**Machine learning Internship Projects by UCT**

**Deliverables for each Project:**

**Code:** on Github

**Project Report:** also on Github, mentioning about company (UCT), Background of the project, Problem statement relevance, Design, Implementation details, Result, your learnings.

Machine learning Internship project scope varies from current scenario, industrial demand, as well as projects running in our company. So we divide them in below areas:

1. Agriculture
2. Predictive maintenance
3. Smart city
4. Industrial Manufacturing and Production

**~~Project 1:~~** ~~Obsolete~~

**~~Project 2:~~** ~~Obsolete~~

**~~Project 3:~~** ~~Obsolete~~

1. **Agriculture**

##### **Project 4: Prediction of Agriculture Crop Production in India**

Context

Agricuture Production in India from 2001-2014

Content

This Dataset Describes the Agricuture Crops Cultivation/Production in india. This is from <https://data.gov.in/> fully Licensed

Acknowledgements

This Dataset can solves the problems of various crops Cultivation/production in india.

Columns

Crop: string, crop name  
Variety:string,crop subsidary name  
state: string,Crops Cultivation/production Place  
Quantity:Integer,no of Quintals/Hectars  
production:Integer,no of years Production  
Season:DateTime,medium(no of days),long(no of days)  
Unit:String , Tons  
Cost:Integer, cost of cutivation and Production  
Recommended Zone:String ,place(State,Mandal,Village)

Inspiration

Across The Globe India Is The Second Largest Country having People more than 1.3 Billion.  
Many People Are Dependent On The Agricuture And it is the Main Resource.  
In Agricuturce Cultivation/Production Having More Problems.  
I want to solve the Big problem in india and usefull to many more people

## Data set Link:

<https://drive.google.com/file/d/1zfqvs8-mAO6E0JpgvhBdueNx8Th03pUp/view?usp=sharing>

##### **Project 5: Crop and weed detection**

# **Content**

This dataset contains 1300 images of sesame crops and different types of weeds with each image labels.  
Each image is a 512 X 512 color image. Labels for images are in YOLO format.

# **Data Preparation**

STEPS:

1. First of we have to collect dataset for it.For that we have to capture photos of weeds and crops. We collected total 589 images
2. After collection of photos we have to clean the dataset. This step is very important because if any bed photo is remain in dataset it causes worse effect in detection model. After cleaning we have 546 images.
3. Now time for image processing. Our photo size is 4000X3000 color which is very large and model will take very long time for training so we convert all images to 512X512X3 size.
4. Now 546 image is not enough for training, so we have done some magic to convert 546 image into 1300 images. We used Data Augmentation technique to increase dataset.(Check it out keras ImageDataGenerator on google)
5. This step is very tedious, Manual labeling of image data!! In this step we have to drow bounding boxes on photos whether it weed or crop.

# **Problem**

Weed is an unwanted thing in agriculture. Weed use the nutrients, water, land and many more things that might have gone to crops. Which results in less production of the required crop. The farmer often uses pesticides to remove weed which is also effective but some pesticides may stick with crop and may causes problems for humans.

# **Aim**

We aim to develop a system that only sprays pesticides on weed and not on the crop Which will reduce the mixing problem with crops and also reduce the waste of pesticides.

## Data set Link:

<https://drive.google.com/file/d/1MNdDKYB0x0PEW7P71bE1Jx_uLllvORA0/view?usp=sharing>

1. **Predictive maintenance**

##### **Project 6: Predict the number of remaining operational cycles before failure for Turbofan engine**

**Experimental Scenario**

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 is considered normal, i.e., it is not considered a fault condition. There are three operational settings 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 during the series. In the training set, the fault grows in magnitude until system failure. In the test set, the time series ends some time prior to system failure. The objective of the competition 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

4) operational setting 2

5) operational setting 3

6) sensor measurement 1

7) sensor measurement 2

...

26) sensor measurement 26

Data Set: FD001

Train trajectories: 100

Test trajectories: 100

Conditions: ONE (Sea Level)

Fault Modes: ONE (HPC Degradation)

Data Set: FD002

Train trajectories: 260

Test trajectories: 259

Conditions: SIX

Fault Modes: ONE (HPC Degradation)

Data Set: FD003

Train trajectories: 100

Test trajectories: 100

Conditions: ONE (Sea Level)

Fault Modes: TWO (HPC Degradation, Fan Degradation)

Data Set: FD004

Train trajectories: 248

Test trajectories: 249

Conditions: SIX

Fault Modes: TWO (HPC Degradation, Fan Degradation)

## Data set Link:

<https://drive.google.com/file/d/1dgWM0KKOnoN9kVObbA-GahsgXPJBCT4c/view?usp=sharing>

##### **Project 7: Predict life time of a bearing in manufacturing industry**

***Data Structure***

Three (3) data sets are included in the data packet (IMS-Rexnord Bearing Data.zip). Each data set describes a test-to-failure experiment. Each data set consists of individual files that are 1-second vibration signal snapshots recorded at specific intervals. Each file consists of 20,480 points with the sampling rate set at 20 kHz. The file name indicates when the data was collected. Each record (row) in the data file is a data point. Data collection was facilitated by NI DAQ Card 6062E. Larger intervals of time stamps (showed in file names) indicate resumption of the experiment in the next working day.

**Set No. 1:**

Recording Duration: October 22, 2003 12:06:24 to November 25, 2003 23:39:56

No. of Files: 2,156

No. of Channels: 8

Channel Arrangement: Bearing 1 – Ch 1&2; Bearing 2 – Ch 3&4;

Bearing 3 – Ch 5&6; Bearing 4 – Ch 7&8.

File Recording Interval: Every 10 minutes (except the first 43 files were taken every 5 minutes)

File Format: ASCII

Description: At the end of the test-to-failure experiment, inner race defect occurred in bearing 3 and roller element defect in bearing 4.

**Set No. 2:**

Recording Duration: February 12, 2004 10:32:39 to February 19, 2004 06:22:39

No. of Files: 984

No. of Channels: 4

Channel Arrangement: Bearing 1 – Ch 1; Bearing2 – Ch 2; Bearing3 – Ch3; Bearing 4 – Ch 4.

File Recording Interval: Every 10 minutes

File Format: ASCII

Description: At the end of the test-to-failure experiment, outer race failure occurred in bearing 1.

**Set No. 3 :**

Recording Duration: March 4, 2004 09:27:46 to April 4, 2004 19:01:57

No. of Files: 4,448

No. of Channels: 4

Channel Arrangement: Bearing1 – Ch 1; Bearing2 – Ch 2; Bearing3 – Ch3; Bearing4 – Ch4;

File Recording Interval: Every 10 minutes

File Format: ASCII

Description: At the end of the test-to-failure experiment, outer race failure occurred in bearing 3

## Data set Link:

<https://drive.google.com/file/d/12rV9AhpqbMivYCu4WVM7DhXpO1aO98k_/view?usp=sharing>

##### **Project 8: Predictive maintenance of Gearbox using vibration sensors**

**Predictive maintenance** allows manufacturers to lower **maintenance** costs, extend equipment life, reduce downtime and improve production quality by addressing problems before they cause equipment failures.

Gearbox Fault Diagnosis Data set include the vibration dataset recorded by using SpectraQuest’s Gearbox Fault Diagnostics Simulator. Dataset has been recorded with the help of 4 vibration sensors placed in four different direction. Data set has been recorded under variation of load from '0' to '90' percent. Data set has been recorded in two different scenario:

1) Healthy condition and

2) Broken Tooth Condition

## Data set Link:

https://drive.google.com/file/d/1nNNnjMPntlo5X0t\_cif7cmlJhikCyWyP/view?usp=sharing

1. **Smart city**

##### **Project 9: Forecasting of Smart city traffic patterns**

We are working with the government to transform various cities into a smart city. The vision is to convert it into a digital and intelligent city to improve the efficiency of services for the citizens. One of the problems faced by the government is traffic. You are a data scientist working to manage the traffic of the city better and to provide input on infrastructure planning for the future.

The government wants to implement a robust traffic system for the city by being prepared for traffic peaks. They want to understand the traffic patterns of the four junctions of the city. Traffic patterns on holidays, as well as on various other occasions during the year, differ from normal working days. This is important to take into account for your forecasting.

## Data set Link:

https://drive.google.com/file/d/1y61cDyuO9Zrp1fSchWcAmCxk0B6SMx7X/view?usp=sharing

**Industrial Manufacturing and Production**

**Project 10:** Quality Prediction in a Mining Process

## Explore real industrial data and help manufacturing plants to be more efficient

### Context

It is not always easy to find databases from **real world** manufacturing plants, specially mining plants. This database comes from one of the most important parts of a mining process: a [flotation plant](https://en.wikipedia.org/wiki/Froth_flotation).

The main goal is to use this data to predict how much impurity is in the ore concentrate. As this impurity is measured every hour, if we can predict how much silica (impurity) is in the ore concentrate, we can help the engineers, giving them early information to take actions (empowering!). Hence, they will be able to take corrective actions in advance (reduce impurity, if it is the case) and also help the environment (reducing the amount of ore that goes to tailings as you reduce silica in the ore concentrate).

### Content

The first column shows time and date range (from march of 2017 until september of 2017). Some columns were sampled every 20 second. Others were sampled on a hourly base.

The second and third columns are quality measures of the iron ore pulp right before it is fed into the flotation plant. Column 4 until column 8 are the most important variables that impact in the ore quality in the end of the process. From column 9 until column 22, we can see process data (level and air flow inside the flotation columns, which also impact in ore quality. The last two columns are the final iron ore pulp quality measurement from the lab.  
Target is to predict the last column, which is the % of silica in the iron ore concentrate.

### Expected submission

* Is it possible to predict % Silica Concentrate every minute?
* How many steps (hours) ahead can we predict % Silica in Concentrate? This would help engineers to act in predictive and optimized way, mitigating the % of iron that could have gone to tailings.
* Is it possible to predict % Silica in Concentrate without using % Iron Concentrate column (as they are highly correlated)?

**Dataset**

This dataset is about a flotation plant which is a process used to concentrate the iron ore. This process is very common in a mining plant.

## Data set Link:

<https://drive.google.com/file/d/1N80d8eTDAf1JMQXGQbHDAUaMGRyA8QG3/view?usp=sharing>

# **Project 11:** Multi-stage continuous-flow manufacturing process

## Real process data to predict factory output

### Context

This data was taken from an actual production run spanning several hours. The goal is to predict certain properties of the line's output from the various input data. The line is a high-speed, continuous manufacturing process with parallel and series stages.

### Expected submission

We are always looking for the best predictive modeling approaches to use in real time production environments. Models are employed for several use cases such as development of real time process controllers (use the models in simulation environments) and anomaly detection (compare model predictions to actual outputs in real time).

**Dataset**

The data comes from a continuous flow manufacturing process with multiple stages. Sample rates are 1 Hz.

* In the first stage, Machines 1, 2, and 3 operate in parallel, and feed their outputs into a step that combines the flows.
* Output from the combiner is measured in 15 locations. These measurements are the primary measurements to predict.
* Next, the output flows into a second stage, where Machines 4 and 5 process in series.
* Measurements are made again in the same 15 locations. These are the secondary measurements to predict.

## Data set Link:

<https://drive.google.com/file/d/1yvZzslpbWw2mpCVF5QqueSkNrNHmtvDE/view?usp=share_link>