**A picture containing shape, arrow

Description automatically generatedSTANDARDIZATION & NORMALIZATION**

Instructions:

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: DHEERAJ MISHRA**

**Batch Id: DS\_01072021**

**Topic: Data Pre-Processing**

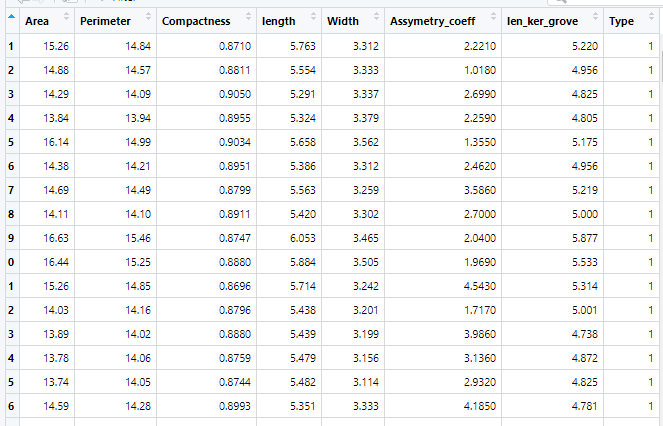
**Problem Statement:**

Data is one of the most important assets. It is often common that data is stored in distinct systems with different formats and scales. These seemingly small differences in how the data is stored can result in misinterpretations and inconsistencies in your analytics. Inconsistency can make it impossible to deliver reliable information to management for good decision making. We have the preprocessing techniques to make the data uniform. Explore the various techniques to have reliable uniform standard data, you can go through this link:

<https://360digitmg.com/mindmap-data-science>

1. Prepare the dataset by performing the preprocessing techniques, to have the standard scale to data

**A picture containing shape, arrow

Description automatically generated**

**SOLN:-**

1. **Business understanding:**

**Max:- Quality of grains**

**Min:- Old farming techniques**

**Constraints:- Natural calamity**

1. **Understanding dataset**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Feature** | **Description** | **Types** | **Relevance** |
| **Area** | **Area of grain** | **Continuous** | **Relevant** |
| **Perimeter** | **Perimeter of grain** | **Continuous** | **Relevant** |
| **Compactness** | **Compactness of grain** | **Continuous** | **Relevant** |
| **length** | **Length of grain** | **Continuous** | **Relevant** |
| **Width** | **Width of grain** | **Continuous** | **Relevant** |
| **Assymetry\_coeff** | **Symmetry of grain** | **Continuous** | **Relevant** |
| **len\_ker\_grove** | **Length of kernel grove of grain** | **Continuous** | **Relevant** |
| **Type** | **Prediction** | **Discrete** | **Relevant** |

**Hints:**

For each assignment, the solution should be submitted in the below format

1. Work on each feature to create a data dictionary as displayed in the image displayed below:
2. Refer to Seeds\_data.csv file
3. Research and perform all possible steps for obtaining solution
4. All the codes (executable programs) should execute without errors
5. Code modularization should be followed
6. Each line of code should have comments explaining the logic and why you are using that function

**A picture containing shape, arrow

Description automatically generatedGrading Guidelines:**

**Note: 1. An Assignment submission is considered complete only when successful executable code(s), and documentation explaining the applied solution and results are provided. Failing to submit either of them will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline date will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline date
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of questions asked in assignments are completed. (or)
  + All assignments were submitted, however, after the given deadline
* **A picture containing shape, arrow

  Description automatically generatedGrade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of questions asked in assignments are completed. (or)
  + Less than 80% of questions asked in assignments are submitted after the deadline
* **Grade D: (>= 60 and < 70):** Assignments submitted after the Deadline and with 50% or less of questions
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of questions asked in the assignments are submitted after the deadline (OR)
  + Less than 30% of questions asked in the assignments are submitted before deadline

**Grade F: (< 50):** Copied submission or No submission