Data science

**Garments Worker Productivity**

Report for Dr.Sherif eletreby

horizontal line

# Placeholder image

# Problem Definition

The Garment Industry is one of the key examples of the industrial globalization of this modern era. It is a highly labour-intensive industry with lots of manual processes. Satisfying the huge global demand for garment products is mostly dependent on the production and delivery performance of the employees in the garment manufacturing companies. So, it is highly desirable among the decision makers in the garments industry to track, analyse and predict the productivity performance of the working teams in their factories. This dataset can be used for regression purpose by predicting the productivity range (0-1) or for classification purpose by transforming the productivity range (0-1) into different classes.

## 

### Method

The method used in this dataset or the data driven approach is regression, I used the read\_csv in the first line to assign the dataset to the dset variable, i used isna().any() to know if there is any missing values in any column , used isnull.sum() to sum all the null values, to fill these null values I have used the fillna() function and the 0 argument to fill them with zeros, i have used the unique() in line 8 to know how many unique values are in the quarter column then i made a dictionary and assigned each quarter with their values after that i used map(q\_dic) to map the dictionary in the quarter column, and so with the other column which has object data types, i dropped the date column because i will not benefit anything from it, in line 19 i used the heatmap in the seaborn library to visualize the correlation of the features, used value\_counts().plot.pie() to visualize the department’s values, finally for the visulaization i made i histogram to know each outliers in each column, for feature selection i called the corr() function to sort values by descending order for showing the least feature with relation which is the department so i dropped it, and for the outliers i used the quantile() with 25 % for the q1 and for q3 75% then subtract them to get the IQR then calculate it and remove any values that are less than 25 and greater than 75, after removing outliers the new dataset shape has 936 records and 13 column, for normalization i used the min max scaler to fit the dataset, and for spitting the data i had to assign the last column which is the label in a new variable the drop it afterward i split the training data which has the features by taking 80% of the dataset and 20% for the testing set, and so on for the label y then i printed them in last line of cell no.31 finally we used alot 3 algorithms one of them is linear regression i fit it with the training features and label, then for testing it i used the predict function with the test set and for measuring the accuracy i used the mean square error with the testing label and the prediction variable and printed root mean square error of linear regression which has a 18% error.

### References

<http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html>

<http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>

<http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>

<http://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html>

<http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>

<https://www.youtube.com/watch?v=ksvJDLdc9eA&t=153s>

<https://www.youtube.com/watch?v=A3gClkblXK8&t=150s>