**Data Preprocessing using Python**

by,

**team\_pyCrastinators**

* **Ajay Edaparambil**
* **Akhilesh L**
* **Amna S**
* **Arun H Das**
* **Dhiya Markose**
* **Fasna**
* **Glaffy**
* **Hasba Habeeb**
* **Rohan Philip Villoth**
* **Rubseena**

**Data Mining**: the practice of examining large pre-existing databases in order to generate new information. It involves discovering patterns in large datasets. And when the right data is mined out, we could use it for analysis, get/set market trends, and above all, predict.

This mined data is not perfect, as it has to be moulded into a perfect data set, or cleaned thoroughly. Hence, the process called ***data preprocessing*** is done.

And here, we use the well known programming language ***python*** for preprocessing.

Python offers some dedicated libraries like Scikit-learn, Pandas etc, exclusively for data analytics.

Data Preprocessing has different stages, which are: Data cleaning, integration, transformation and discretization.

**Data cleaning**

We handle the missing data here. Some datasets might contain them.

**Deleting missing tuple**: If the dataset is very large, and if very few values are missing; it can be dropped i.e. deleted. Since the dataset is huge, deleting some values are not going to affect the whole data.

The function dropna() is used.

**Value substitution**: The missing values can be replaced by some values, which either the user can input or through specific functions like mean,median or mode.

The function fillna(<value>) is used to replace the missing NaN value with <value>.

The function fillna(object.mean()) is used to replace the missing NaN value with the mean of that column/row.

Similarly fillna(object.median())and fillna(object.mode()) is used for median and mode replacement respectively.

**Data transformation**

The cleaned data is transformed to a scaled dataset here.

There are some methods like:  
**MinMax scaling** where the cleaned data is scaled to a value between 0 and 1.

The function MinMaxScaler() is used.

**Binarization** where the cleaned data is transformed to either 0s or 1s. A threshold value is kept, and if the values are greater than the threshold; it is transformed to 1 else transformed to 0.

The function Binarizer() is used.