Challenge 6

Supervisory Control And Data Acquisition (SCADA) systems are computer-based process control systems that control and monitor remote physical processes, and normally span a large geographical area such as a gas pipeline, power transmission or water distribution system. As SCADA systems are widely being used in critical infrastructure, they're strategically important to countries. Thus, identifying anomalous incidents and cyberattacks against such systems is of crucial importance.

In this challenge, you are provided with a dataset of transactions from a gas pipeline system in Attribute Relationship File Format (ARFF). There are a total number of 97,019 transactions in this dataset. Each transaction has 26 different features that can be categorized into three main groups, including network traffic, process control and process measurement. These features are summarized as below:

Network traffic features collected from MODBUS systems:

- **command_address**: Device ID in command packet
- **response_address**: Device ID in response packet
- **command_memory**: Memory start position in command packet
- **response_memory**: Memory start position in response packet
- **command_memory_count**: Number of memory bytes for R/W command
- **response_memory_count**: Number of memory bytes for R/W response
- **command_length**: Total length of command packet
- **response_length**: Total length of response packet
- **time**: Time interval between two packets
- **crc_rate**: CRC error rate (a measurement of the rate of CRC errors identified in command and response packets.)

Process control features:

- **measurement**: Pipeline pressure
- **pump**: Compressor state
- **control_mode**: Automatic (2), manual (1) or shutdown (0)
- **sub_function**: Value of sub-function code in the command/response

- **comm_read_function**: Command read function code
- **comm_write_fun**: Command write function code
- **resp_read_fun**: Response read function code
- **resp_write_fun**: Response write function code

Process measurement features:

- **setpoint**: Target gas pressure in the pipe
- **control_scheme**: Control scheme of the gas pipeline
- **solenoid**: State of solenoid used to open the gas relief valve
- **gain**: Gain parameter value of the PID controller
- **reset**: Reset parameter value of the PID controller
- **deadband**: Dead band parameter value of the PID controller
- **rate**: Rate parameter value of the PID controller
- **cycletime**: Cycle time parameter value of the PID controller

Prepare a report and submit a PDF file by Tuesday (**11/17/2020 before 6 pm**) considering the below details:

1. Apply the Isolation Forest algorithm to provided dataset and report the number of outliers found by the algorithm.
2. Use the "decision_function" method of this library to calculate the average anomaly score of all transactions and report the number of transactions that have an average score of -0.2 or smaller values (i.e., Avg_Score <= -0.2). Also, report the exact transactions (i.e., data instances or rows) that have such average anomaly scores.
3. Apply the Local Outlier Factor algorithm to the same dataset and report the number of outliers found by the algorithm (set number of neighbors, k, to 3).
4. Now, this time, calculate the LOF score of all transactions and report the number of data points (or transactions) that have an LOF score smaller or equal to -40 (i.e., LOF_score <= -40).