**DA 515 – Case Studies in Data Analytics**

**Summer 2022 (202103)**

Birol Yüceoğlu

Migros T.A.Ş (IT - Research and Development)

byuceoglu@migros.com.tr

byuceoglu@sabanciuniv.edu

**Course Objectives:**

This course aims at discussing the key principles of knowledge discovery process through various case studies arising from different application areas. The students are expected to learn the main steps to traverse when they face new data analytics problems. With each case study, the tools for cleaning, processing and altering the data shall be visited. A particular attention shall be given to data inspection, manipulation and model selection. Each case study will be completed by a thorough discussion of the results.

**This syllabus may be subject to changes.**

**Course Outline**:

Python and Important Packages

Review of Machine Learning Methods

Twitter Sentiment Analysis

Unsupervised Learning – Case Study: Customer Segmentation

Supervised Learning – Case Study: Credit Scoring

Supervised Learning – Linear Regression

Anomaly Detection

Character Recognition from Images – Case Study: Google Street View

Forecasting: Predicting Egg Sales for a Retailer

Interpretable ML

**Lecture Notes:**

The lecture notes will be available on SUCourse before each class. A solution to the case study will be available after the related lectures.

**Software:**

We shall use [Anaconda (Python 3.X)](http://continuum.io/downloads). This software installs most of the necessary packages on your computer. You should install [Lightgbm](https://lightgbm.readthedocs.io/en/latest/) and [Tensorflow](https://www.tensorflow.org/install) separately. Please install these packages before coming to the class.

For the term project and homework you are advised to use [Jupyter Notebook](http://jupyter.org/) and provide adequate documentation of your work.

**Software:**

The reading summaries are for preparing before coming to class. Therefore, I will not allow for late submissions. The submissions will go through Turnitin, so be careful about using your own statements

**Grading:**

Term Project & Report 25%

Homework 10%

Reading Summary 5% \* 8

Quiz 25% (The date will be announced)

**Detailed Outline**

**Case Study 0 – Preliminaries (1 day)**

* Revisiting Python: tools, packages, programming environment, Jupyter Notebook
* Machine learning review: supervised and unsupervised learning
* Warm-up case study: Text mining on citation data of Turkish academics
* Data representation and its meaning for machine learning

**Case Study 1 – Sentiment Analysis with Twitter Data (1 day + Reading Summary)**

* I will provide a notebook on using Twitter API (Application Programming Interface) but will not discuss this further as it is getting harder to gain access to Twitter API.
* Simple sentiment analysis with word counts and NLTK
* Advanced topics in NLP: n-grams, stemming

**Case Study 2 - Supervised Learning: Personal Credit Scoring (3 days + Reading Summary)**

* Main approaches in classification
* Features, feature engineering
* Model selection, model evaluation
* Logistic regression, random forests, gradient boosting and plotting
* Parameter tuning
* Classification using Lightgbm
* Classification using Tensorflow

**Case Study 3 - Unsupervised Learning: Customer Segmentation in Retail Industry (2 days + Reading Summary)**

* Main approaches in clustering
* Features, correlations, data manipulation
* Model selection
* Clustering and plotting
* Soft clustering
* Discussion

**Case Study 4 – Supervised Learning: Predicting House Prices (1 day + Reading Summary)**

* Main approaches in regression
* Model selection, model evaluation
* Regularization

**Case Study 5 – Anomaly Detection (1 day + Reading Summary)**

* Main approaches in anomaly detection
* Review of autoencoders
* Data anonymization
* Anomaly detection
* Variational autoencoders

**Case Study 6 – Character Recognition from Images (1 day + Reading Summary)**

* Main approaches in image processing
* Review of classification methods and convolutional neural networks
* Model selection
* Performance improvements using data augmentation
* Discussion

**Case Study 7 – Predicting Egg Sales for a Retailer (1 day + Reading Summary)**

* Main approaches in regression and time series
* Review of regression methods and neural networks (LSTM)
* Model selection
* Discussion

**Case Study 8 – Interpretable Machine Learning (1 day + Reading Summary)**

* Main approaches in regression and time series
* Review of regression methods and neural networks (LSTM)
* Model selection
* Discussion

**Case Study 9 – Your choice (1 day)**

**Project Presentation (1 day)**