**FINAL PROJECT REPORT**

**LOW RANK MATRIX COMPLETION & ENSEMBLE CORRELATION BASED LOW RANK MATRIX COMPLETION**

**Submitted By:**

Saurav Subham (104871034)

Arushi (105067234) 

**Submitted To: Dr. Roozbeh Razavi Far**

Data Mining

ELEC 8900-5-R1-2019W



University of Windsor

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| **CONTENTS** | **PAGE NO.** |
| 1.1 Data Mining |  |
| 1.2 Introduction to the Project |  |
| 1.3 Language Used: R |  |
| 1.3.1 Features of R |  |
| 1.3.2 Benefit of R |  |
| 1.3.3 R studio |  |
| 2.1 Low Rank Matrix Completion |  |
| 2.2 Correlation based Low Rank Matrix Completion |  |
| 2.3 Ensemble Learning |  |
| 2.4 Package Used: SoftImpute |  |
| 2.5 Implementation Procedure |  |
| 3.1 Complexity Analysis |  |
| 3.2 Results |  |
| 3.3 Applications |  |
| References |  |

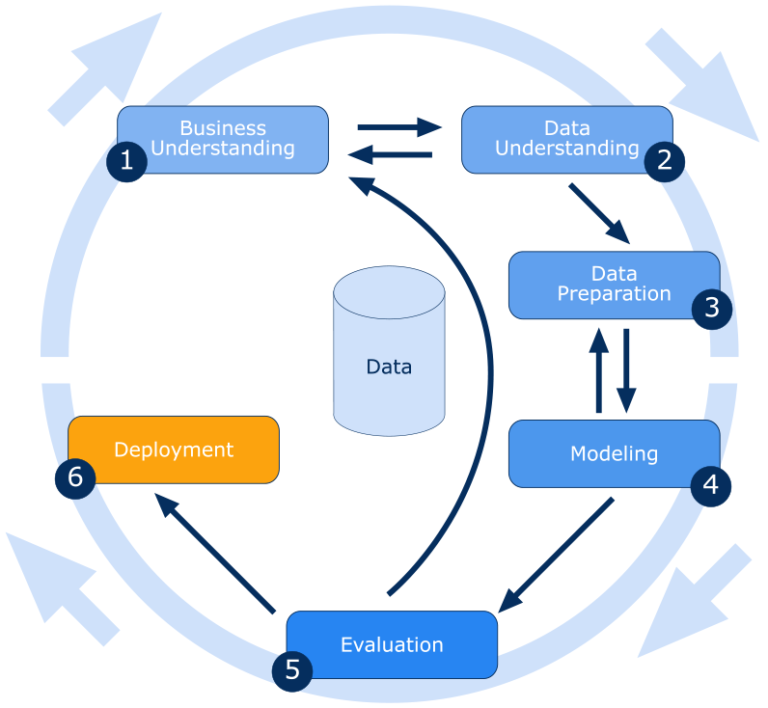
**1.1 DATA MINING**

Data mining is the process of performing various activities on data such as sorting, finding relationship among different sets of data etc. Various data mining tools are used by different companies to predict various future trends. It is also used to identify different patterns in the data sets. It goes through various steps such as data cleaning, data integration, data transformation, data mining, data evaluation.

Steps of data mining:

The various steps followed at data mining are as follows:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Data Modelling
5. Data Evaluation
6. Deployment



Techniques of data mining:

The various techniques used for data mining are as follows:

* Clustering
* Association
* Outlier detection
* Regression
* Prediction
* Classification

**1.2 INTRODUCTION TO THE PROJECT**

Every day data is being generated at a large speed. The technology is improving rapidly. There are advancements in technology almost every day. Just a simple scroll on the internet leads to a lot of data. The data generated at every stage is in huge sizes and is in varied forms. Work is done on the data in order to improve the technology. But working with data is not an easy task. It is due to multiple reasons. One reason is that the data present is in inconsistent forms. Other reason is that the data present is incomplete. Incomplete data is present due to various reasons. It could be there because people are reluctant to providing some particular piece of information pertaining to their personal life. Missing data can also be a result of the negligence of the person filling the information. Data with missing values can be hard to work with because it is a possibility that the missing values are important to work with in order to generate some important information required. It is also harmful as working with the data with missing values can also lead to the inconsistency of results. Therefore, imputing missing values of the data is of utter importance. Missing data can be of various types. They are as follows:

1. Missing completely at random: It is when the missing value is not dependent at any other value.
2. Missing at random: It is when the missing value
3. Missing not at random: It is when the missing values depends on the observed values.

Data missing completely at random and the data missing at random are ignorable whereas the data values missing not at random are not ignorable. It is so because the values that are not missing at random could have an important piece of information required for the analysis of the data.

**1.3.1 LANGUAGE USED: R**

Language R is an open source programming language that is widely used for data statistics. The environment of R follows a command line scripting and the steps followed can be saved. It is used by the those dealing with data statistics and the field of data mining for data analysis. It is even possible to do looping and branching with the help of R that has the capability of handling complex computations. R has the feature of supporting various operating systems like Linux, Microsoft, etc. It is also possible to work with offline tools such as Excel, Adobe etc.



**1.3.2 FEATURES OF R**

* R is a simple programming language which is effective and has conditions, iterative, loops etc.
* R can handle data effectively.
* R has the storage facility as well.
* R contains a series of operators that can be used to compute on data.
* R has a number of tools that can be used for data analysis.
* R also has the facilities for the representation of the data using various kinds of graphs, charts etc.

**1.3.3 BENEFITS OF R**

* R is an open source and is available for free of cost.
* R has a number of packages to help the users with the job of data analytics.

**1.3.4 R STUDIO**

R is an Integrated Development Environment. It is available for working with R language. It is user friendly and easy to work with. It is available for free on internet. Another advantage of R Studio is that it is available in two versions, desktop and web-based version. Desktop version can be used directly on the desktop while web-based version can be used via any web browser.



**2.1 LOW RANK MATRIX COMPLETION**

Low rank matrix completion refers to the activity of completing matrices by filling in the spaces that have missing values. Most of the data occurring is in the form of matrices. Therefore, it is important to have a technique that could deal with the missing values of the matrices. For example, Netflix uses the matrices in order to store the ratings [2]. The values of the ratings in the matrices are generally considered inconsistent since some users are not attentive while rating a show or movie at Netflix and some skip the rating while watching content on Netflix. It is important to deal with such kind of datasets as using these inconsistent matrices can result in the loss of accuracy of results obtained by working on such a matrix.

The algorithm of low rank matrix completion takes while dataset into consideration while ignoring the subtle differences that might be in the datasets. It focuses on the dataset as a whole.

**2.2 CORELATION BASED LOW RANK MATRIX COMPLETION**

In this algorithm, correlation between data sets is computed to find the relation between them. It is done to establish the relationship between the data elements. This algorithm is the advanced version of the low rank matrix completion. The major difference between low rank matrix algorithm and correlation based low rank matrix completion is that the latter takes subtle differences between data into consideration whereas low rank matrix completion ignores it.

**2.3 ENSEMBLE LEARNING**

Ensemble learning is the concept of machine learning where a number of learners are employed on a single job to get the best output. It is constructed in a way so as to get the best result with the help of algorithms used in the field of machine learning. Most of the algorithms of machine learning aim to generate a result using just a learner whereas the basic concept used by ensemble learning is that a large number of learners are employed that are trained to work on the same problem. The advantage of this technique is that since so many workers are working on the same problem working together to assume the values, the probability of getting the right value increases. The learners working on the problem are known as base learners. Generalization is used to combine the results of the base learners and get the final result. Most of the algorithms of the ensemble learning contain homogeneous base learners that use a single algorithm.

**2.4 PACKAGE USED: SOFTIMPUTE**

SoftImpute is a package used in the language R to complete the incomplete matrices. SoftImpute has the ability to deal with all types of matrices, be it large or smaller ones. Since we were dealing with the matrices in the project, therefore we used SoftImpute in the imputation of missing values. It considers the incomplete dataset as matrices and complete it.

**2.5 IMPLEMENTATION PROCEDURE**

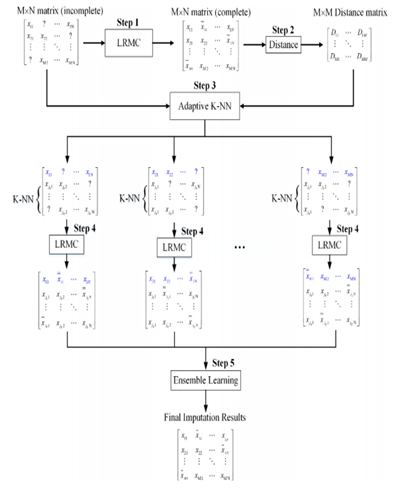
Step 1: Low rank matrix completion is applied to get the estimated values of the matrix.

Step 2: Then a distance matrix containing pairwise distance between the estimated values is calculated to get a correlation between the estimated values.

Step 3: For a specific sample, its most similar samples based on the distance matrix constructed are chosen by using Adaptive K-NN.

Step 4: CLRMC is applied on these with much stronger correlation to obtain refined estimation of missing values.

Step 5: Ensemble learning is then applied using a lot of learners, and then the values are integrated in order to get better values for the results of imputations.



**Flow of the algorithm**

**3.1 COMPLEXITY ANANLYSIS**

* The time complexity of K-nearest neighbour algorithm is

O(mn) for each sample.

* The time complexity of conventional LRMC performed on the whole data is

O (t {mn², m²n}).

* The time complexity of CLRMC with K+1 sample is O (t{(K+1) d², (K+1) Cd}).
* Then, the total time complexity of the proposed method can be estimated as O(m²d + mt{(K+1) n², (K+1) ²n})
* We assume the total number of samples is m, the dimensionality is n, t denotes the number of iterations and the number of nearest neighbours is K, where m >>K.

**3.2 APPLICATIONS**

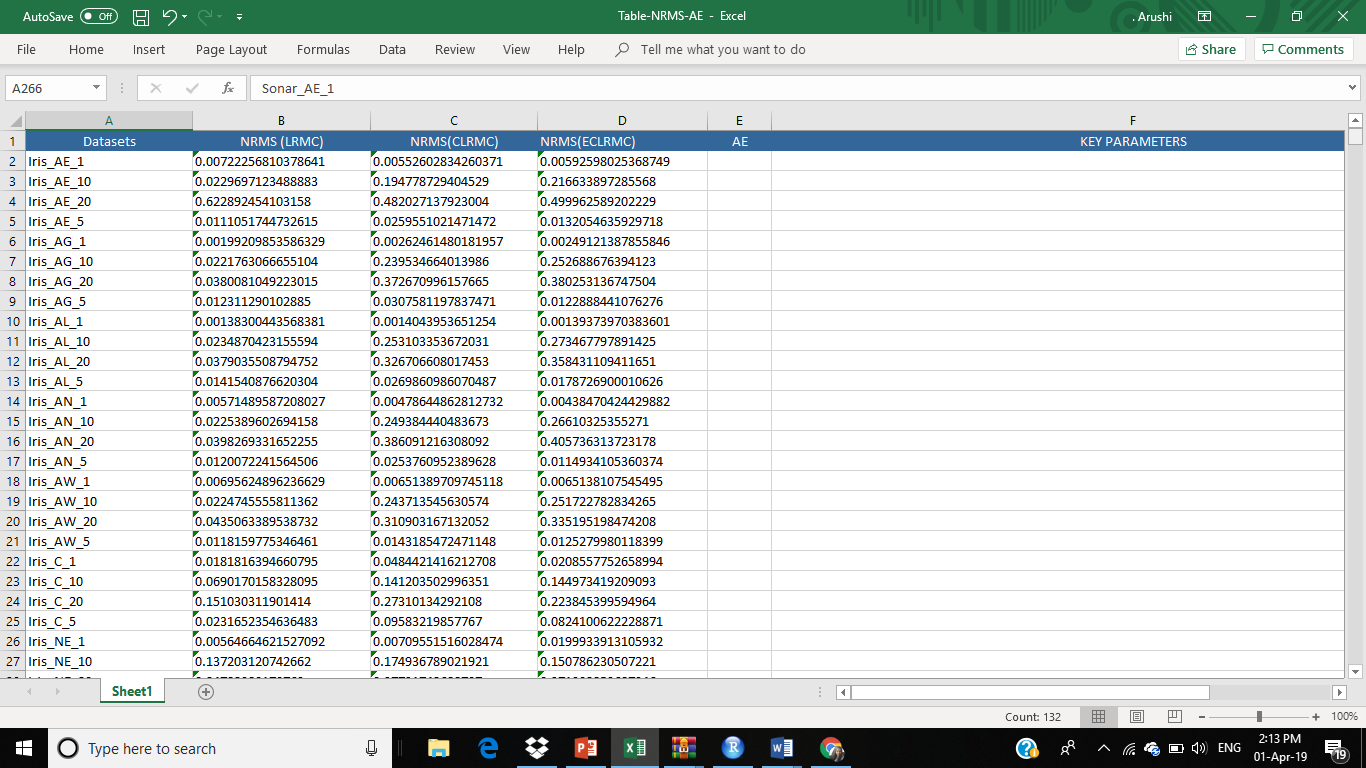
Some successful applications of LRMC include:

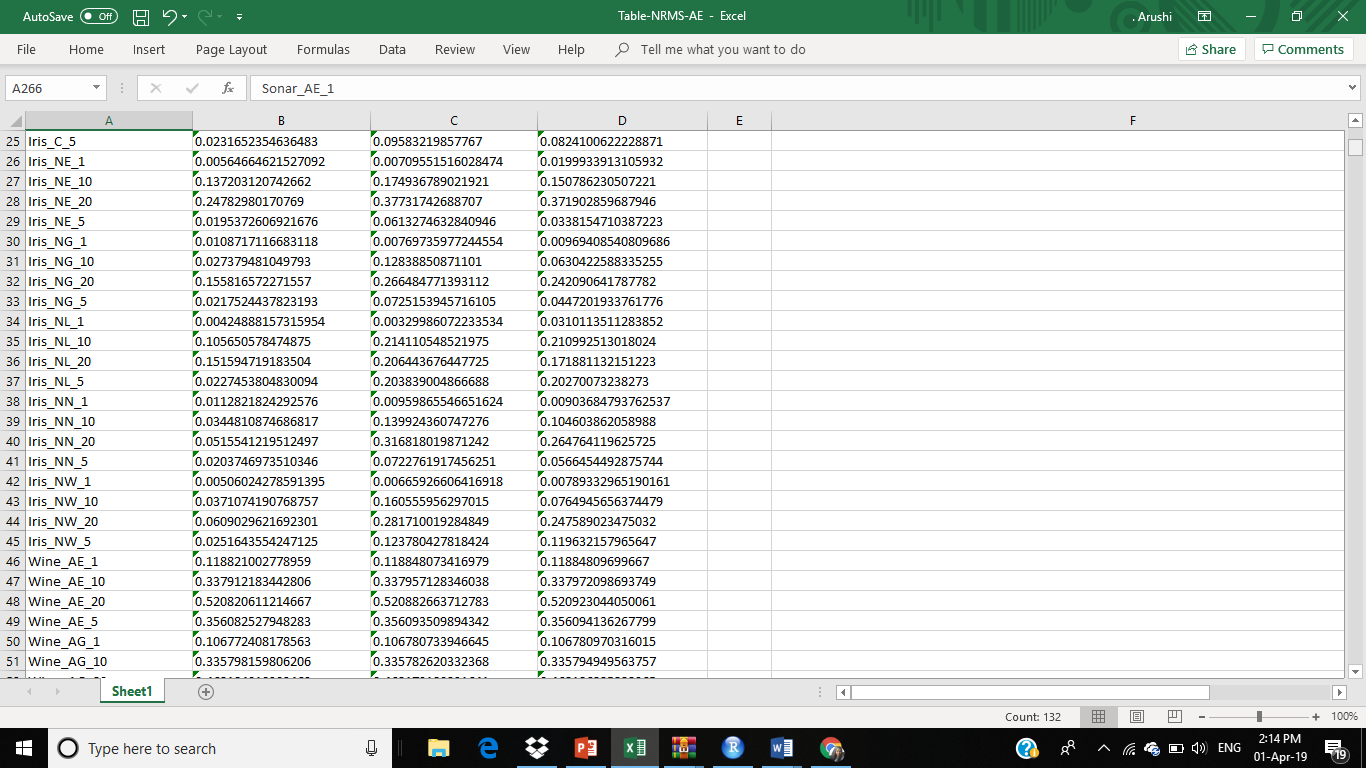
1. Collaborative filtering - Recommender systems (Netflix, Spotify, Amazon...)
2. Data Science
3. Model reduction
4. Environmental monitoring of soil and crop conditions, water contamination, and air pollution, also sensor calibration.
5. Source localization and target tracking in radar and sonar.
6. Imaging: denoising, reconstruction in medical, hyper-spectral imaging.
7. Anomaly detection in network flows. [2]

**3.2 RESULTS**

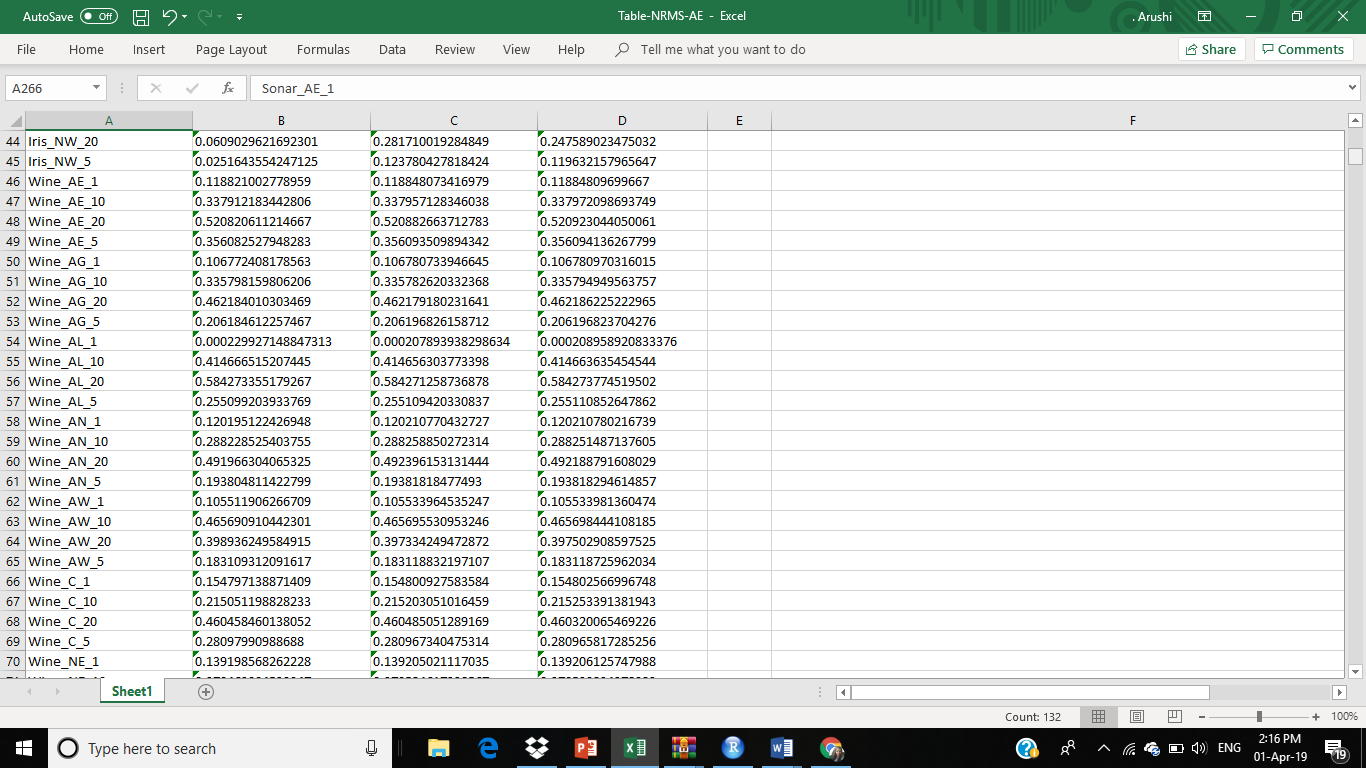
The project we are working on computes three different kinds of imputed values. One is the values obtained after applying the algorithm of low rank matrix completion. The second table of imputed values obtained after applying correlation based low rank matrix completion. The third set of values obtained is after applying the algorithm of ensemble learning. Corresponding to the three sets of imputed values, the three sets of NRMS values have been created. Following are the screenshots of the NRMS values obtained.

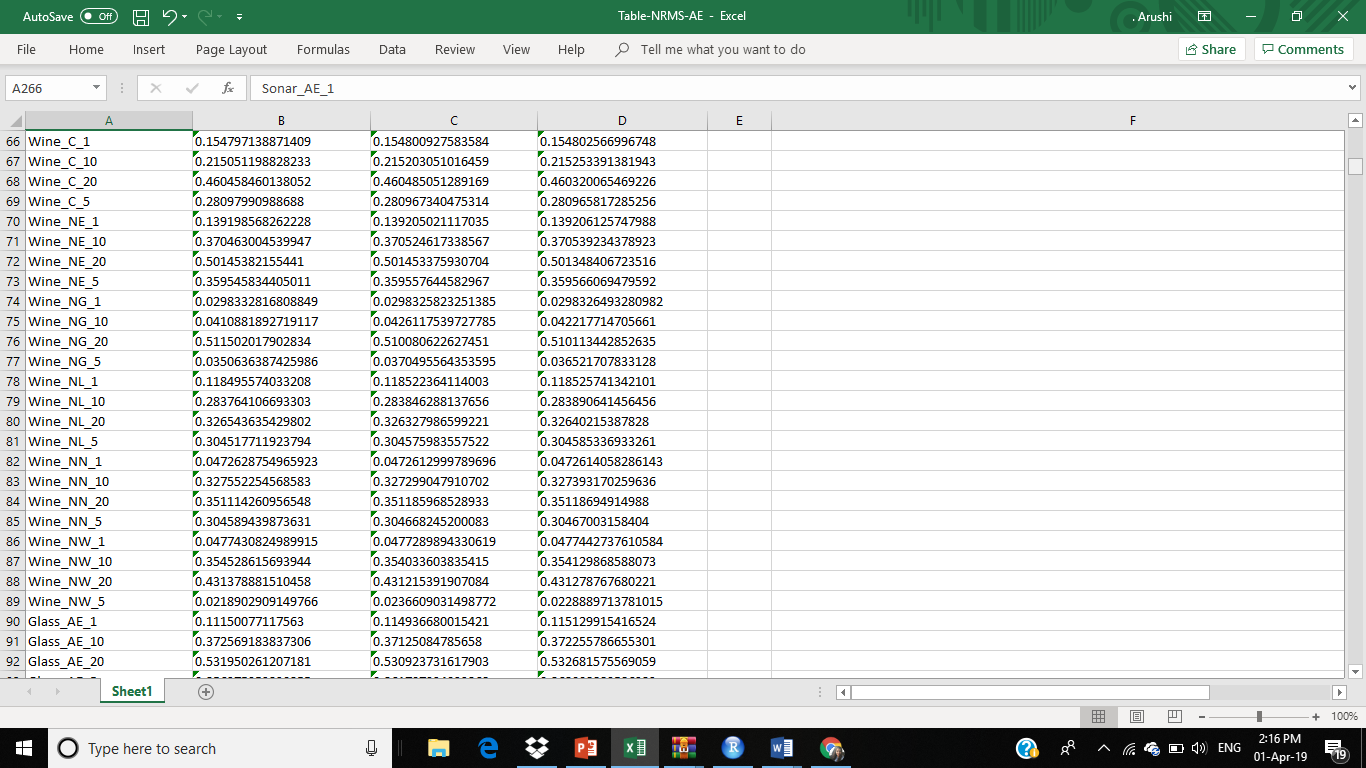
**DATASET: IRIS**



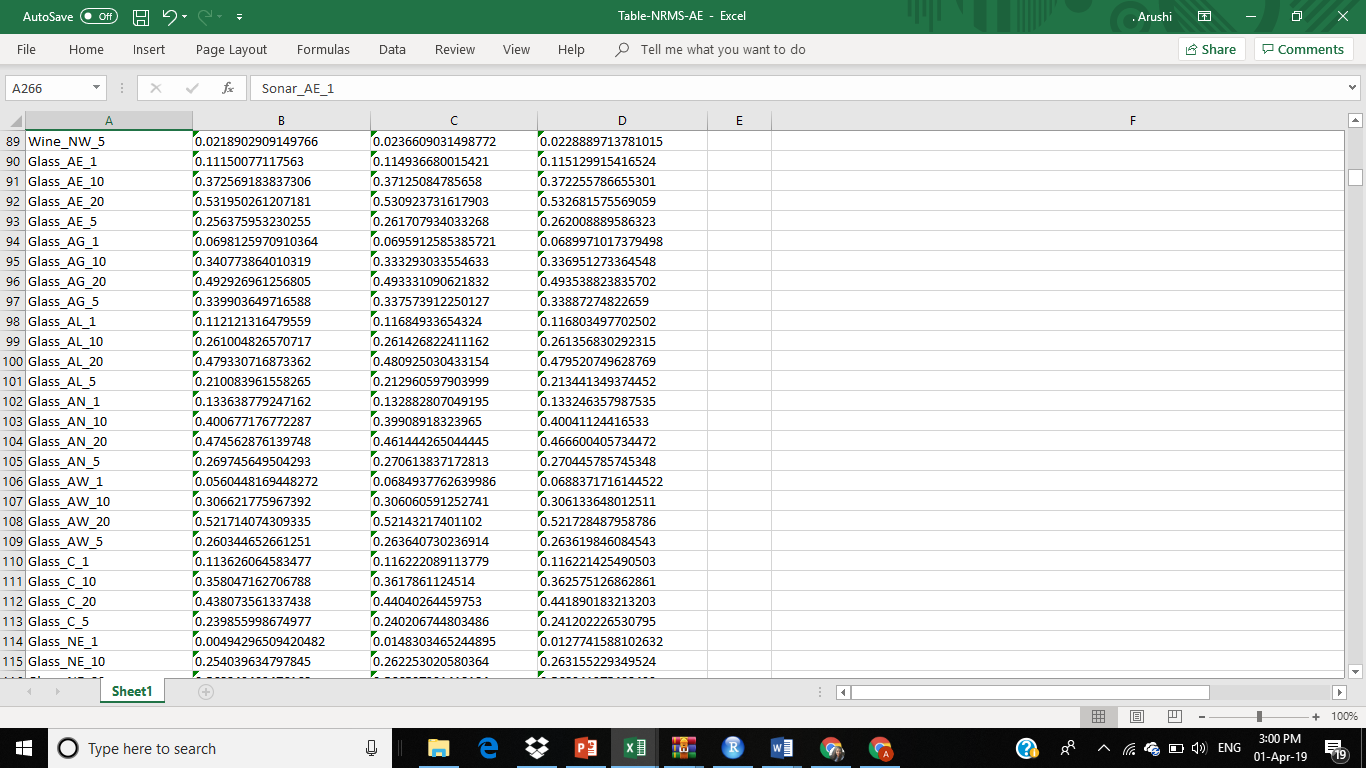


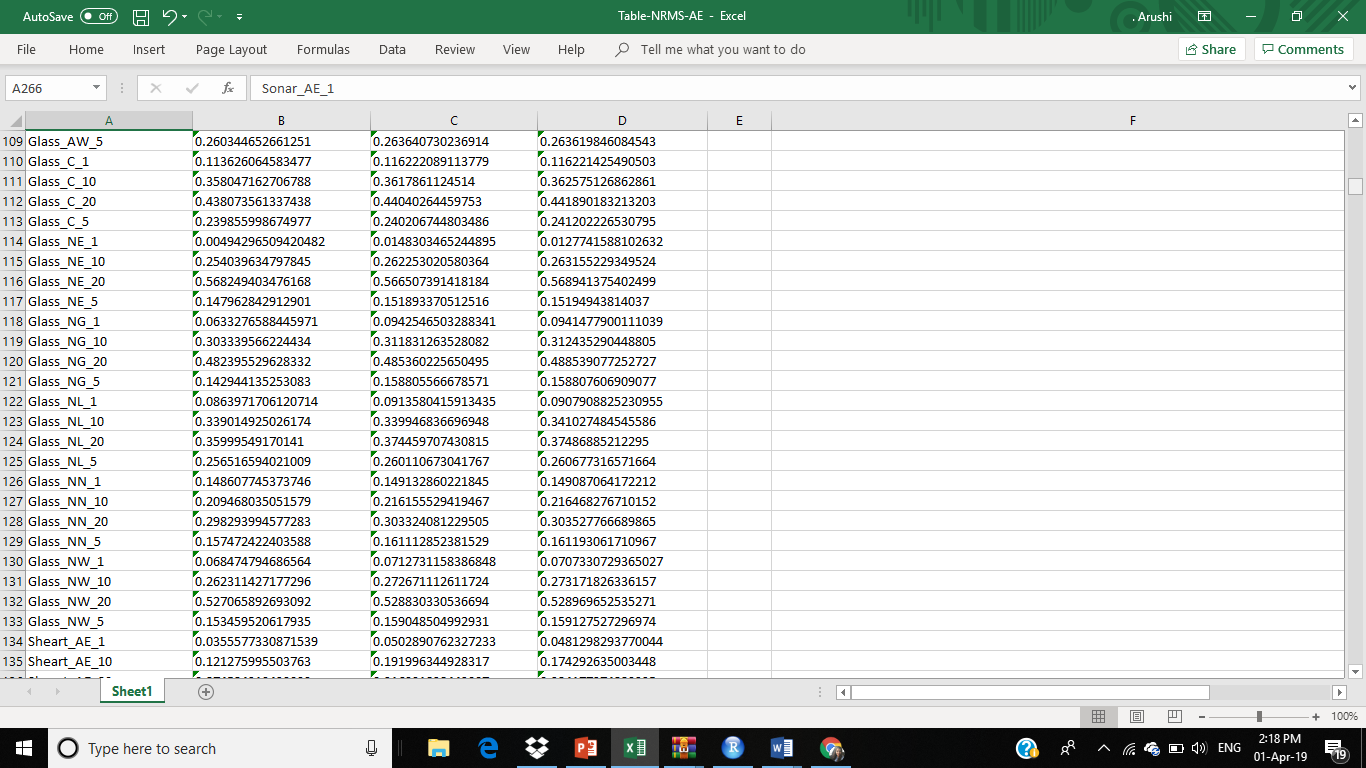
**DATASET: WINE**



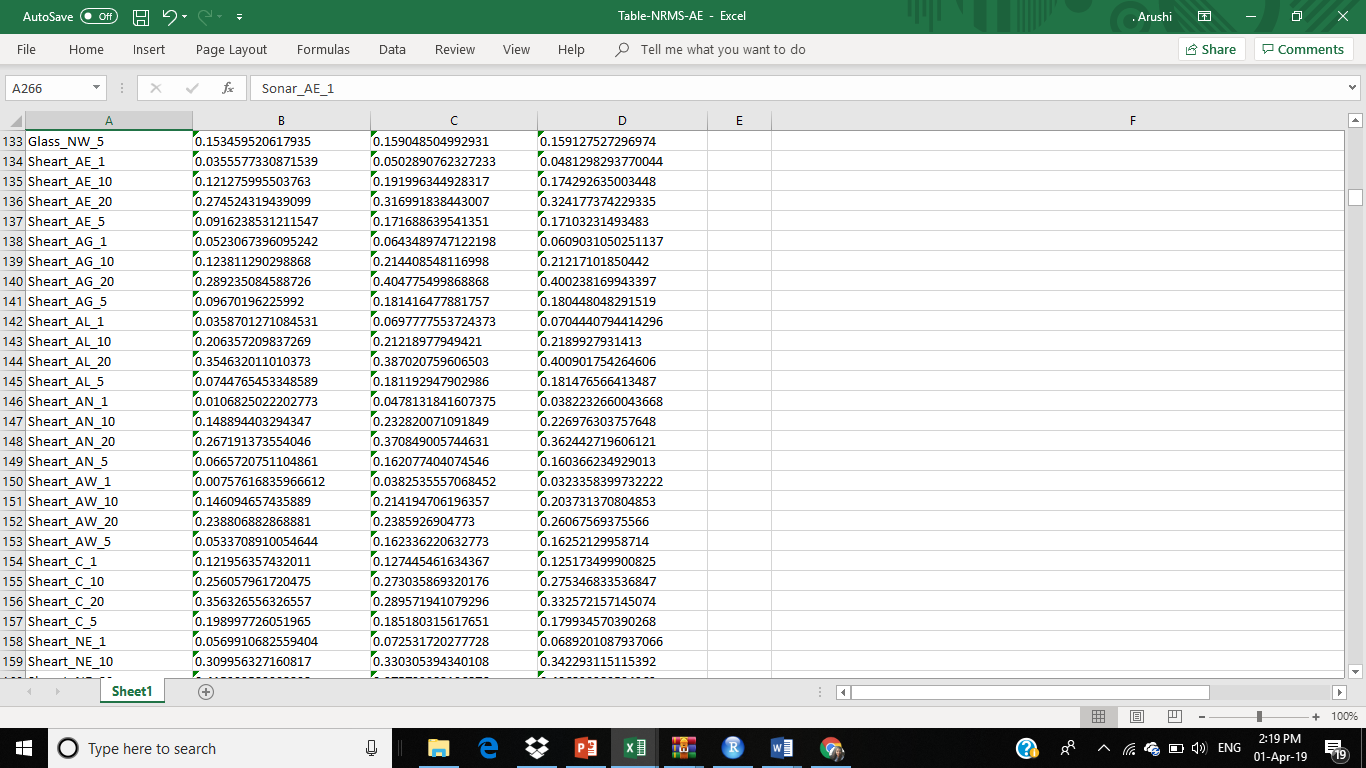


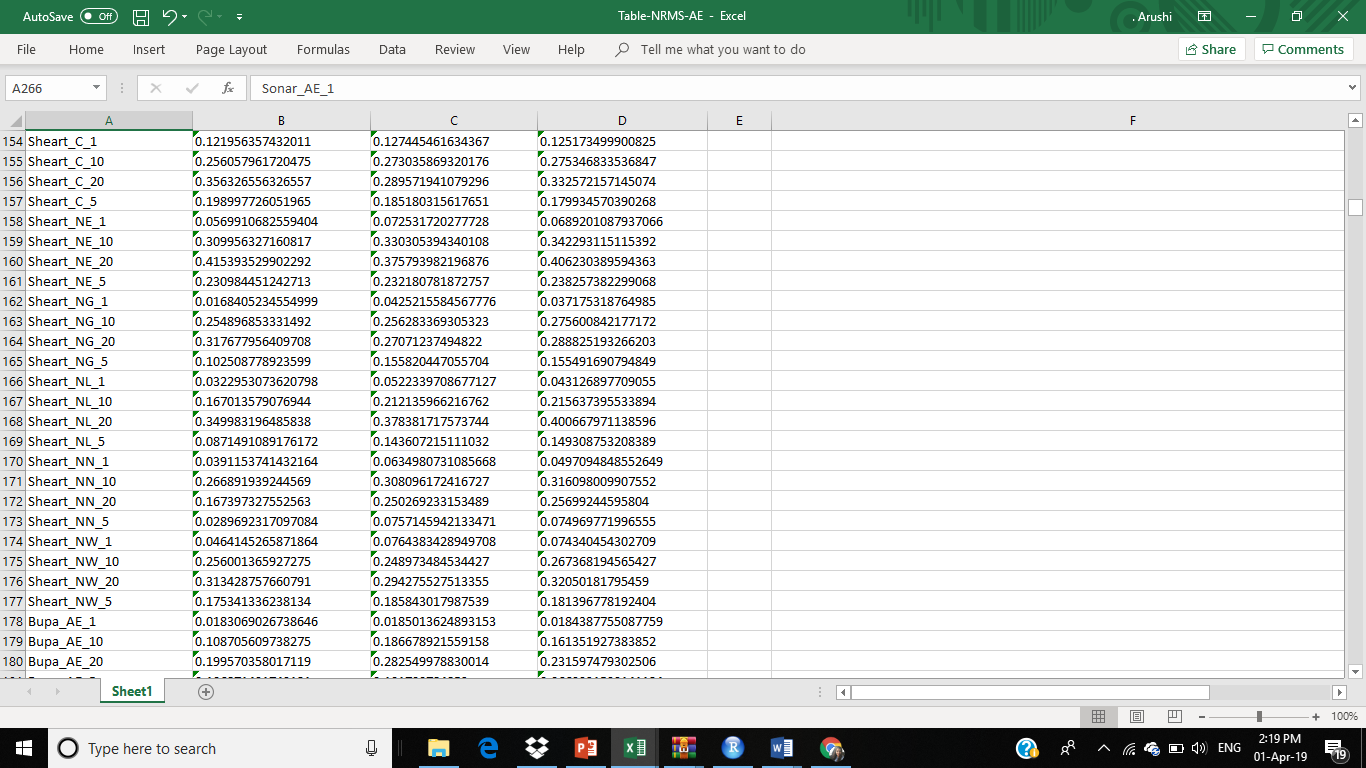
**DATASET: GLASS**



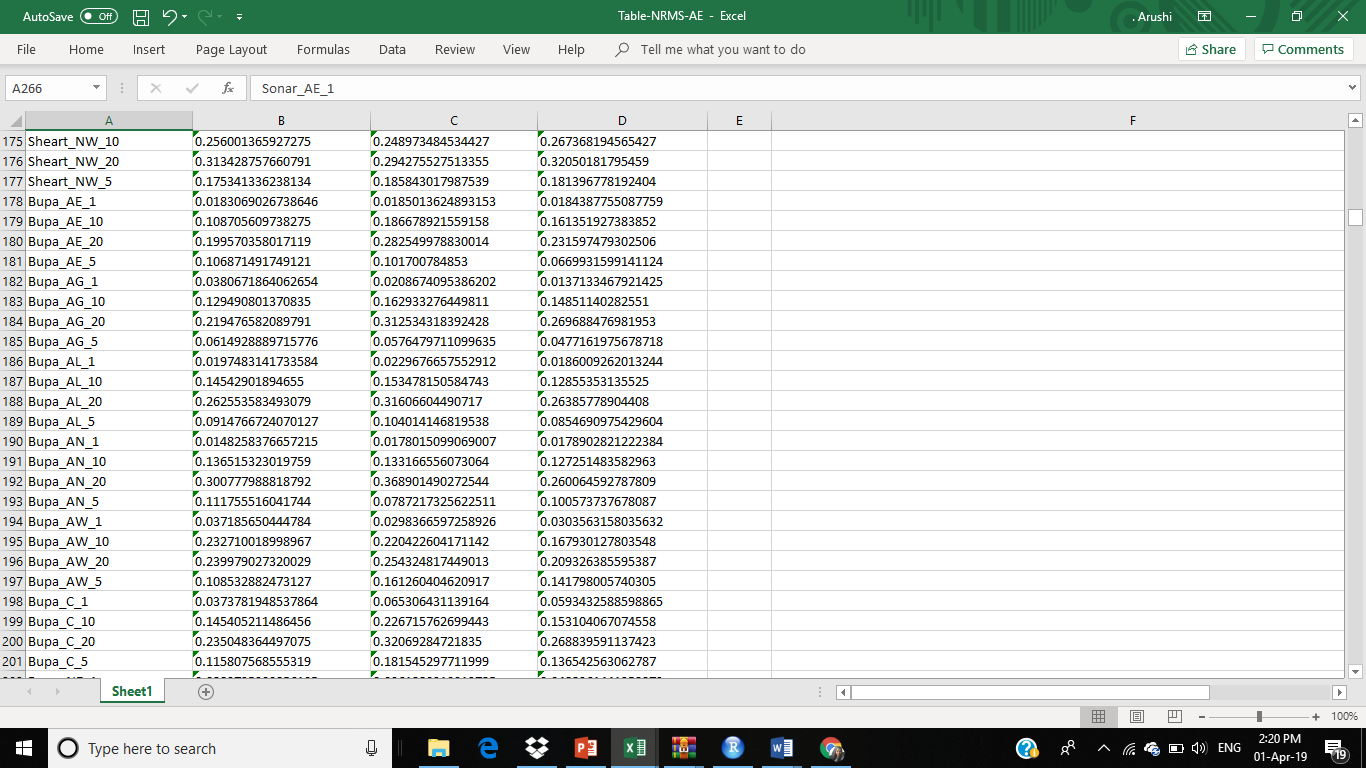


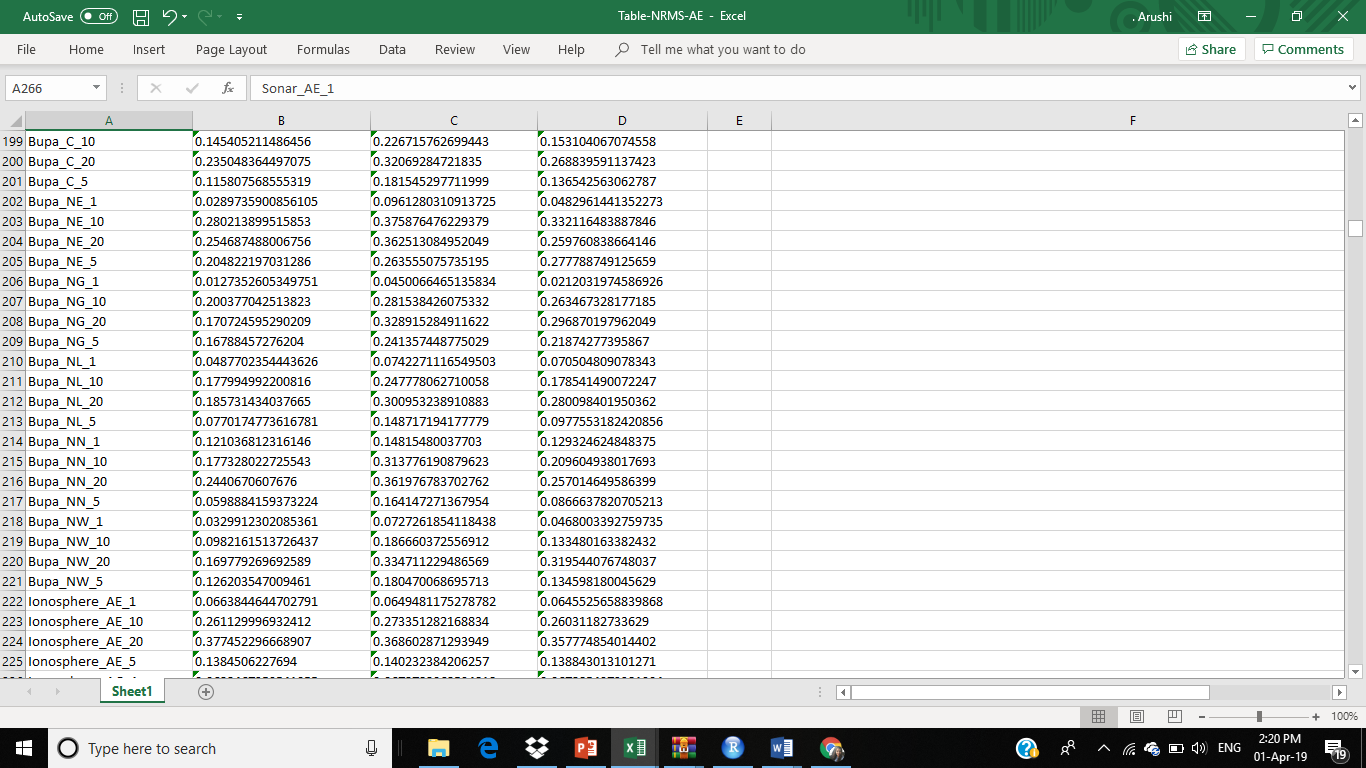
**DATASET: STATLOG HEART**



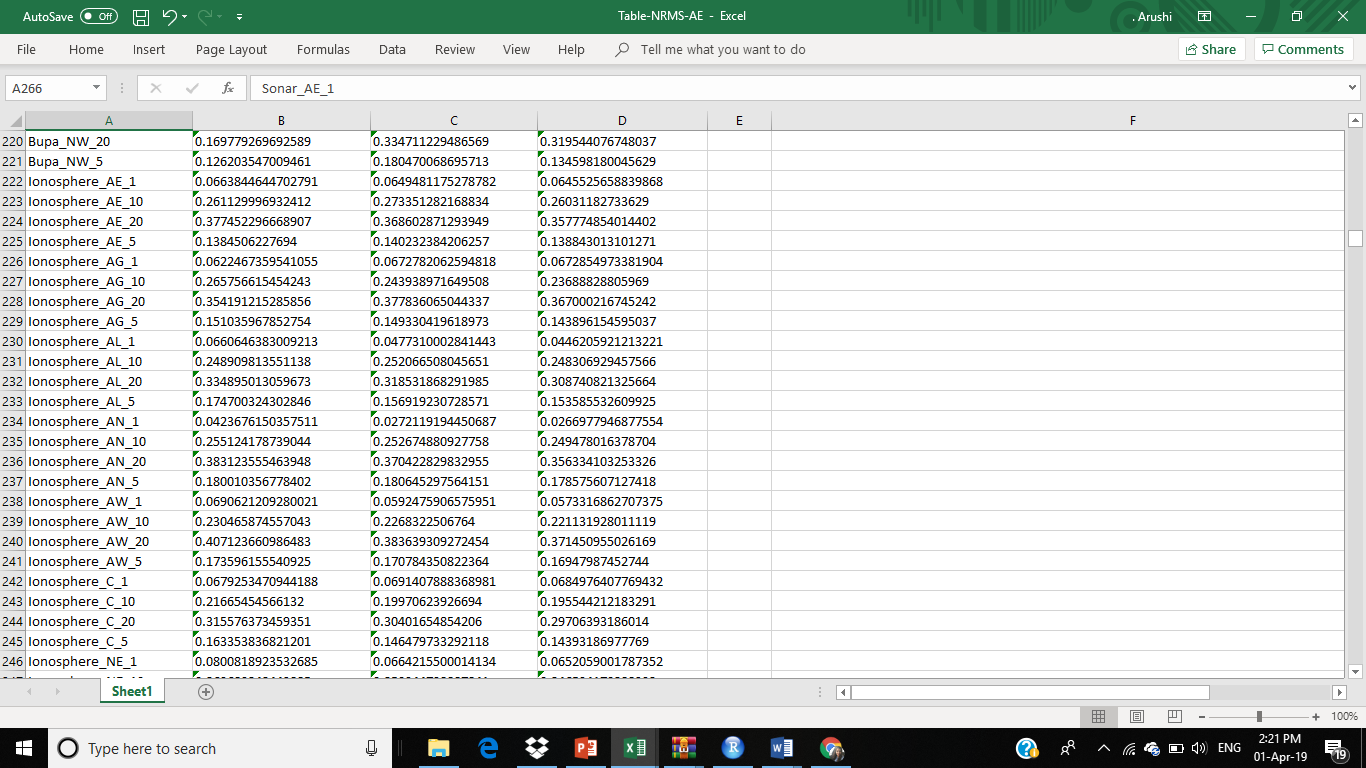


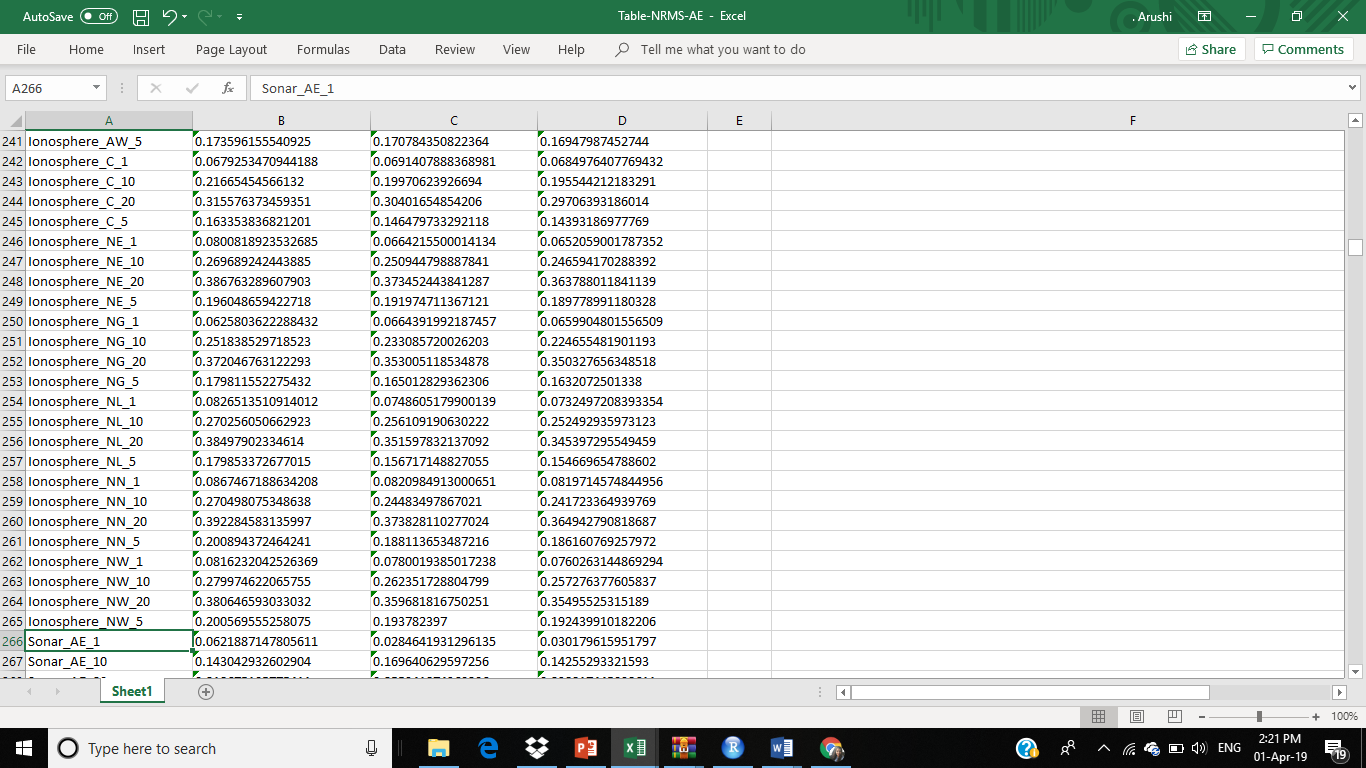
**DATASET: BUPA LIVER DISORDERS**



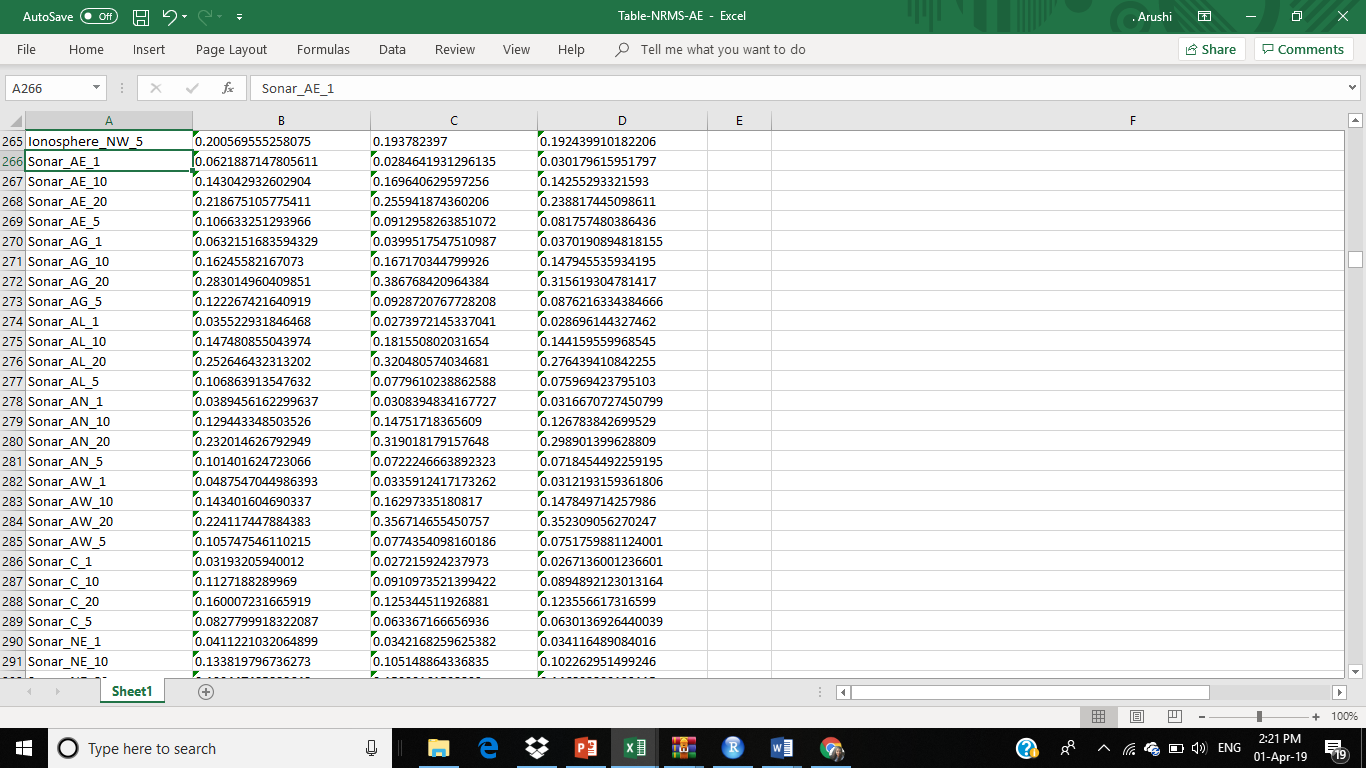


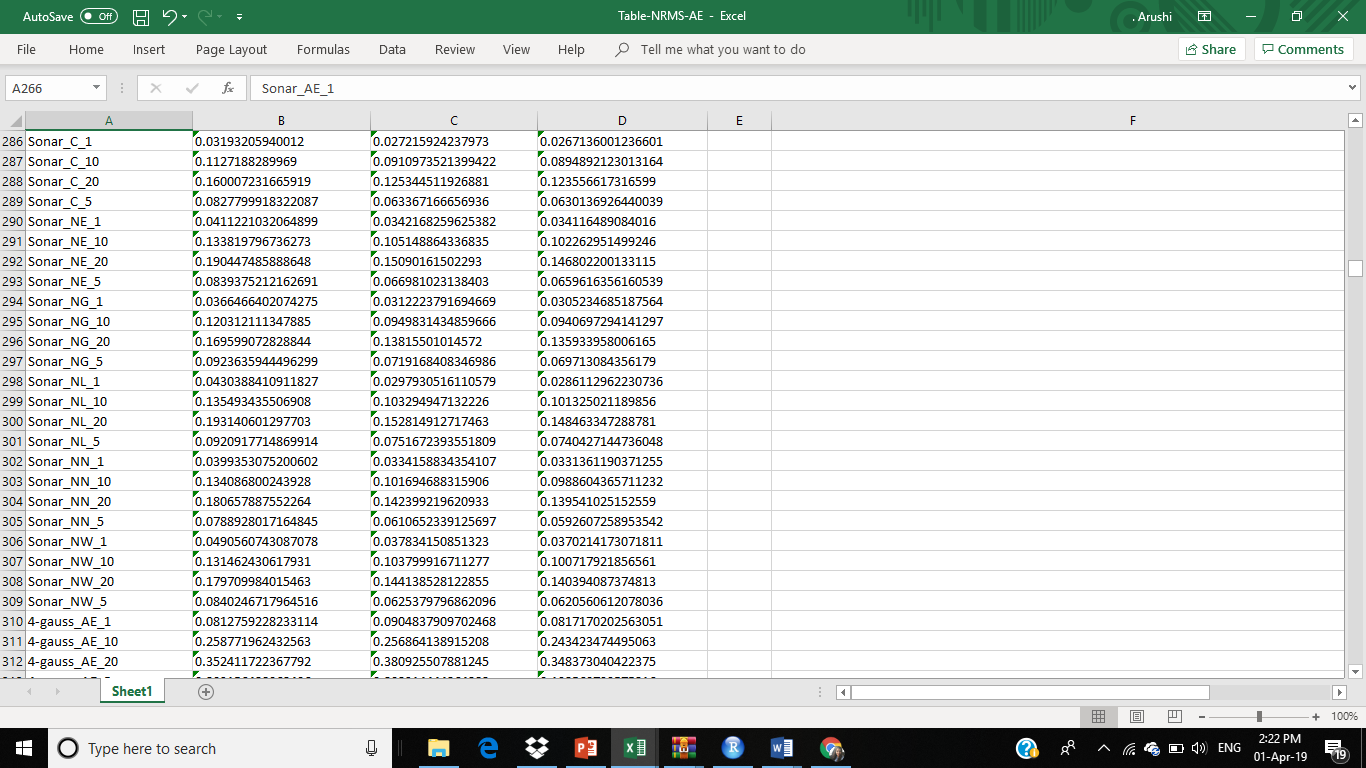
**DATASET: IONOSPHERE**



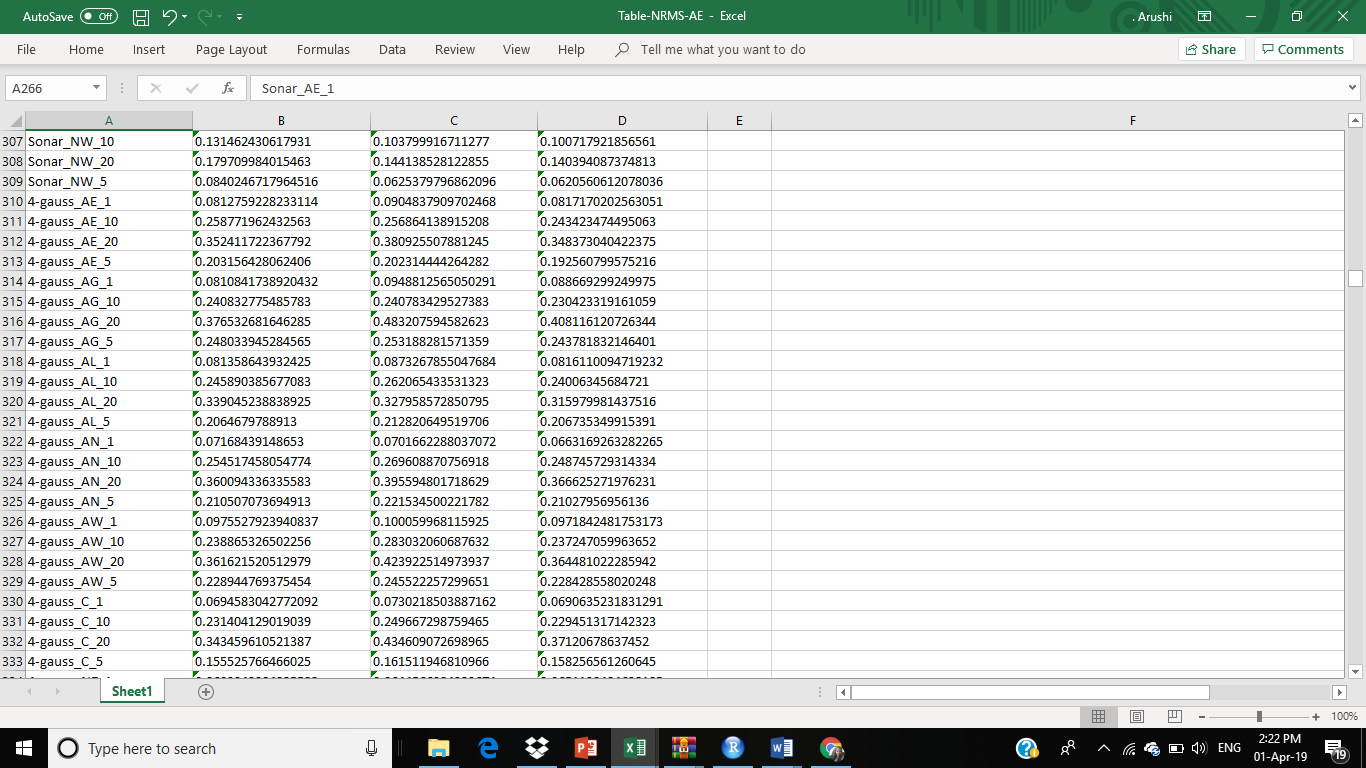


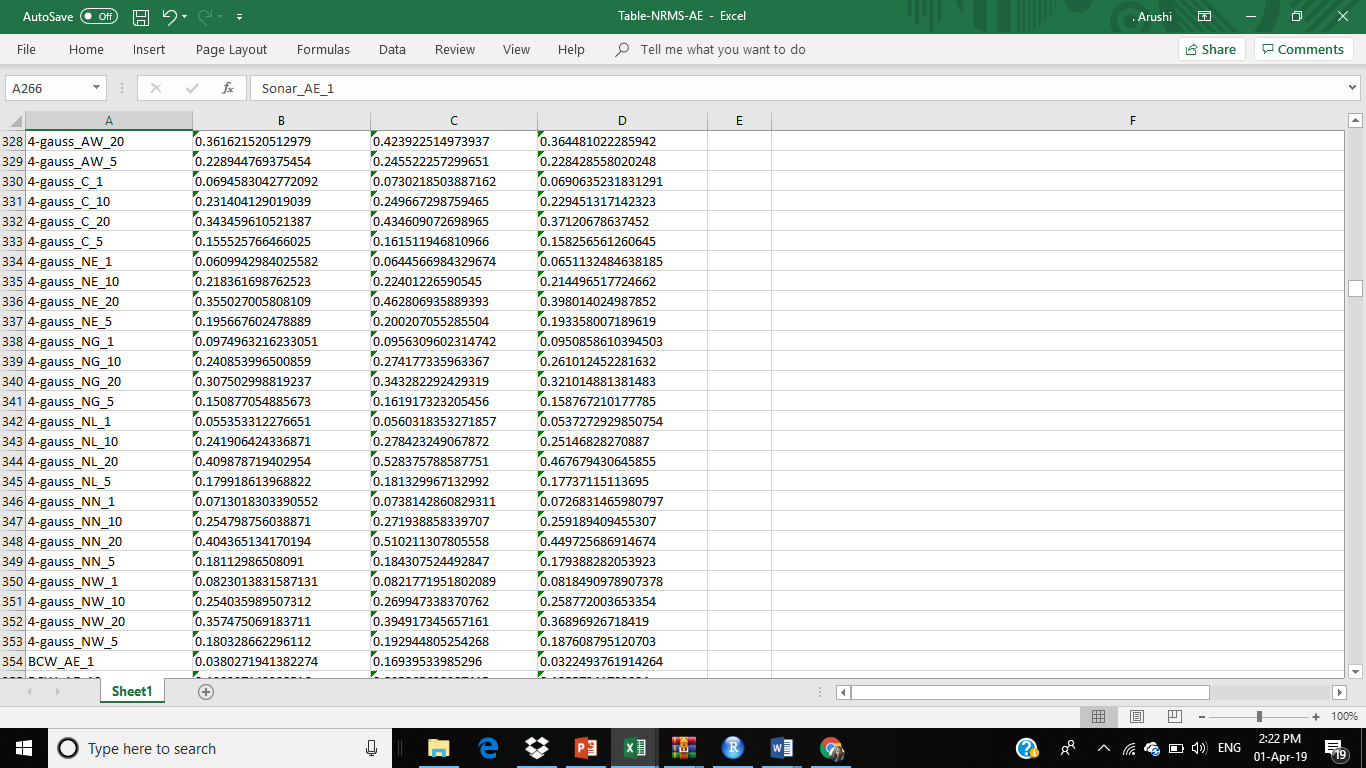
**DATASET: SONAR**



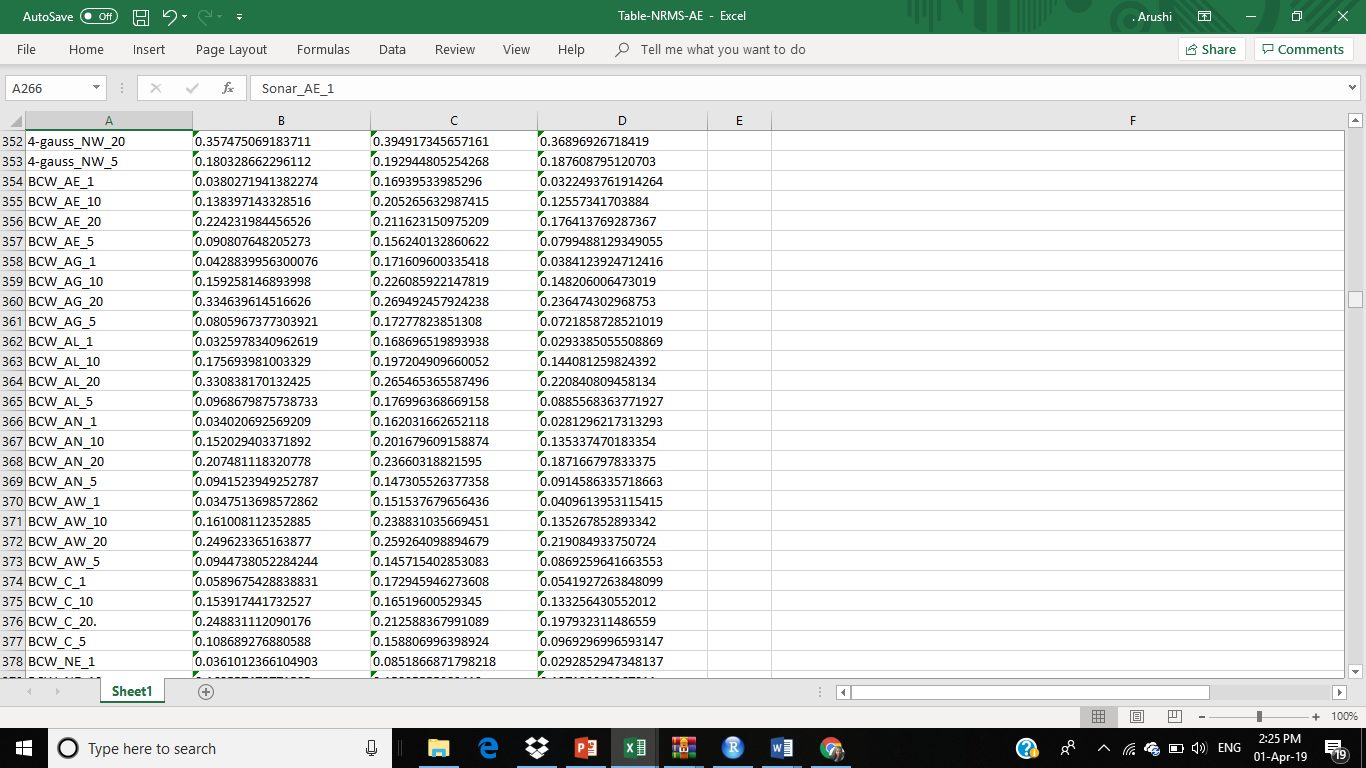


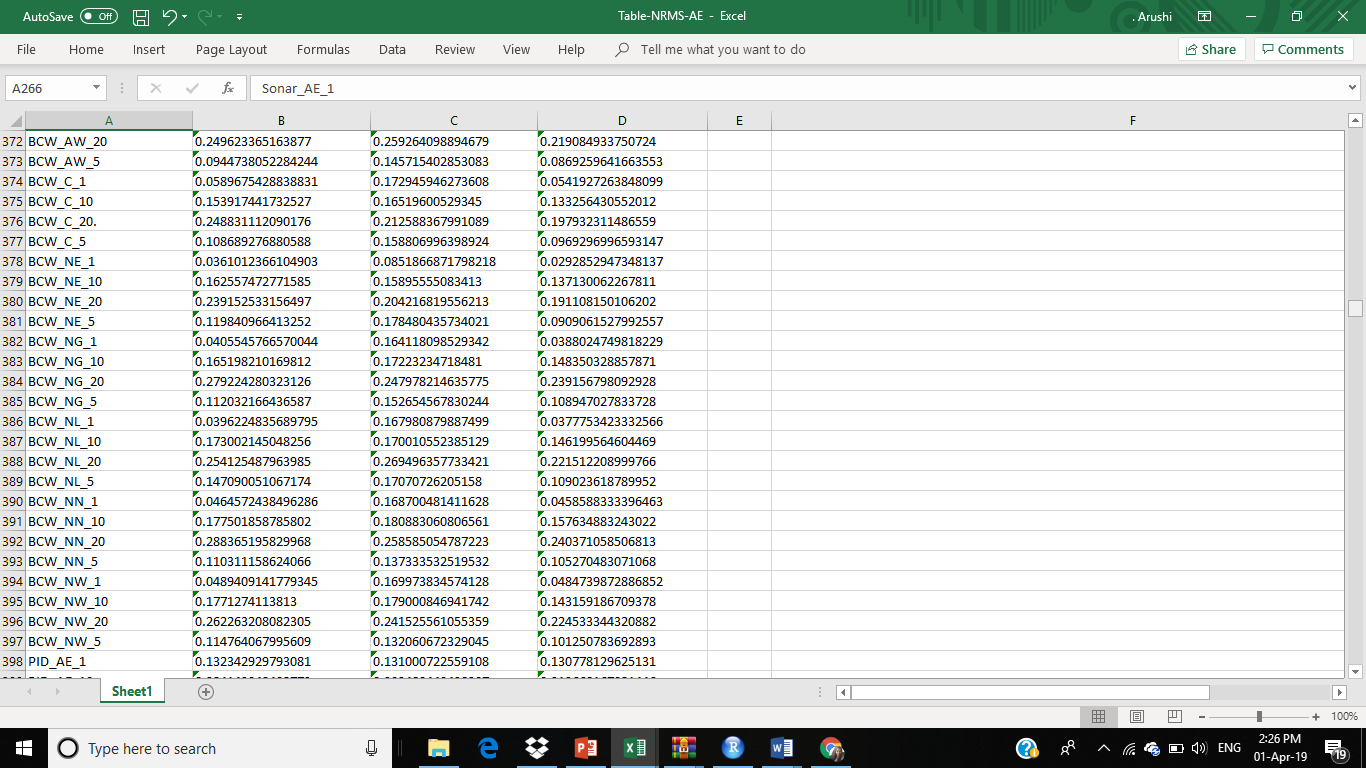
**DATASET: 4-GAUSS**



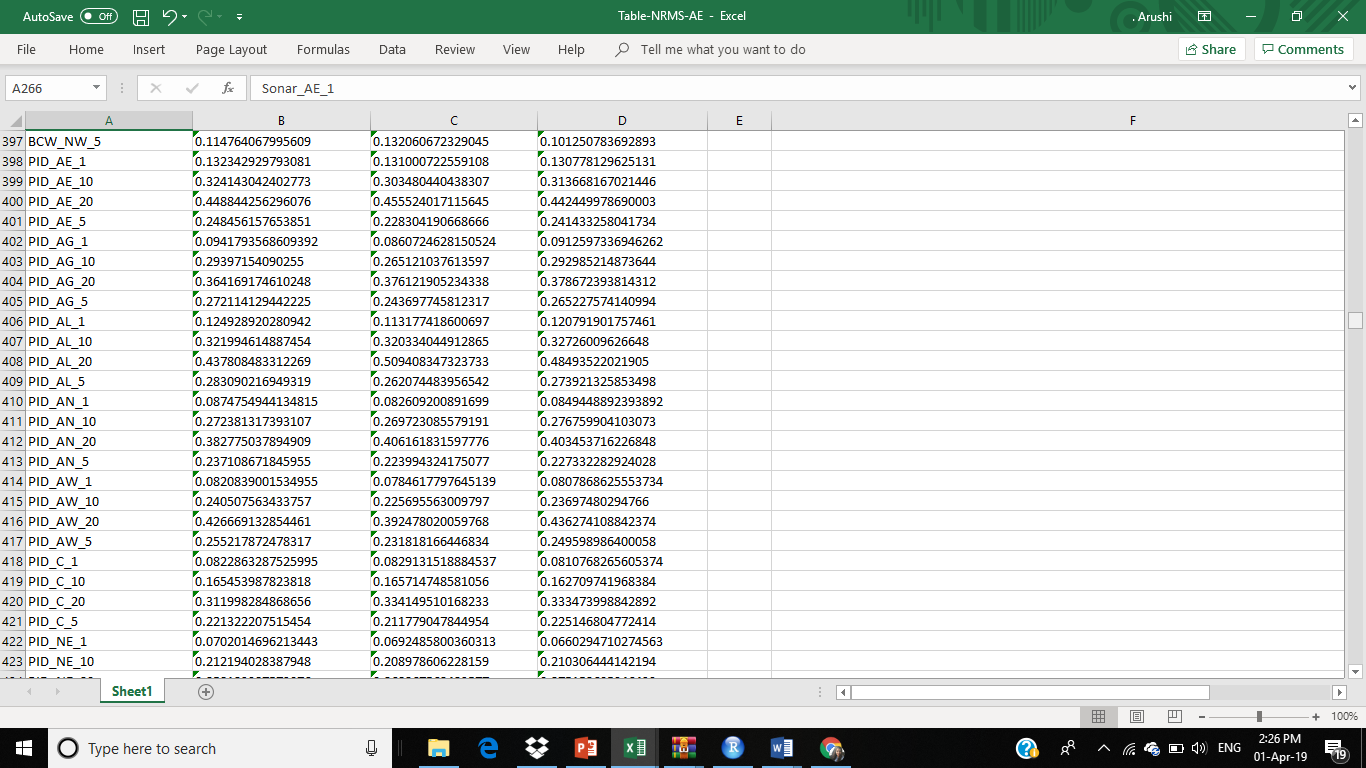


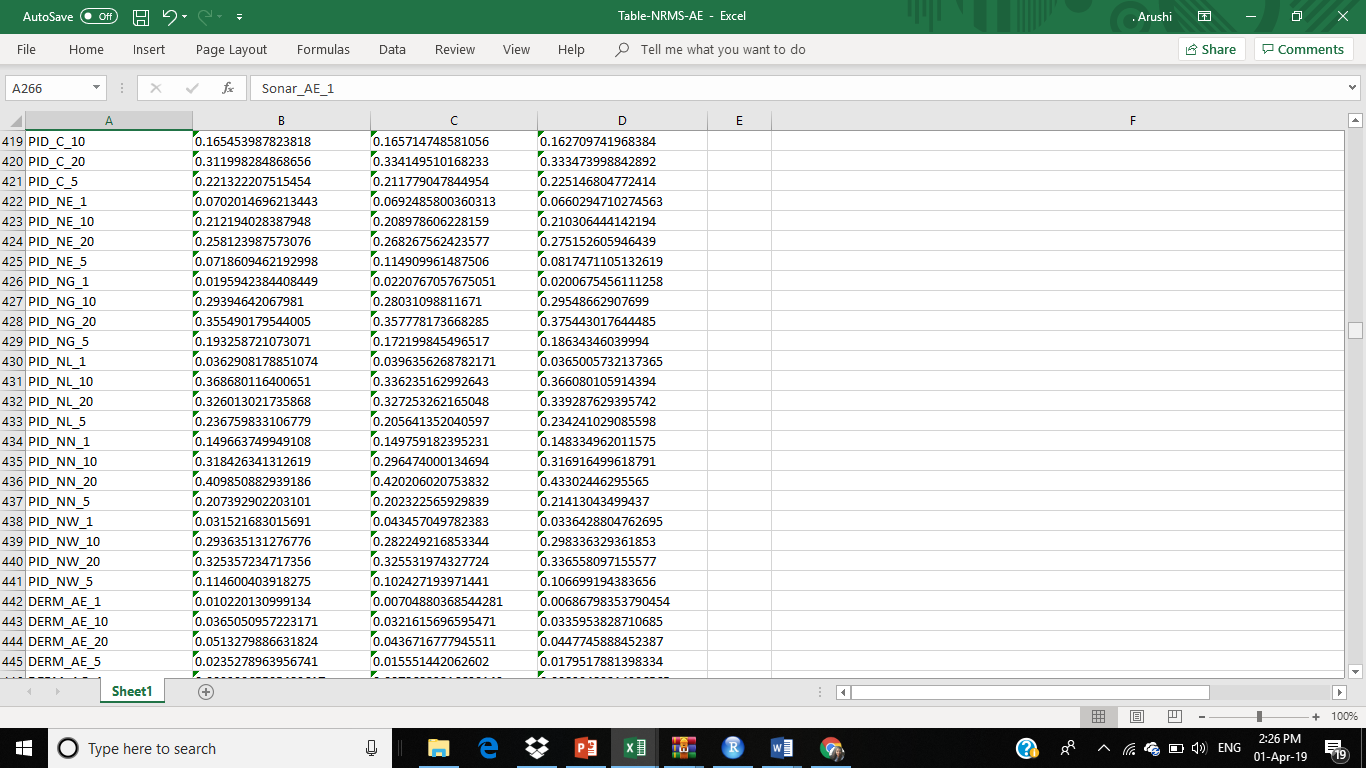
**DATASET: BCW**



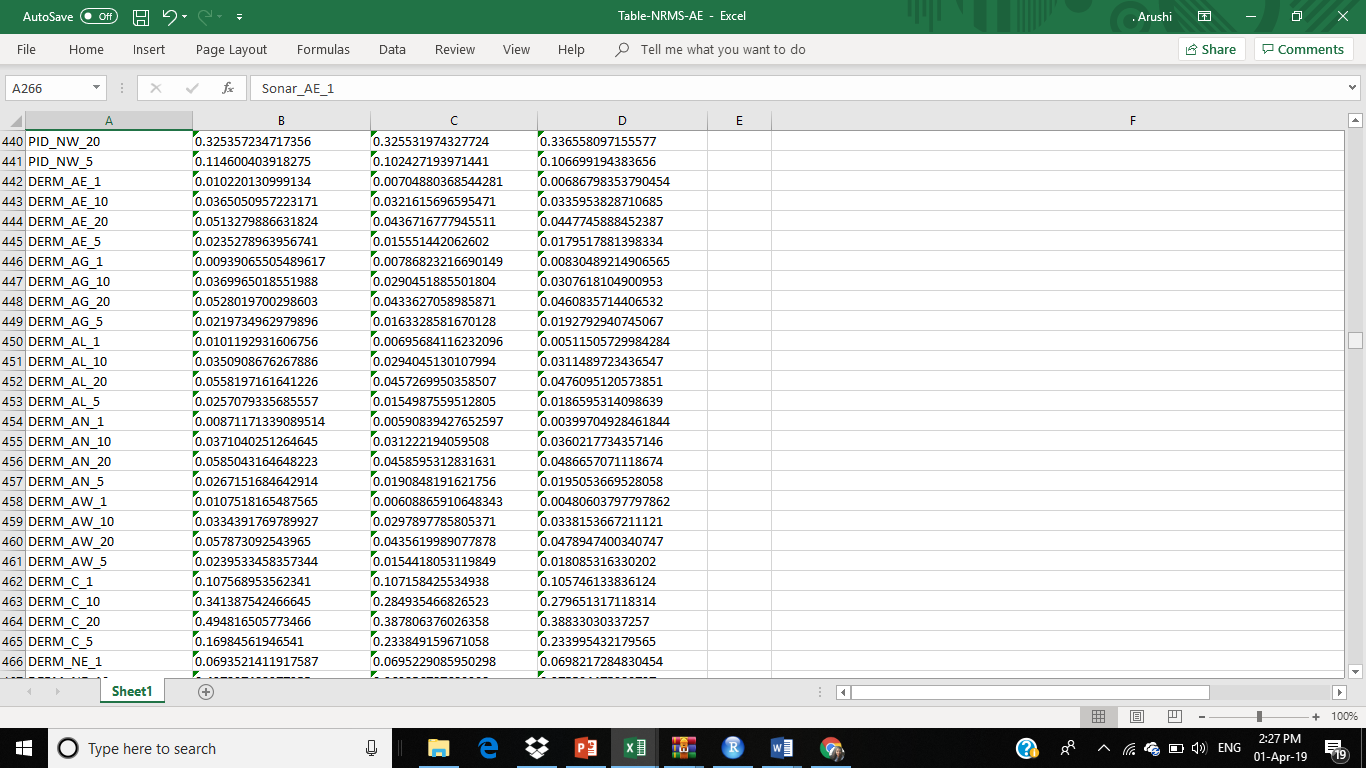


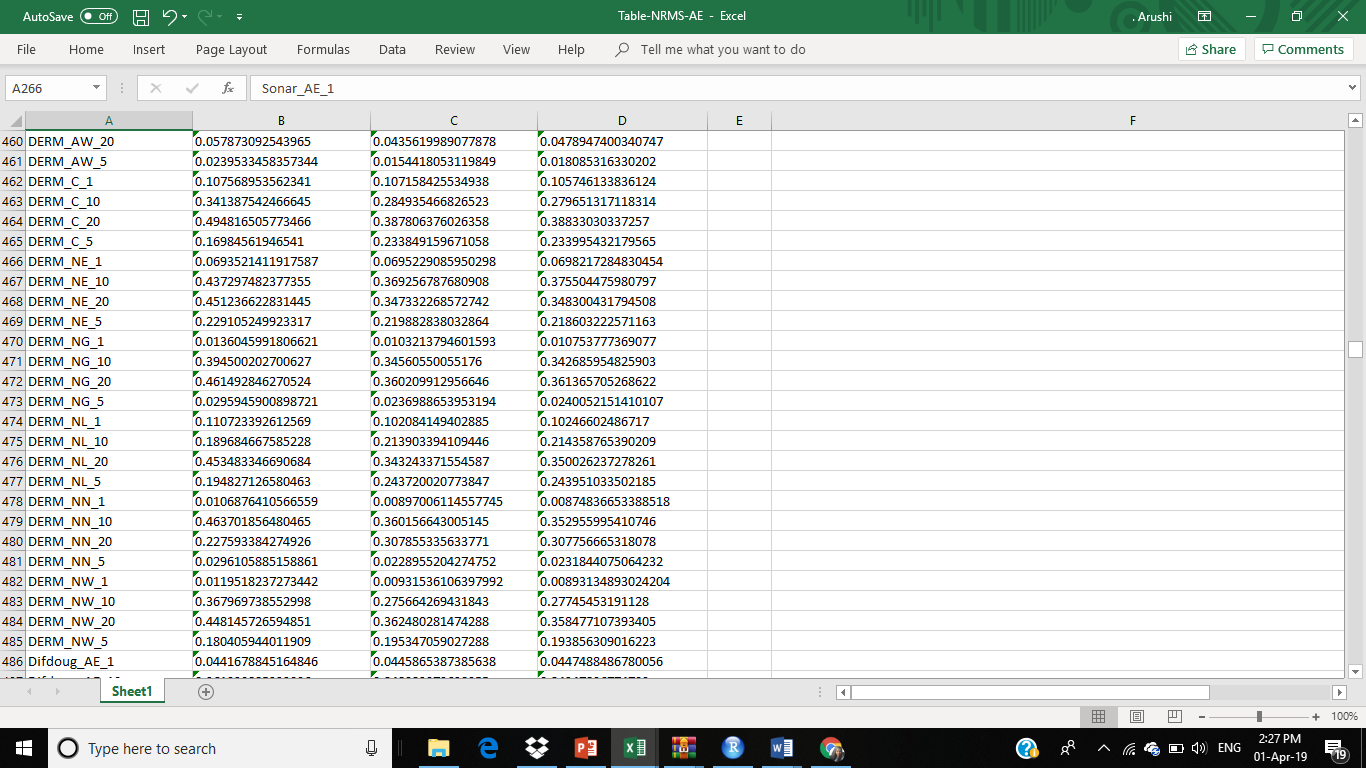
**DATASET: PID**



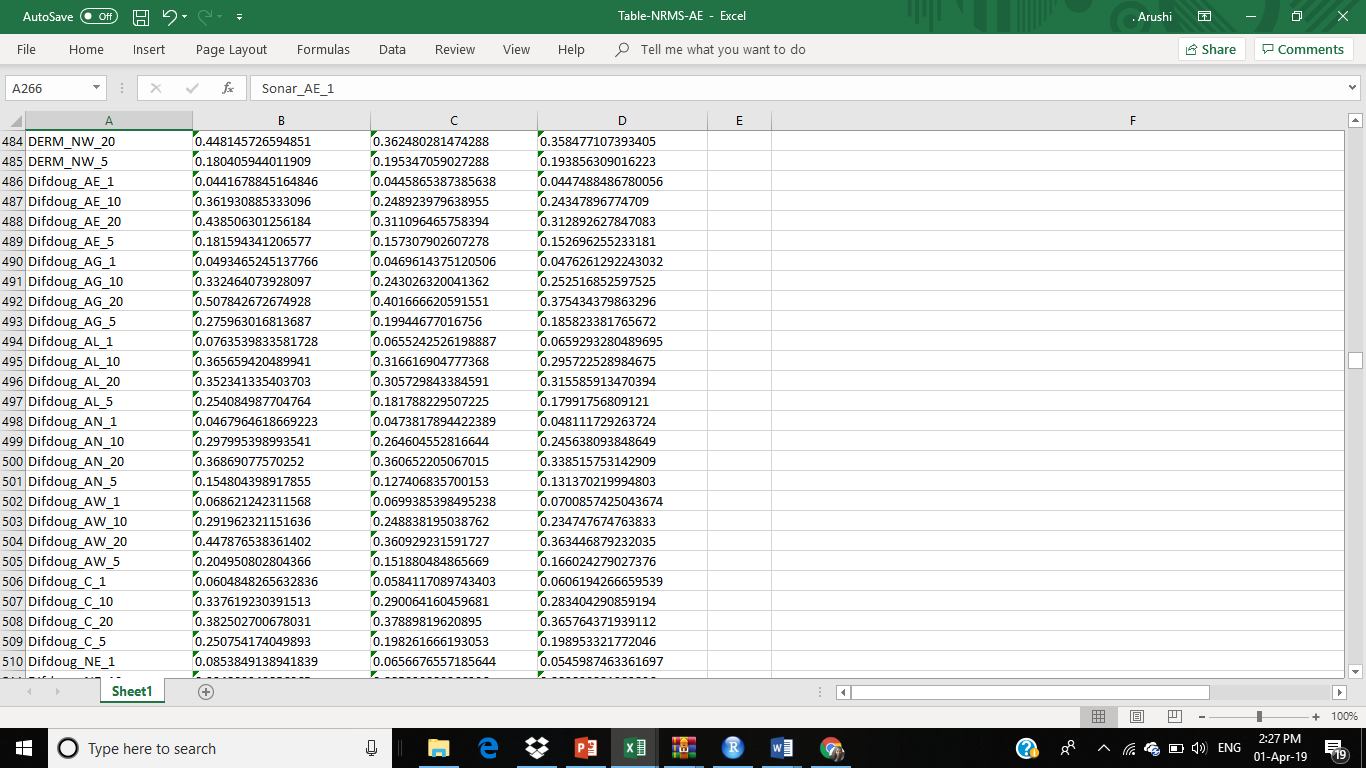


**DATASET: DERM**



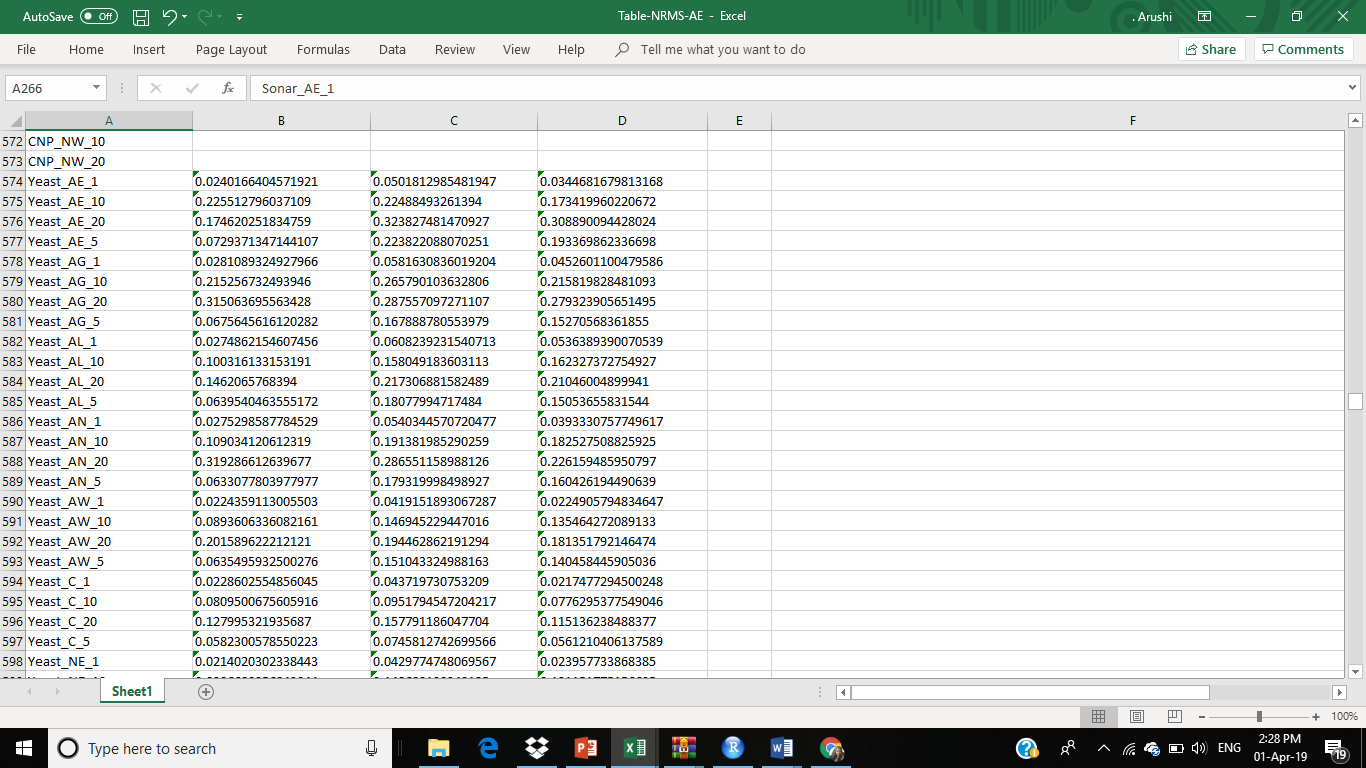


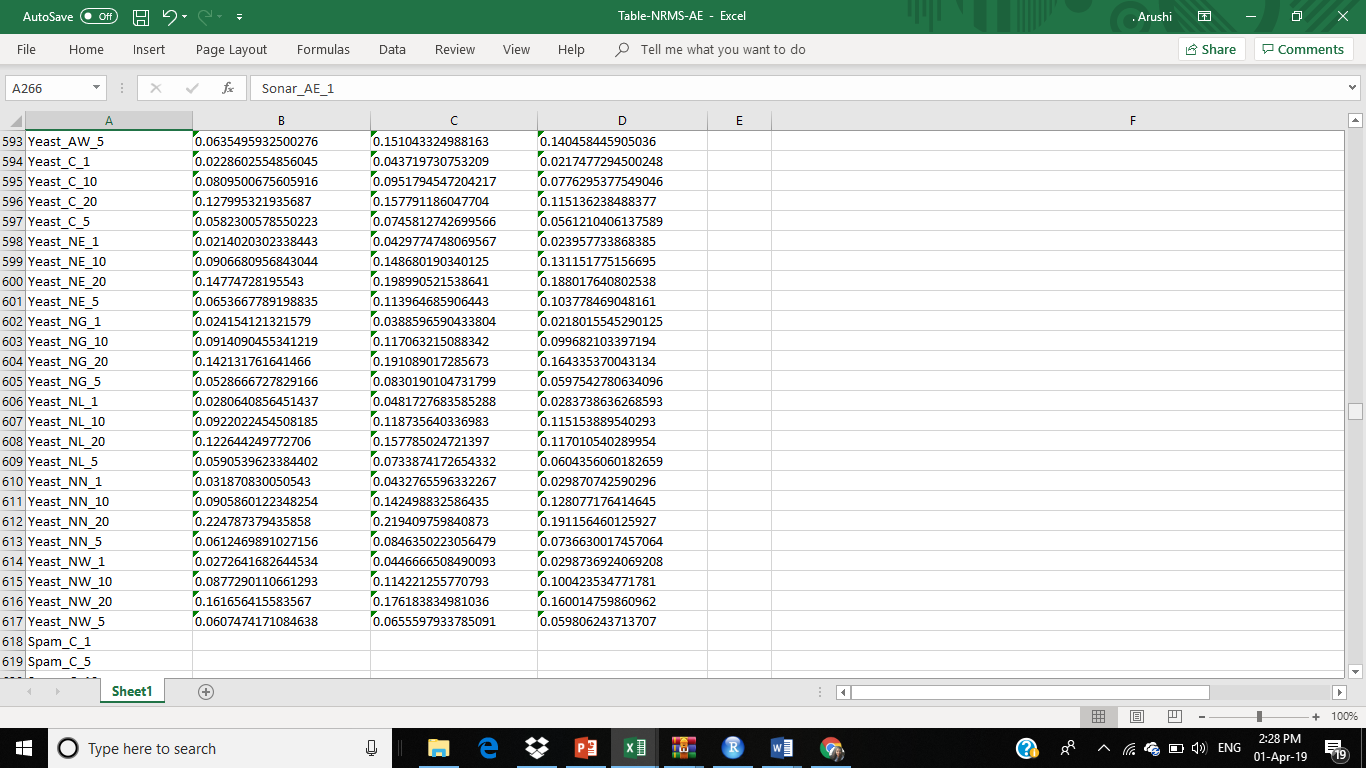
**DATASET: DIFDOUG**





**DATASET: YEAST**





**3.3 CONCLUSION**

A method is performed based on LRMC for missing data imputation. It takes consideration of the specific characteristics of each sample and performs LRMC on the set of samples with strong correlation. Moreover, an ensemble learning technique is presented which is commonly used in the field of machine learning to boost the imputation performance.

It has been proved that this algorithm provides better and more accurate results than the competing algorithms and makes LRMC method more effective in case of nonlinear dependency between rows and columns. Although, the proposed method has relatively high complexity which makes it slow in processing large data.

One of the improvements which is required in this method is the ability to handle categorical data.

**3.4 References**

[1] Xiaobo Chen, Zhongjie Wei, Zuoyong Li, Jun Liang, Yingfeng Cai, Bob Zhang, “Ensemble correlation based low-rank matrix completion with applications to traffic data imputation,” Knowledge-Based Systems, Volume 132, pp. 249-262, 2017.

[2] J. Fessler, November 28, 2018, 17:23

<https://web.eecs.umich.edu/~fessler/course/551/l/n-09-complete.pdf>

[3] Ryan Kennedy,“Low-rank matrix completion,”Written Preliminary Examination II, University of Pennsylvania,2013.