# SIT225 Data Capture Technologies

## Pass Task: Data analysis and interpretation

## Overview

Plotly charts can be used for displaying various types of regression models, starting from simple models like Linear Regression. Scikit-learn is a popular Machine Learning (ML) library which can be used to train various regression models. The library was designed to be accessible, and to work seamlessly with popular libraries like NumPy and Pandas.

## Hardware Required

i. Arduino Nano 33 IoT device,
ii. USB cable,
iii. DHT22 sensor.

## Software Required

i. Python 3,
ii. Plotly Python library,
iii. Numpy and Pandas Python libary
iv. Sklearn Python ML library

## Pre-requisites: You must do the following before this task

Week 7 activities in the unit site.

## Task Objective

In recent weeks, you have learned how to use Plotly charts and Dash for visualisation and monitoring. You have also learned how machine-learning models such as Linear Regression (LR) can be used to learn correlation between data variables and generate predictions on unknown data points. In this task, you will need to use LR to learn relationships between sensor variables you have captured data such as DHT22 having 2 variables – temperature and humidity so you can use the learned LR model to predict humidity for unknown temperature.

Steps:

1. Connect the sensor for collecting data samples for more than 30 minutes, possibly overnight, if you do not have collected so previously. The more data samples you have, the LR will have a better chance of learning patterns.
2. Data should be available in CSV format.

3. Create LinearRegression model from sklearn, train with your data points (by calling model's ***fit(X, y)*** function) considering temperature values as independent variable (X axis) and humidity as dependent variables (Y axis).
4. Find the min and max temperature values and interpolate to create 100 equally distant temperature values in between, which will be the test temperature values, and predict humidity for all 100 test temperature values.
5. Scatter plot temperature vs humidity of your data and on the same plot, create a line-plot of your test temperature and humidity which will show the trend of temperature and humidity. Analyse the trend line and comment on below points -
    a. The trend line follows the original data points?
    b. Can you find any outlier in the sample points? Outliers are sample points which are farther apart from the trend lines.
6. Try to filter a few of the outlying samples, possibly by removing 5-10 high temperatures and low temperatures. You can manually identify them or free to use Pandas dataframe data filter capabilities to remove samples above a maximum temperature value, as well as below a minimum temperature value. You can search Pandas filter for more detail (a quick search popped up https://www.geeksforgeeks.org/ways-to-filter-pandas-dataframe-by-column-values ).
7. Repeat steps 3-5 to train LR model, create test temperature/humidity values, create plot with original training samples and trend line from test samples. Compare these 2 scenarios -
    a. The trend line is still the same or slope of the line changed a bit, did the LR model learning different pattern? Justify your answer.
8. Filter a few more outliers again and repeat 3-5 and compare the trend line w.r.t. the filtered reduced training sample points and provide your insight.

## Submission details

Q1. Perform week 7 activities mentioned in the unit site and produce outputs.

Q2. Describe different scenarios mentioned in the steps in the task objective section. Try to be as detailed as possible. Note that the graphs and trend line might be different for other students based on the day/time of data capture, the choice of min/max temperature threshold for outlier removal and the number of samples used in LR training.

Q3. Create a video in Panopto/CloudDeakin showing your program execution, graph output for different scenarios, and share the video link here.

Q4. Create a subdirectory 'week-7' under directory 'SIT225_2024T2' in your drive where you copy the Python script file, Arduino sketch file if any, data file and the generated graphs. Commit and push to changes to GitHub. Include the link to your repository here with a GitHub page screenshot of weekly folder content. A tutor may try to access your GitHub link, if necessary. Give access to your tutor by adding tutor's email address as a collaborator of your private repository.

# Instructions

Consolidate outputs following the submission details above into a single PDF file.

## Submit your work

When you are ready, login to OnTrack and submit your pdf which consolidates all the items mentioned in the submission detail section above. Remember to save and backup your work.

## Complete your work

After your submission, your OnTrack reviewer (tutor) will review your submission and give you feedback in about 5 business days. Your reviewer may further ask you some questions on the weekly topics and/or about your submissions. You are required to address your OnTrack reviewer's questions as a form of task discussions. Please frequently login to OnTrack for the task ***Discuss/Demonstrate*** or ***Resubmit*** equivalent to fix your work (if needed) based on the feedback to get your task signed as ***Complete***.