**Data Mining, ITC**

In this course you will get a chance to mine data using decision trees, rule based systems, statistical approaches, instance based approaches, linear techniques and clustering. Using small sample datasets, you will explore simplified versions of the above algorithms, to get a solid understanding of what each involves. You will get a chance to explore and report on data mining success stories, competitions and specialized techniques. A substantial portion of this class will be applying the data mining process, including data aquisition, data cleaning, transformation & integration, feature extraction, data mining and evaluation, to an area of your choosing, in order to gain insight on a question of your choice. You are free to use **Python, R, Weka or other data mining packages**.

Software used : **Python, Jupyter notebook, R or weka**

**Data Sets**

- kaggle, world's largest data mining and machine learning community https://www.kaggle.com/datasets
- Aviation Weather Center https://www.aviationweather.gov/metar
- World Bank Open Data https://data.worldbank.org/
- Awesome Public Datasets on github https://github.com/awesomedata/awesome-public-datasets
- Robert Wood Johnson Foundation County Health Rankings & Roadmaps http://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation

**WEKA**

- Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classificiaton, regression, clustering, association rules and visualizaiton. It is free software developed at University of Waikato and licensed under the GNU General Public License.
- A manual, Weka 3.8.1 manual is available on the Weka website at, http://www.cs.waikato.ac.nz/~ml/weka/documentation.html, which can be accessed via the "Help" option within Weka.
- Weka packages can be gotten via the package manager that is under Tools in the Weka GUI Chooser. A useful package is "simpleEducationLearningSchemes".
- Attribute Relation File Format (ARFF file is an ASCII text file that describes a list of instances sharing a set of attributes.

**R**

- R project: R: The R Project for Statistical Computing (r-project.org)
- Manual of R: R-intro.pdf (r-project.org)

**Data Mining Competitions**

- DrivenData: Data Science Competitions for Social Good https://www.drivendata.org/
- Kaggle, the leading platform for data prediction competitions https://www.kaggle.com/

- CrowdANALYTIX, converts business challenges into analytics competitions https://www.crowdanalytix.com/
- Innocentive, mainly focusing on life sciences, but has other interesting competitions https://www.innocentive.com/
- TunedIT, education, research and industrial contests. http://tunedit.org/