques and key algorithms in DM.

## Different components involved in the project

1. **Dataset Selection**
   We will be working with Indian Government open data sets available at https://data.gov.in/ Few sample datasets are suggested for projects (https://docs.google.com/spreadsheets/d/1FGL6lsThp_xs9tpb_f1YIBLVBsKRyiOlcS63YIaW7jc/edit?usp=sharing), but you are free to further explore and choose your own data from this source based on your own interests. Multiple datasets can be merged together to form a larger dataset and problem definition.
   It is important that you get the dataset verified before the deadline in case you decide to choose the data yourself. For the sample datasets provided, no more than 3 teams are allowed to work on the same dataset. Please specify your dataset choices in the spreadsheet (https://docs.google.com/spreadsheets/d/1uUSPKJyHw9o_TqjRm4LoVYb7-9_fJeWtR1fOp1JgRKw/edit?usp=sharing) by the given deadline.
   In order to select the dataset, you would need to study and understand the data in details and identify the key questions or insights that you would like to draw from the dataset. This should be mentioned clearly in the data and final report. Please keep in mind that you will be working with real world data that can be messy and incomplete, so decide your dataset really carefully.
2. **Data Preprocessing and Visualization**
   You would need to perform exploratory data analysis alongwith suitable visualizations and identify/employ different preprocessing techniques suitable for the dataset.
   You should implement ***atleast 3 data preprocessing techniques*** studied in the class in addition to mandatory data cleaning and show the results.
3. **Data Analysis**
   Based on the insights that you wish to draw from the dataset, you would need to identify the key DM tasks like association analysis, clustering, classiication or outlier analysis that are applicable for the dataset.
   You should implement ***atleast 2 data analysis techniques and corresponding algorithms*** studied in the class and show the results.

## Weightage of Individual Components

Total Weightage 30%

1. Problem Defintion, Data Prepocessing and Visualization 10%
2. Data Analysis 10%

3. Demo and Report 5%
4. Extra Credits for extra effort and exceptional work 5%

**Important Deadlines**

1. ***Team selection*** Deadline January 20th
   Teams should consist of 3-4 students.
2. ***Dataset selection*** Deadline January 30th
   You would need to upload a **data report** with dataset details consisting of project title, problem definition, data description and development tools to be used (2 page pdf document named as DMData_TeamID_XX.pdf)
3. ***Midterm evaluation*** for component 1 (Problem Defintion, Data preprocessing and visualization) Deadline February 16th
   You would need to upload the **zipped data, code and initial version of report** on CMS submission link through one of the team member's account before deadline. Name of the file should be DMMidTerm_TeamID_XX.zip
4. ***Final evaluation*** for component 2 (Data analysis) Deadline April 12th
   You would need to upload the **zipped data, code and final version of report** on CMS submission link through one of the team member's account before deadline. Name of the file should be DMFinal_TeamID_XX.zip
   *NOTE- NO LATE SUBMISSIONS WILL BE ACCEPTED.*

**Few Important Things to Remember**

1. Teams once decided cannot be changed.
2. Dataset needs to be selected carefully and cannot be changed once decided.
3. You can use Python for executing the project and submit code as Jupyter Ipython Notebooks with detailed markdown text, comments and plots. Use of standard libraries is not allowed for data processing/analysis other than for data handling/visualization purposes. In case you are planning to use any other languages or analysis/visualization tools, please inform accordingly and mention in your data and final report clearly.
4. Report has to be written using the IEEE research paper format including a problem motivation, objectives, background, methodology, results and discussion, bibliography. Report will be evaluated both at mid and end semester.
   *NOTE- NO OTHER REPORT FORMAT WILL BE ACCEPTED.*
5. All source code and reports will be checked for plagiarism and any kind of plagiarism will lead to severe penalization.
6. The entire team will be expected to demo the code and present the results as per the schedule that will be made available on CMS later. A **team self-evaluation report** including team achievements and individual team contributions should be submitted in a separate document named DMSelf_TeamID_XX.pdf during the final submission.
   *NOTE- THOSE UNABLE TO ATTEND SCHEDULED DEMO WILL NOT BE MARKED.*

**For any further queries**

Please contact

Srijanee Mookherji - p20190023@hyderabad.bits-pilani.ac.in
Deepak Praturi - p20190421@hyderabad.bits-pilani.ac.in