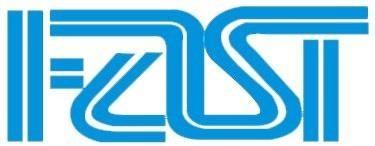
***National University of Computer &***

***Emerging Sciences Karachi Campus***



DATA CLEANSING

***Group Members***

1. **Shahzaib Khan (19k-0273)**
2. **Taha Yousuf (19k-1352)**
3. **Fatima Jamal (19k-1107)**

# **INTRODUCTION**

# Situations in which we handle large amounts of information, data cleansing is a must as it properly organizes the data, making algorithms easier to implement and run. So, if we have a well-cleaned dataset, we can get desired results even with a very simple algorithm, which can prove very beneficial at times. Data cleaning is the process consisting of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

# **PROBLEM STATEMENT**

# Data when obtained from sources is no doubt processed but not accurate as it contains information that is not according to desired layout, redundancy requiring extra space and irrelevant data which is either mistyped or out of context. Also, sometimes the data is not completely updated.

# **PROPOSED SOLUTION**

In our project, the program will traverse the whole data-set and format all the misplaced data in correct order. The ***outdated*** information will either be ***deleted*** or ***updated***. Also, the ***duplicate*** data would be searched and removed. Hence, the entire data will be reformed completely based on optimum memory efficiency, while implementing all the required ***steps*** of ***Data-Cleansing***.

**SALIENT FEATURES**

* The ***Data*** will be collected from any website e.g., Kaggle, which contain ***data-sets*** in ***.csv*** format, and after cleansing, it will be stored in separate .***csv*** file.
* Our aim is to provide ideal solution that is able to cleanse ***any type*** of data.
* Improvement levels might be introduced e.g., level 1 improvements will be basics, level 2 will include some A.I (if/else) based decisions, etc.
* Data will be handled either in separate different structures or in single 2D abstract node structure, while unnecessary data will be removed.
* ***Outdated data*** will be ***updated*** and missing information will be replaced with an ***average value*** or organized some in another way.
* ***Data Outliers*** might be dealt if they affect the data.

**TOOLS & TECHNOLOGIES**

1. DevC++ compiler /Visual studio 2019
2. C++11 programming language
3. Operating System Microsoft Windows 10.
4. All major concepts, algorithms and terms learned in Data Structure course will be applied in best possible way.