CSC 3304

TRIAL ASSIGNMENT FOR WEEK 6

SUBMISSION DATE: 3 June 2020 ( Time: 11.30 a.m.)

**Instructions: For this trial assignment, no marks will be given but submission will be recorded.**

Questions:

1. With your own words, explain why data manipulation is an important process in machine learning experiments. What will happen if the data is not explored and manipulated before performing any machine learning experiments?

Data manipulation is an important process in machine learning experiments because machine learning does not work well without the proper data preparation process due to the fraud and invalid values in certain datasets.

1. Given the following data set:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **YEAR** | **MODEL** | **PRICE** | **MILEAGE** | **COLOUR** | **TRANSMISSION** |
| 2011 | SEL |  | 7413 | Silver | AUTO |
| 2011 | SEL | 20995 | 10926 | Green | AUTO |
| 2011 | SEL | 19995 | 7351 | Red | AUTO |
| 2011 | SEL | 17809 | 11613 | White | AUTO |
| 2012 | SE |  | 8367 | Silver | AUTO |
| 2010 | SEL | 17495 | 25125 | **₸≠≥×** | MANUAL |
| 2011 | SEL | 17000 | 27393 | **₸≠≥×** | AUTO |
| 2010 | SEL | 16995 | 21026 | **₸≠≥×** | AUTO |
| 2011 | SES | 16995 | 32655 | **₸≠≥×** | AUTO |
| 2010 | SES |  | 36116 | **₸≠≥×** | AUTO |
| 2010 | SES | 16995 |  | **₸≠≥×** | AUTO |
| 2011 | SES | 16992 | 9199 | Silver | AUTO |
| 2011 | SEL | 16950 | 9388 | Green | MANUAL |
| 2010 | SES | 16950 | 32058 | Red | AUTO |
| 2011 | SE | 16000 | 111115367 | White | AUTO |

1. Explain 4 data manipulation task that you will do for the dataset.

1. Removing duplicates: To remove any redundant or same data in a particular dataset.

2. Treating missing values: To identify any missing values and treating it by removing the column or rows, or filling the data with new value.

3. Identifying errors and outliers: To identify certain erroneous values in the datasets and changing it to a proper value or using the same method as treating missing values.

4. Scaling: To make sure the largest value in the dataset does not become an important value by using z-score method and min-max method.