CSCI - 4146 - The Process of Data Science - Fall 2020
Assignment 2

The submission must be done through Brightspace.
Due date and time as shown on Brightspace under Assignments.

- To prepare your assignment solution use the assignment template notebook available on Brightspace.
- The detailed requirements for your writing and code can be found in the evaluation rubric document on Brightspace.
- Questions will be marked individually with a letter grade. Their weights are shown in parentheses after the question.
- Assignments can be done by a pair of students, or individually. If the submission is by a pair of students, only one of the students should submit the assignment on Brightspace.
- We will use plagiarism tools to detect any type of cheating and copying (your code and PDF).
- Your submission is a single Jupyter notebook and a PDF (With the compiled results generated by your Jupyter notebook). File names should be:
  - **A2-<your_name1>-<your_name2>.ipynb**
  - **A2-<your_name1>-<your_name2>.pdf**
- **Forgetting to submit both files results in 0 markings for both students.**

Predictive maintenance (PdM) is gaining traction in the industry. In PdM, components are replaced as they approach failure, not at prescribed intervals (Preventative Maintenance). For PdM, equipment is monitored by sensors, and machine learning models are used to predict the remaining useful life (RUL) (Fig 1.) of the equipment based on data streams generated by the sensors. The data is typically a time series of sensor measurements collected until failure.

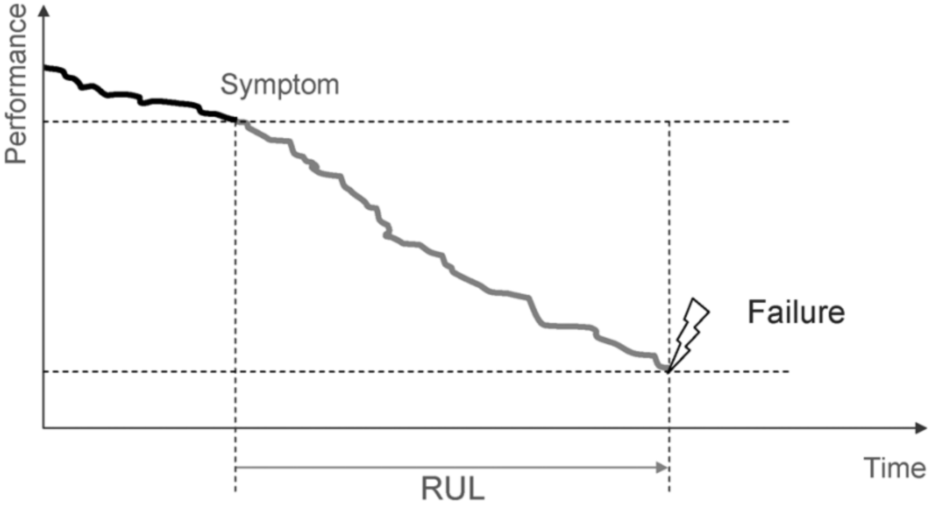
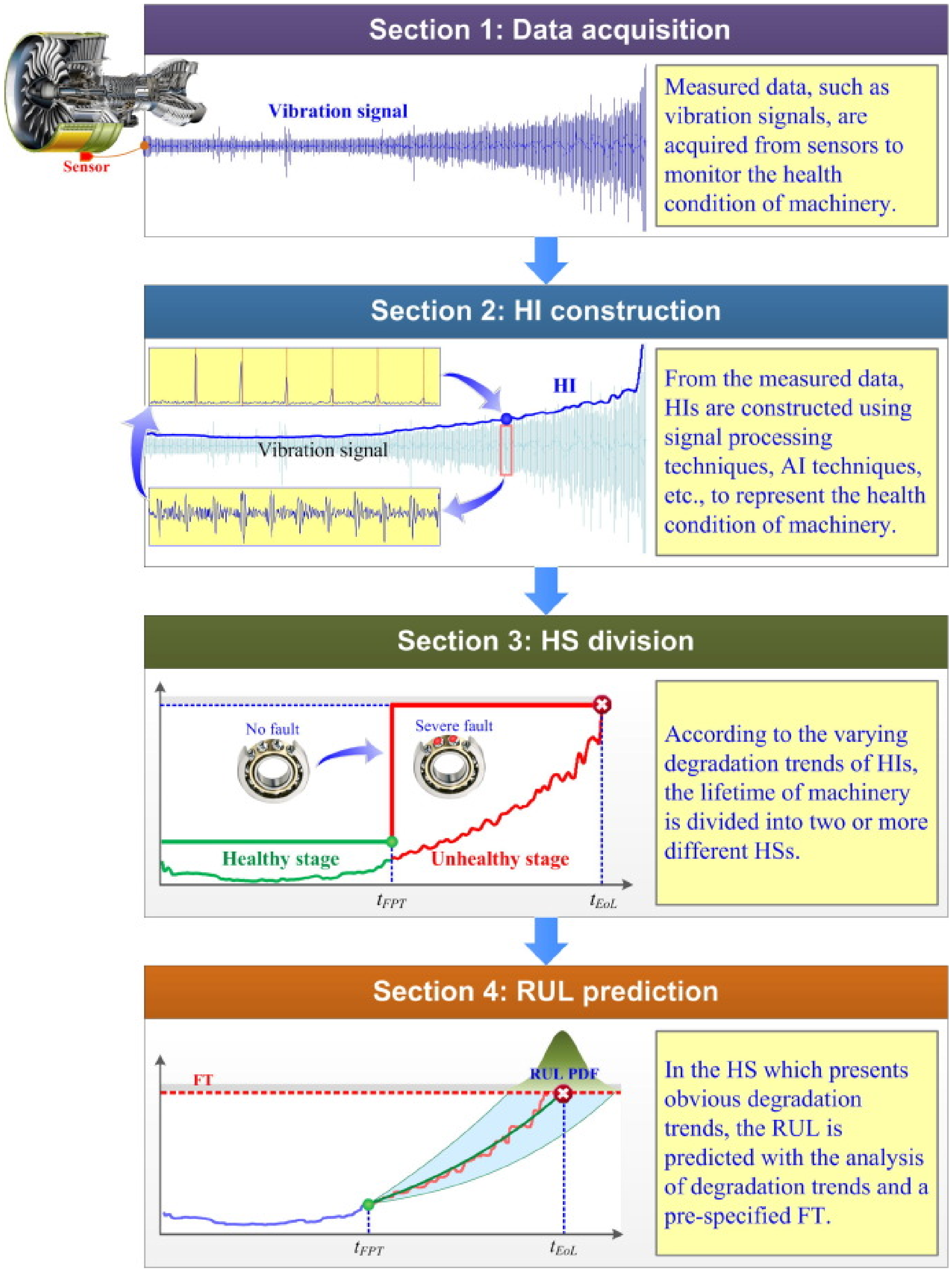


Figure 1: Illustration of an RUL.[1]

As shown in (Fig 2), a machinery health prognostic program is generally composed of four technical processes, i.e., data acquisition, health indicator (HI) construction, health stage (HS) division and RUL prediction. At first, measured data, such as vibration signals, are acquired from sensors to monitor the health condition of machinery. Then, from the measured data, HIs are constructed using signal processing techniques, artificial intelligent (AI) techniques, etc., to represent the health condition of machinery. After that, according to the varying degradation trends of HIs, the whole lifetime of machinery is divided into two or more different HSs. Finally, in the HS which presents an obvious degradation trend, the RUL is predicted with the analysis of the degradation trends and a pre-specified failure threshold (FT).[2]



Section 1: Data acquisition

Measured data, such as vibration signals, are acquired from sensors to monitor the health condition of machinery.

Section 2: HI construction

From the measured data, HIs are constructed using signal processing techniques, AI techniques, etc., to represent the health condition of machinery.

Section 3: HS division

According to the varying degradation trends of HIs, the lifetime of machinery is divided into two or more different HSs.

Section 4: RUL prediction

In the HS which presents obvious degradation trends, the RUL is predicted with the analysis of degradation trends and a pre-specified FT.

Figure 2: Four technical processes in a machinery health prognostic program.[2]

In this assignment, you will need to predict an RUL of Turbofan Engine Degradation. For the specific data set on turbofan engine(#6 of the datasets in the NASA Prognostics Center repository, https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/), Data sets consists of multiple multivariate time series. Each data set is further divided into training and test subsets. Each time series is from a different engine – i.e., the data can be considered to be from a fleet of engines of the same type. Each engine starts with different degrees of initial wear and manufacturing variation, which is unknown to the user. This wear and variation are considered normal, i.e., it is not considered a fault condition. There are three operational settings (altitude (0-42K ft.), Mach number (0-0.84), and TRA (20-100)) that have a substantial effect on engine performance. These settings are also included in the data. The data is contaminated with sensor noise.

The engine is operating normally at the start of each time series and develops a fault at some point. In the training set, the fault grows in magnitude until system failure. In the test set, the time series ends at some time prior to system failure. The objective of the assignment is to predict the number of remaining operational cycles before failure in the test set, i.e., the number of operational cycles after the last cycle that the engine will continue to operate. Also provided a vector of true Remaining Useful Life (RUL) values for the test data.

The data are provided as a zip-compressed text file with 26 columns of numbers, separated by spaces. Each row is a snapshot of data taken during a single operational cycle, each column is a different variable. The columns correspond to:
1) unit number
2) time, in cycles
3) operational setting 1 (altitude)
4) operational setting 2 (Mach number)
5) operational setting 3 (TRA - throttle resolver angle)
6 - 26) sensor measurements: The value of the sensor will act as the feature for predicting RUL
**Note -> The meaning of the sensors are not relevant to the exercise.
These sensor record the following parameter not necessarily in the same order

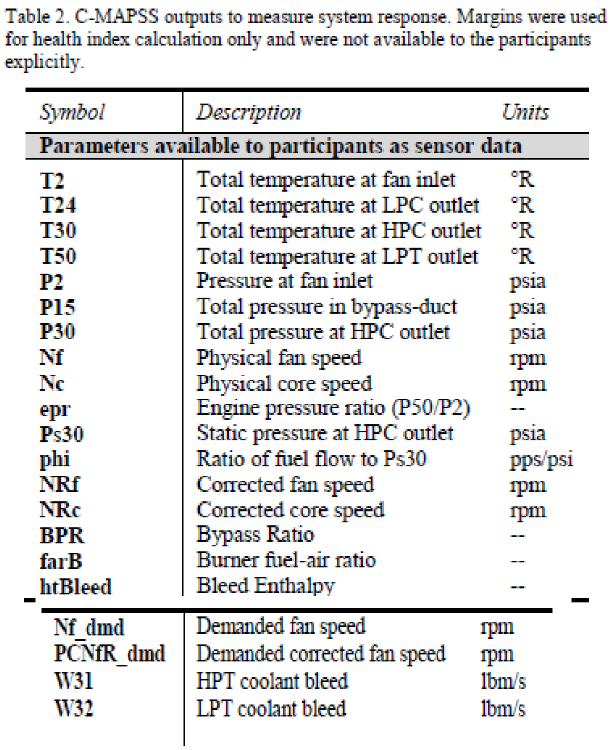


Table 2. C-MAPSS outputs to measure system response. Margins were used for health index calculation only and were not available to the participants explicitly.

Symbol	*Description*	*Units*
Parameters available to participants as sensor data		
T2	Total temperature at fan inlet	°R
T24	Total temperature at LPC outlet	°R
T30	Total temperature at HPC outlet	°R
T50	Total temperature at LPT outlet	°R
P2	Pressure at fan inlet	psia
P15	Total pressure in bypass-duct	psia
P30	Total pressure at HPC outlet	psia
Nf	Physical fan speed	rpm
Nc	Physical core speed	rpm
epr	Engine pressure ratio (P50/P2)	--
Ps30	Static pressure at HPC outlet	psia
phi	Ratio of fuel flow to Ps30	pps/psi
NRf	Corrected fan speed	rpm
NRc	Corrected core speed	rpm
BPR	Bypass Ratio	--
farB	Burner fuel-air ratio	--
htBleed	Bleed Enthalpy	--
Nf_dmd	Demanded fan speed	rpm
PCNfR_dmd	Demanded corrected fan speed	rpm
W31	HPT coolant bleed	lbm/s
W32	LPT coolant bleed	lbm/s

1. **Data understanding and feature engineering (0.1)**
    a. We will extract features from each channel of each of the data files of train_FD002. Calculate the RUL for the training data knowing the fact that the end of the cycle for each engine is reported as the failure of the engine.
    b. Build the data quality report
    c. Identify data quality issues and build the data quality plan
    d. Analyze your data. Plot the 21 features as functions of time or cycle for each of the engines. Compute and plot the trends of the sensor output for each engine against the RUL. Describe your observations. How similar are the plots of the different engines? Is there any evidence in the plots for which features are the most useful for the RUL prediction task? Is the normalization of the data useful?
    e. Preprocess your data according to the data quality plan

2. Build a baseline model to predict RUL (0.35). For FD002, three files are training data, test data, and the true RUL for each engine in testing data.

a. Explain what the task you're solving is (e.g., supervised x unsupervised, classification x regression x clustering or similarity matching x etc)
b. Use a feature selection method to select the features to build a model.
c. Select the evaluation metric. Justify your choice.
d. Perform hyperparameter tuning if applicable.
e. Train and evaluate your model on test data from Test set 1
f. How do you make sure not to overfit?
g. Plot learning curve
h. Analyze the results

3. Build a NN model to predict RLU (0.35). Repeat question #2 above but now use a neural network model to predict RLU. You can use a simple feedforward neural network or 1D CNN from tutorial 6. Compare the model to your baseline model with a statistical significance test. Use a box plot to visualize your comparison.

4. Concept drift detection (0.2). Use concept drift methods and find out if there is any drift in the data that can be detected. If so, what type of drift is that? Suggest specific actions to adapt your model to the new concept.

References:

[1] D. A. Tobon-Mejia, K. Medjaher, N. Zerhouni and G. Tripot, "A Data-Driven Failure Prognostics Method Based on Mixture of Gaussians Hidden Markov Models," in IEEE Transactions on Reliability, vol. 61, no. 2, pp. 491-503, June 2012

[2] Machinery health prognostics: A systematic review from data acquisition to RUL prediction. 2018. Yaguo Lei ⇑ , Naipeng Li, Liang Guo, Ningbo Li, Tao Yan, Jing Lin

[4] A. Saxena and K. Goebel (2008). "Turbofan Engine Degradation Simulation Data Set", NASA Ames Prognostics Data Repository