Project 1
Covertype dataset
https://archive.ics.uci.edu/ml/datasets/covertype
The Covertype dataset records the types of forest-covering parcels of land in Colorado, USA. Each example contains several features describing each parcel of land—like its elevation, slope, distance to water, shade, and soil type—along with the known forest type covering the land. The forest cover type is to be predicted from the rest of the categorical and numerical features, of which there are 54 in total. There are 581,012 recordings in this dataset. This is a *classification* task

Project 2
Census Income dataset
https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29
This data set contains weighted census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. The data contains 41 demographic and employment related variables. There are 199,523 instances in the data file and 99,762 in the test file. Note that Incomes have been binned at the $50K level to present a binary classification problem.

Project 3
YearPredictionMSD Data Set
https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD
This dataset is a subset of the Million Song Dataset: (http://labrosa.ee.columbia.edu/millionsong/). The dataset has been pre-processed. In particular, the numerical audio features are extracted (by using. the Echo Nest API). The first attribute is the release year. Other attributes include two group: the timbre average (12 columns) and the timbre covariance (78 columns). The task is to predict the release year of a song from audio features. Use the first 463,715 examples as the training dataset and the last 51,630 examples as the test dataset. This is a regression problem.

Project 4
Record linkage dataset
https://archive.ics.uci.edu/ml/datasets/Record+Linkage+Comparison+Patterns
The dataset contains element-wise comparison of records with personal data from a record linkage setting. The task is to decide from a comparison pattern whether the underlying records belong to one person. The records represent individual data including first and family name, sex, date of birth and postal code, which were collected through iterative insertions in the course of several years. The comparison patterns in this data set are based on a sample of 100,000 records. Data pairs were classified as “match” or “non-match”. Thus “is_match” is the outcome variable. Note that “id_1” and “id_2” should not be used for prediction but could be used to construct connected components from the found matches.

Project 5
UNSW network intrusion dataset
https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/
Several datasets are available for model development and model testing for IDS. This project will utilize the UNSW-NB15 dataset. The UNSW-NB15 dataset is published by Cyber Range Lab of the Australian Centre for Cyber Security. The data was collected over 15 hours by an IXIA traffic generator in 2014, then pre-processed and labelled as “normal” and various types of “attack”. Download the *training dataset* and the *test dataset* from the above link. The task is to predict whether a record represents “normal” or “attack” (a binary classification problem). Note that the last two columns represent the targe variables, which should not be used as training features.