**Build a model to predict performance of employees, based on the visualisation and analysis of past data of employee performance**

TASK

**Reading your data into a Dataframe in Python**

To read the data into a dataframe in Python, you must execute the following steps:

* Import packages
* Read your data from CSV to a dataframe using Pandas
* Take a closer look at the data using the str() and head() functions and check if the data got imported properly

**Data Cleaning**

***Data cleaning*** is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted.  
  
In this task, you can execute the below steps & proceed further

* Treat NULL Values (if any)
* Remove duplicate data / check all the columns and look for the useful metrics which can be employed in the model

**Visualize Count Plots & Unique Values to infer from Datasets**

**Data visualization** is the graphical representation of information and **data**. By using visual elements like charts, graphs, and maps, **data visualization** tools provide an accessible way to see and understand trends, outliers, and patterns in **data.**In this task, you must execute the below steps:

* Use Charts (Count plots) to examine and compare the features present in the dataset
* Identify trends and insights from the above plots

**Data Scaling**

**Scaling** is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing.

In this task, We use the below-mentioned Scalers to check their effect on our data and choose the most appropriate Scaler based on the results. More on Scalers and their purpose has been provided as a resource in the resource hub.  
  
RobustScaler, StandardScaler, MinMaxScaler, MaxAbsScaler

* Create a function to run the given four scalers on your dataset and return the resultant scaled dataset. Make changes to parameters in such a way that you can iterate over various scalers to get datasets.

**Create Baseline ML Model for Binary Classification Problem**

In this task:

* We will be using 3 models for this problem -  XGBoost, CatBoost, LightGBM
* Run the dataset through XGBoost to predict the target variable and check the accuracy
* Run the dataset through CatBoost to predict the target variable and check the accuracy
* Run the dataset through LightGBM to predict the target variable and check the accuracy

**Report Results**

In this task, you need to check for various scalers and check which gives better accuracy.  
  
You can follow the below steps:

* Iterate over the data scaling functions and ML Models to understand which model under which scaling technique is giving better results using the function in Task 1 and the baseline models in Task 2.
* Document the accuracy using Confusion matrices and determine the best Scaling Function and best model for your dataset.
* **Make a final decision on the model to be employed for this purpose**
* Based on your work done in the previous components and tasks in this Menternship, you must now make a final decision on which model must be used to predict employee performance.  
    
  You will be required to document the analysis you have from the application of the models in the precious component, and as to why the algorithm you have chosen makes the most sense.