Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7699812

IMPROVISED K-MEANS CLUSTERING ALGORITHM TO CATEGORIZE THE COVID19 DATASET

Dr. P. Bavithra Matharasi
Associate Professor, Dept. of MCA, Mount Carmel College Autonomous, Bengaluru-560052, India, p.bavithra.matharasi@mccblr.edu.in

Dr. A. Clementking
Human Resource Development, Mount Carmel College Autonomous, Bengaluru-560052, India, clementking1975@gmail.com

Dr. S. Rani
Department of Information Technology (BCA/IT), Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai 600117 , India, srani.scs@velsuniv.ac.in

Dr. R. Roseline
Assistant Professor, PG Department of Computer Applications, St.Joseph's College of Arts and Science (Autonomous), Cuddalore-1, roseline_r@sjctnc.edu.in

S. Anupriya
Department of MCA, Vels Institute of Science, Technology & Advanced Studies (VISTAS) Chennai-600117, India, anupriya2778@gmail.com

Abstract
The Coronavirus Disease 2019 (COVID19) has brought severe stress on all human lives, especially to healthcare systems worldwide. The Healthcare system includes personnel, equipment and infrastructure, medicines, and voluminous dataset. This is a contagious and dreadful disease that timely treatment must be given to cure. All the patients affected by coronavirus cannot be treated the same way and provide the same resource. Depending on the severity of the disease, treatment should be given and resources should be allocated. Doctors and healthcare personnel interrogating the patients and identifying them based on their severity is a time-consuming process. In such a scenario, the help of IT and computer are the need of the hour. A computer-based system is needed that automatically categorizes the patients according to their severity. The machine learning approach will be best suited to category coronavirus-affected patients based on their severity. Unsupervised learning is a technique that categorizes the dataset without the label. Clustering is one of the techniques of unsupervised learning. There are ample number of clustering algorithms available in the market for this purpose. K-means algorithm is one of the clustering algorithms that is simple and efficient to use. This research project aims at improving the performance of the K-means clustering algorithm that categorizes the data more accurately.

Keywords: k-means clustering, covid19 dataset, machine learning, clustering algorithm

I. INTRODUCTION

Coronavirus disease has laid a huge responsibility on the healthcare sector. Healthcare sector includes doctors, nurses, hospitals, equipment, medicines and voluminous dataset. The dataset are generated and curated in such a way, that more complex tools and analysis techniques are necessary to find the insight. Doctors and researchers are inquisitive in finding the treatment and curing the corona virus infected patients. This is a contagious and dreadful disease that timely treatment should be given to cure. All the patients infected by this virus cannot be treated under the same umbrella or provide the same resource, rather, depending on the severity, treatment should be given and resource should be allocated. As the number of patients affected are large in number, manual identification and classification of the patients based on their severity is a time consuming process. A computer based system is needed that automatically categorizes the patients according to their severity.

II. LITERATURE REVIEW

Machine learning has become the most important technique used in every area of computational work. This approach will be best suited to categorize the corona virus infected patients based on their severity. It is often used in various applications such as computer security, engineering, biomedicine, and healthcare [1]. Clustering is a machine learning technique that could be used for grouping or classifying the dataset. It groups the dataset with similar characteristics into clusters [2]. There are number of clustering available for implementation. K-means algorithm is one of the simplest yet powerful algorithms for clustering [3]. This paper presents a modified, yet more efficient version of k-means algorithm with respect to the covid-19 dataset.

III. MACHINE LEARNING APPROACH

This approach is capable of performing tasks that are complex for human being to do manually. A machine learning system uses statistical model to analyze the voluminous of dataset. Machine learning approach is suitable in places where there is rapid increment in the generation of dataset. A traditional program approach takes input data, process the data and produces the output. Whereas the machine learning approach takes the dataset, model it using an algorithm and produces the output, where the out may be prediction, classification or grouping. A machine learning system learns from the past data or historical dataset, uses a statistical model and predicts the output.

A. K-Means algorithm

This is one of the most popular and efficient clustering algorithm that applies unsupervised learning. It works on unlabelled dataset, when it is given as input, the algorithm dives them into k clusters. It divides the given dataset into different clusters of similar characteristics based on the centroid. A centroid is a central data point towards which other data points with related or similar characteristics are grouped. There can be more than one and up to k centroids, such that there are k groups. The number of k or the groups must be specified in this algorithm. It is an iterative algorithm that classifies the input dataset into k clusters that has similar properties. This is shown in Figure 1(a) and 1(b).

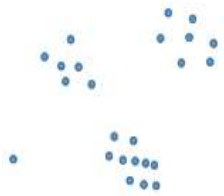


Figure 1 (a) Dataset before clustering

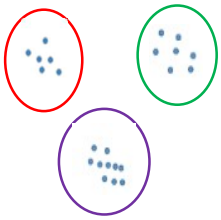


Figure 1 (b) Dataset after clustering

K-means algorithm works in two phases

1. By iterative process, the number of k centroids are computed
2. The data points that are closer to these k centroids are grouped

The steps involved in implementing k-means algorithm is given in figure 2.

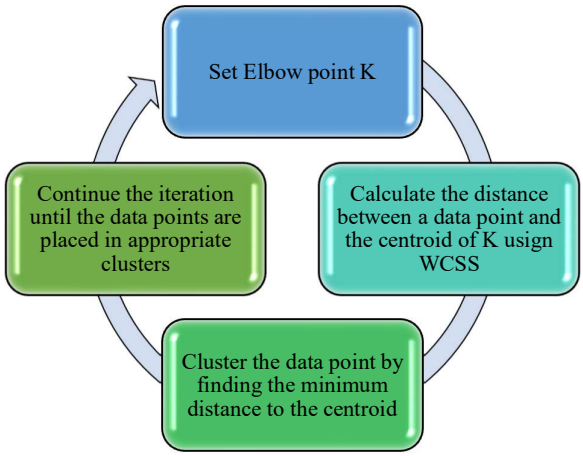


Figure 2: Steps involved in K-means clustering algorithm

The efficiency of this algorithm is in finding the number of centroids. There are various methods followed to calculate the centroid, like Euclidean distance or Manhattan distance method. But elbow method is one of the prominent and efficient techniques to find the centroids.

Elbow Method

This elbow method is based on the WCSS (Within Cluster Sum of Squares) value. Initially the centroids are randomly chosen. Then for each data point, the sum of squares of distance

between each data point and the centroid within each cluster are calculated. Formula to calculate the WCSS for three clusters, is given in equation 1.

$$WCSS = \sum_{k=1}^{3} distance\ between\ centroid\ and\ data\ points\ (Pk\ Ck)^2 \quad \text{-------- equ. (1)}$$

B. ***Key Features of K-means Clustering Algorithm***

- This k-means algorithm is very robust but simple to implement complex and voluminous dataset
- It is best suited for interpretations and resolutions
- It is faster than hierarchical clustering

C. ***Limitations of K-means Clustering***

- Though there are number of advantages in this algorithm, one of the adverse properties is, the performance goes down when implementing non-linear and noisy dataset.
- The second limitation is the user should specify the k value to the algorithm

IV. EXPERIMENTAL SETUP

This research project is implemented using Python. The code for this algorithm is written from the scratch, no packages are used. The dataset is stored in Microsoft Excel file. Though this is an unsupervised algorithm, labels are used for testing purposes only. The centroid values that are calculated in the program is store in the same excel file. The accuracy rate and the confusion matrix is also displayed in this same file. The results obtained from the Python shell is given in detailed in the 'Results and Interpretation' section.

V. LIFE CYCLE OF A MACHINE LEARNING SYSTEM

The machine-learning life cycle starts with the dataset rather than starting with program as in traditional programming paradigm. There are six steps involved in developing any project using machine learning approach. Given below in the Figure 3.

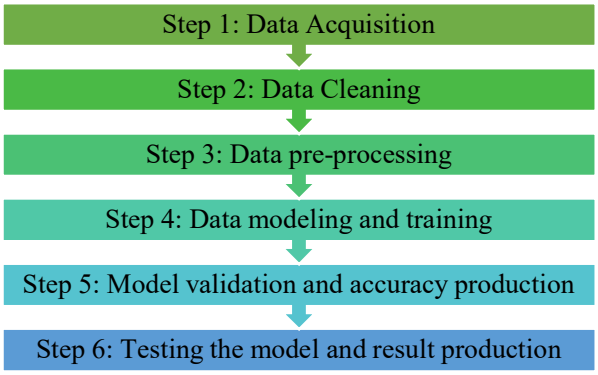
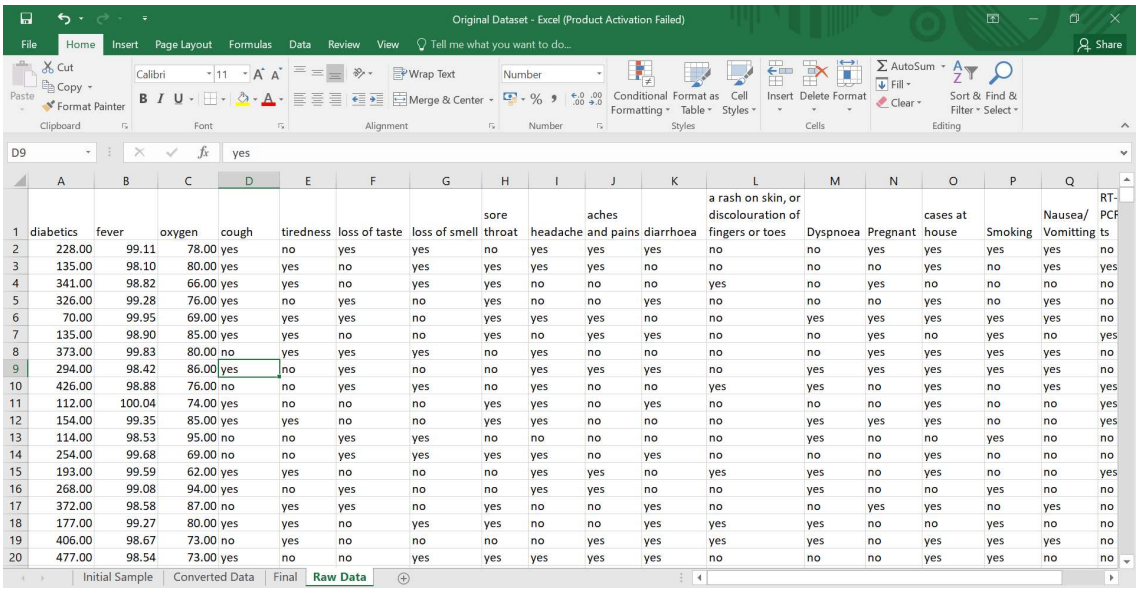


Figure 3: Steps in Machine learning approach

Step 1: Data Acquisition

The effectiveness and the efficiency of a machine learning solution depend on the nature and characteristics of data and the performance of the learning algorithms [4]. The accuracy of the model depends on the dataset that is fed into the system. The more dataset fed into the system, the more accurate result we get.

Data is generated from almost every parts of the world. Data is readily available for public on any domain. There are ample number of tools available in the market to scrap the data. Most of the time spent developing an ML application is spent on data preparation, e.g., Merck data scientists invest at least 90 % of their time seeking the appropriate data sets relevant for the task [5.]. This research deals with the covid19 dataset which is obtained from a healthcare sector.



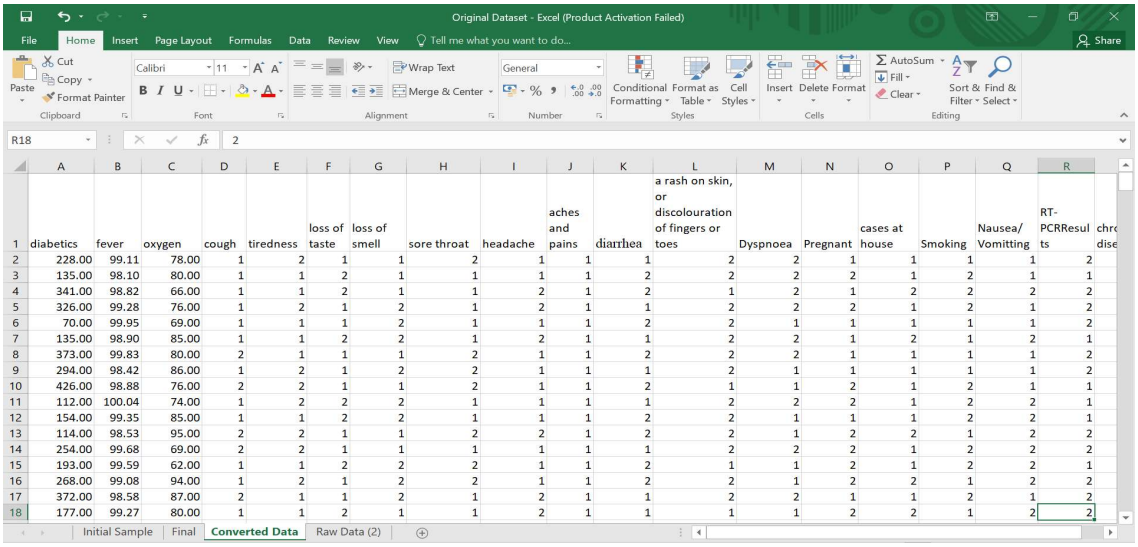
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1	diabetics	fever	oxygen	cough	tiredness	loss of taste	loss of smell	sore throat	headache	aches and pains	diarrhoea	a rash on skin, or discolouration of fingers or toes	Dyspnoea	Pregnant	cases at house	Smoking	Nausea/ Vomitting	RT-PCR ts
2	228.00	99.11	78.00	yes	no	yes	yes	no	yes	yes	yes	no	no	yes	yes	yes	yes	no
3	135.00	98.10	80.00	yes	yes	no	yes	yes	yes	yes	no	no	no	no	yes	no	yes	yes
4	341.00	98.82	66.00	yes	yes	no	yes	yes	no	no	no	yes	no	yes	no	no	no	no
5	326.00	99.28	76.00	yes	no	yes	no	yes	no	no	yes	no	no	no	yes	no	yes	no
6	70.00	99.95	69.00	yes	yes	yes	no	yes	yes	yes	no	no	yes	yes	yes	yes	yes	no
7	135.00	98.90	85.00	yes	yes	no	no	yes	no	yes	yes	no	no	yes	no	yes	no	yes
8	373.00	99.83	80.00	no	yes	yes	yes	no	yes	no	no	no	no	yes	yes	yes	yes	no
9	294.00	98.42	86.00	yes	no	yes	no	no	yes	yes	yes	no	yes	yes	yes	yes	yes	no
10	426.00	98.88	76.00	no	no	yes	yes	no	yes	no	no	yes	yes	no	yes	no	yes	yes
11	112.00	100.04	74.00	yes	no	no	no	yes	yes	no	yes	no	no	no	yes	no	no	yes
12	154.00	99.35	85.00	yes	yes	no	no	yes	yes	no	no	no	yes	yes	yes	no	no	yes
13	114.00	98.53	95.00	no	no	yes	yes	no	no	no	no	no	yes	no	no	yes	no	no
14	254.00	99.68	69.00	no	no	yes	yes	yes	yes	no	yes	no	no	no	yes	no	no	no
15	193.00	99.59	62.00	yes	yes	no	no	no	yes	yes	no	yes	yes	no	yes	no	no	yes
16	268.00	99.08	94.00	yes	no	yes	no	no	yes	yes	no	no	yes	no	no	yes	no	no
17	372.00	98.58	87.00	no	yes	yes	no	yes	no	no	yes	no	no	yes	yes	no	yes	no
18	177.00	99.27	80.00	yes	yes	no	yes	yes	no	no	yes	yes	yes	no	no	yes	no	no
19	406.00	98.67	73.00	no	yes	no	no	no	no	yes	yes	yes	yes	no	yes	yes	yes	no
20	477.00	98.54	73.00	yes	no	no	yes	yes	yes	yes	yes	no	no	no	yes	yes	no	no

Above screenshot is the original data collected from the source. There are totally twenty fields like diabetics, fever, oxygen, cough, tiredness, loss of taste, loss of smell, sore throat, headache, aches and pains, diarrhea, a rash on skin, or discolouration of fingers or toes, dyspnoea, pregnant, cases at house, smoking, nausea/ vomiting, and RT-PCR result

Step 2: Data Cleaning

This phase of the lifecycle involves removing the irrelevant data like special character, symbols, incomplete or unfilled data, and noisy data. Data cleaning is the initial stage of any machine learning project and is one of the most critical processes in data analysis. It is a critical step in ensuring that the dataset is devoid of incorrect or erroneous data [6]. Data collected from the various resources are dirty and this will affect the accuracy of prediction result. Data cleansing offers a better data quality which will be a great help for the organization to make sure their data is ready for the analyzing phase. [7]. Data cleaning process involves removing of those data or filling with default and appropriate data.

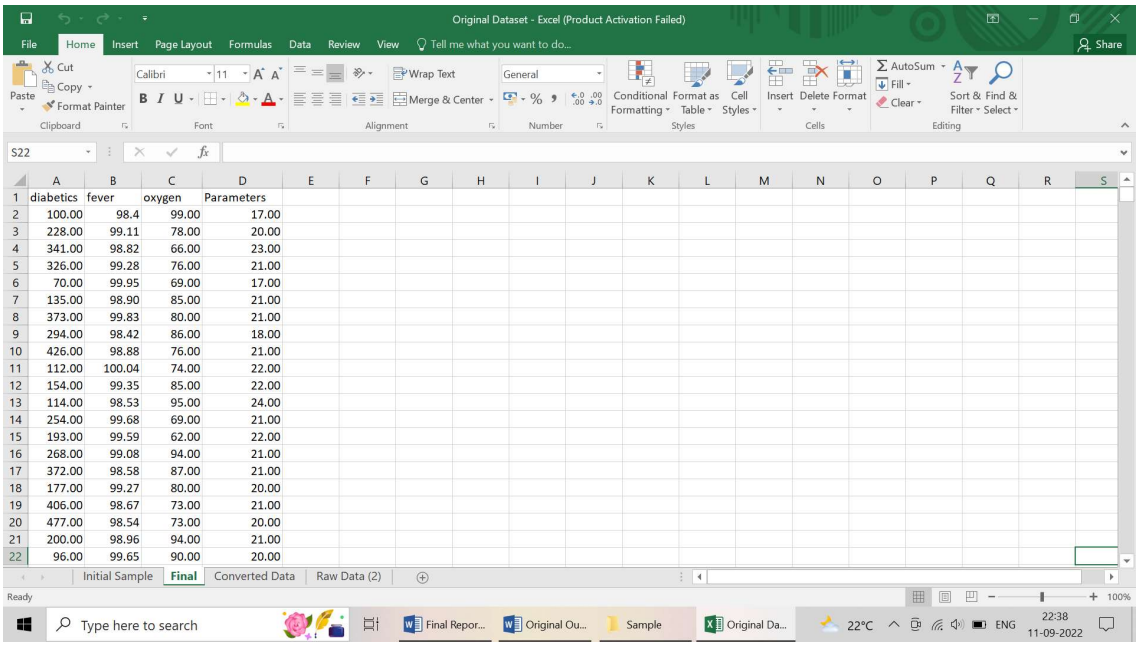
In this dataset, there are no unfilled data or empty column, whereas there are non-numeric datasets, which are to be cleaned. The cleaning process was done manually in the excel sheet by applying formulae. The below screenshot shows the cleaned datasets



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	diabetics	fever	oxygen	cough	tiredness	loss of taste	loss of smell	sore throat	headache	aches and pains	diarrhea	a rash on skin, or discolouration of fingers or toes	Dyspnoea	Pregnant	cases at house	Smoking	Nausea/ Vomitting	RT-PCRResults
2	228.00	99.11	78.00	1	2	1	1	2	1	1	1	2	2	1	1	1	1	2
3	135.00	98.10	80.00	1	1	2	1	1	1	1	2	2	2	2	1	2	1	1
4	341.00	98.82	66.00	1	1	2	1	1	2	1	2	1	2	1	2	2	2	2
5	326.00	99.28	76.00	1	2	1	2	1	2	1	1	2	2	2	1	2	1	2
6	70.00	99.95	69.00	1	1	1	2	1	1	1	2	2	1	1	1	1	1	2
7	135.00	98.90	85.00	1	1	2	2	1	2	1	1	2	2	1	2	1	2	1
8	373.00	99.83	80.00	2	1	1	1	2	1	1	2	2	2	1	1	1	1	2
9	294.00	98.42	86.00	1	2	1	2	2	1	1	1	2	1	1	1	1	1	2
10	426.00	98.88	76.00	2	2	1	1	2	1	1	2	1	1	2	1	2	1	1
11	112.00	100.04	74.00	1	2	2	2	1	1	1	1	2	2	2	1	2	2	1
12	154.00	99.35	85.00	1	1	2	2	1	1	1	2	2	1	1	1	2	2	1
13	114.00	98.53	95.00	2	2	1	1	2	2	1	2	2	1	2	2	1	2	2
14	254.00	99.68	69.00	2	2	1	1	1	1	1	1	2	2	2	1	2	2	2
15	193.00	99.59	62.00	1	1	2	2	2	1	1	2	1	1	2	1	2	2	1
16	268.00	99.08	94.00	1	2	1	2	2	1	1	2	2	1	2	2	1	2	2
17	372.00	98.58	87.00	2	1	1	2	1	2	1	1	2	2	1	1	2	1	2
18	177.00	99.27	80.00	1	1	2	1	1	2	1	1	1	1	2	2	1	2	2

Step 3: Data pre-processing

In this data preprocessing phase, only the fields that are necessary in classification or prediction is taken into consideration. Data reduction can be conducted in two directions, first, row-wise for data sample reduction and second, column-wise for data variable reduction. Data reduction is applied to reduce data dimensions and therefore, reducing the computational costs associated. [8]. The fields are dropped out. For example, in this research of clustering the covid19 dataset, patients name and address will not be used. This field has to be dropped and will not be used for any purpose. Further, due to confidentiality also, this field will not be used. There are two categories of data: the primary features like diabetics, fever and oxygen level and all other fields are secondary features. The primary features are used in classification, whereas the secondary features are only symptoms. When the secondary features are examined, one can infer that all the data are binary data and it cannot be ignored but it should be considered for classification. In such condition, a new technique is considered taking the sum of all the secondary features. The sum will range from 17 to 34, if all the symptoms are true, then the maximum sum value is 34. Moreover, if all the symptoms are false, then the minimum value is 17. After manipulation of the parameter field, the resultant values are given in the screenshot. Thus data pre-processing [9] is a major and essential stage whose main goal is to obtain final data sets which can be considered correct and useful for further data mining algorithms.

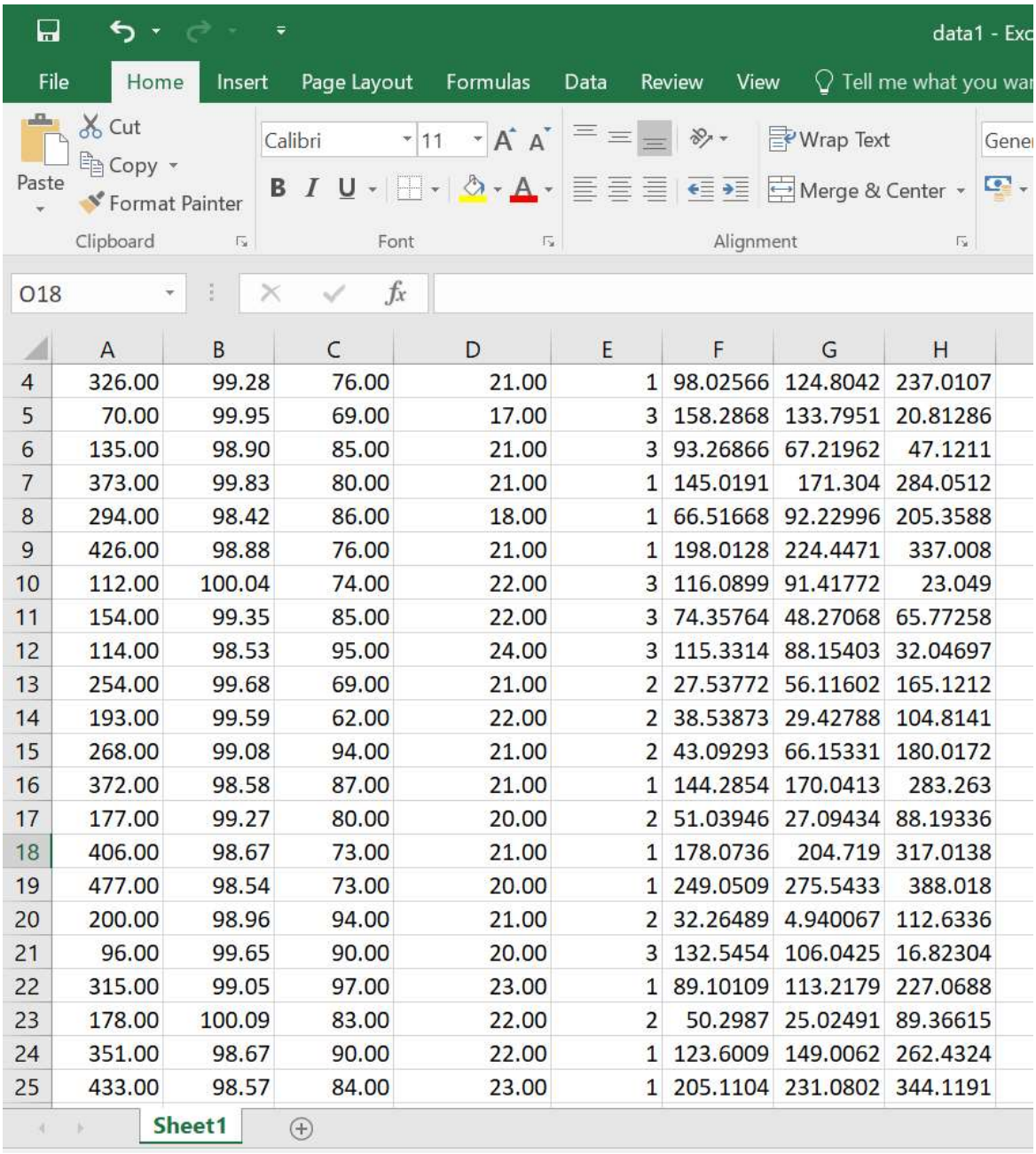


	A	B	C	D
1	diabetics	fever	oxygen	Parameters
2	100.00	98.4	99.00	17.00
3	228.00	99.11	78.00	20.00
4	341.00	98.82	66.00	23.00
5	326.00	99.28	76.00	21.00
6	70.00	99.95	69.00	17.00
7	135.00	98.90	85.00	21.00
8	373.00	99.83	80.00	21.00
9	294.00	98.42	86.00	18.00
10	426.00	98.88	76.00	21.00
11	112.00	100.04	74.00	22.00
12	154.00	99.35	85.00	22.00
13	114.00	98.53	95.00	24.00
14	254.00	99.68	69.00	21.00
15	193.00	99.59	62.00	22.00
16	268.00	99.08	94.00	21.00
17	372.00	98.58	87.00	21.00
18	177.00	99.27	80.00	20.00
19	406.00	98.67	73.00	21.00
20	477.00	98.54	73.00	20.00
21	200.00	98.96	94.00	21.00
22	96.00	99.65	90.00	20.00

Step 4: Data modeling and training

Machine learning involves creating a **model** that is trained on a set of **training data** and is then applied to additional data to make **predictions**. Various **types of models** have been used and researched for machine learning based systems [10]. Machine learning uses data and algorithms to build models that carry out certain tasks without being explicitly programmed [11]. While machine learning focuses on technologies, its application is part of data science. More precisely, data science uses principles, processes, and techniques for understanding phenomena via analysis of data [12]

It is in this phase the algorithm is implemented. K-means algorithm is implemented by finding the centroids. In this research, the datasets are clustered into three groups, namely ‘Acute’, ‘Mild’ and ‘Trivial’. The model is trained with the historic dataset and the results are produced for that training dataset. A newer way of identifying the clustering the dataset using K-means clustering algorithm is implemented in this research. The results of both methods is compared and the time taken to execute both codes is produced. This research projects aims at giving a better accuracy result, when compared to the traditional K-means clustering algorithm. The centroid calculations are performed and stored in excel file itself for visibility. The columns A, B, C and D are the features whereas columns F, G, and H are the centroid calculations. The screenshot is given below

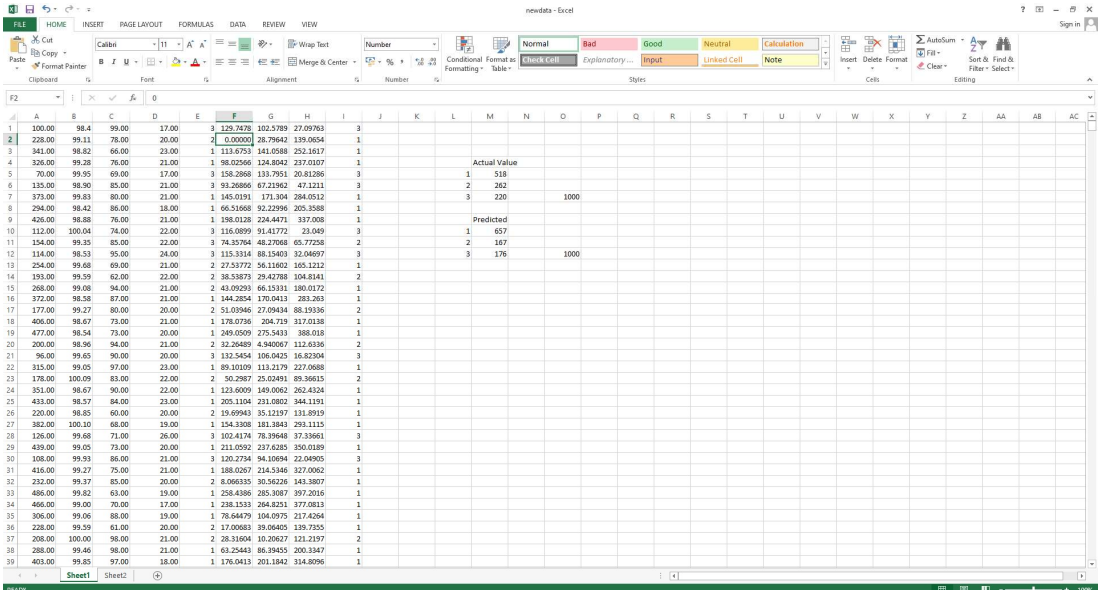


	A	B	C	D	E	F	G	H
4	326.00	99.28	76.00	21.00	1	98.02566	124.8042	237.0107
5	70.00	99.95	69.00	17.00	3	158.2868	133.7951	20.81286
6	135.00	98.90	85.00	21.00	3	93.26866	67.21962	47.1211
7	373.00	99.83	80.00	21.00	1	145.0191	171.304	284.0512
8	294.00	98.42	86.00	18.00	1	66.51668	92.22996	205.3588
9	426.00	98.88	76.00	21.00	1	198.0128	224.4471	337.008
10	112.00	100.04	74.00	22.00	3	116.0899	91.41772	23.049
11	154.00	99.35	85.00	22.00	3	74.35764	48.27068	65.77258
12	114.00	98.53	95.00	24.00	3	115.3314	88.15403	32.04697
13	254.00	99.68	69.00	21.00	2	27.53772	56.11602	165.1212
14	193.00	99.59	62.00	22.00	2	38.53873	29.42788	104.8141
15	268.00	99.08	94.00	21.00	2	43.09293	66.15331	180.0172
16	372.00	98.58	87.00	21.00	1	144.2854	170.0413	283.263
17	177.00	99.27	80.00	20.00	2	51.03946	27.09434	88.19336
18	406.00	98.67	73.00	21.00	1	178.0736	204.719	317.0138
19	477.00	98.54	73.00	20.00	1	249.0509	275.5433	388.018
20	200.00	98.96	94.00	21.00	2	32.26489	4.940067	112.6336
21	96.00	99.65	90.00	20.00	3	132.5454	106.0425	16.82304
22	315.00	99.05	97.00	23.00	1	89.10109	113.2179	227.0688
23	178.00	100.09	83.00	22.00	2	50.2987	25.02491	89.36615
24	351.00	98.67	90.00	22.00	1	123.6009	149.0062	262.4324
25	433.00	98.57	84.00	23.00	1	205.1104	231.0802	344.1191

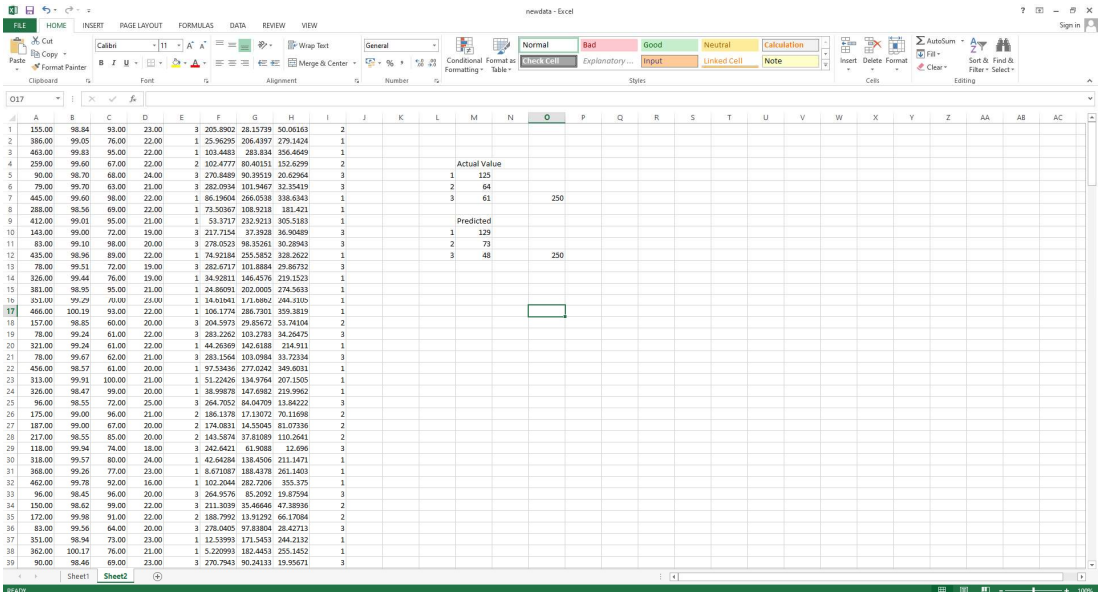
Step 5: Model validation and accuracy production

Data errors are common and can be difficult to detect when developing and operating ML-enabled software systems. Data validation in ML projects is the process of ensuring the high quality of data that is fed into the ML algorithm(s). The aim is to continuously check and monitor the data in order to assess its quality and identify underlying issues in data quality [13][14][15]. Important data quality dimensions of consideration are with respect to accuracy, completeness, consistency, timeliness [16][17]. Recently, studies have demonstrated that the ML model performance increases when data quality is continuously monitored and corrected according to the data quality measurement results [18].

Once the model is trained with the dataset, it has to be validated for its accuracy prediction. Usually the dataset that is used for training the system will use for both training and testing in the ratio of 80% training and 20% of testing. There are various methods like cross validation, finding scores and confusion matrix are used for validating the mode. In this research project, once the model is trained with the given dataset, the accuracy rate is validated using cross validation. The validation and the calculations made are stored in the excel sheet itself. This is shown in below screenshots.



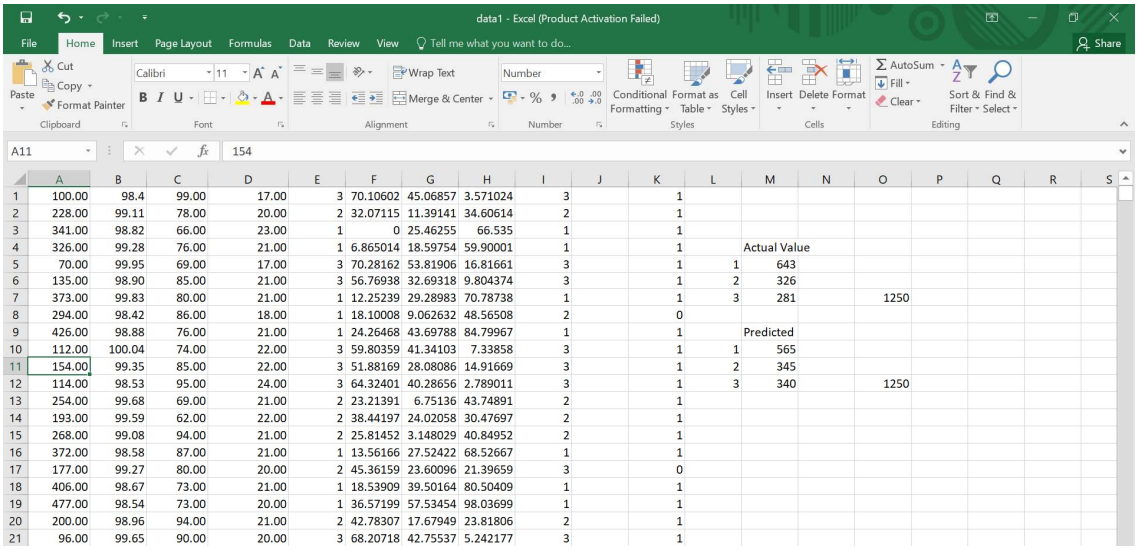
Training dataset 1000 (80%)



Testing dataset 250 (20%)

Step 6: Testing the model and result production

The model is designed and once the required level of accuracy is achieved, the real testing dataset is fed into the system and the result is predicted. In this project, an application will be developed that receives the input from the user and produces the result. For the purpose of testing the model, the actual classification is given in the excel sheet itself and shown as column E in the sheet. Whereas, the column I is the actual classified value. Further, the accuracy rate is calculated using confusion matrix and stored in the excel sheet itself. It is given in columns L through O. This is shown in the below screenshot.



data1 - Excel (Product Activation Failed)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	100.00	98.4	99.00	17.00	3	70.10602	45.06857	3.571024	3		1								
2	228.00	99.11	78.00	20.00	2	32.07115	11.39141	34.60614	2		1								
3	341.00	98.82	66.00	23.00	1	0	25.46255	66.535	1		1								
4	326.00	99.28	76.00	21.00	1	6.865014	18.59754	59.90001	1		1		Actual Value						
5	70.00	99.95	69.00	17.00	3	70.28162	53.81906	16.81661	3		1	1	643						
6	135.00	98.90	85.00	21.00	3	56.76938	32.69318	9.804374	3		1	2	326						
7	373.00	99.83	80.00	21.00	1	12.25239	29.28983	70.78738	1		1	3	281		1250				
8	294.00	98.42	86.00	18.00	1	18.10008	9.062632	48.56508	2		0								
9	426.00	98.88	76.00	21.00	1	24.26468	43.69788	84.79967	1		1		Predicted						
10	112.00	100.04	74.00	22.00	3	59.80359	41.34103	7.33858	3		1	1	565						
11	154.00	99.35	85.00	22.00	3	51.88169	28.08086	14.91669	3		1	2	345						
12	114.00	98.53	95.00	24.00	3	64.32401	40.28656	2.789011	3		1	3	340		1250				
13	254.00	99.68	69.00	21.00	2	23.21391	6.75136	43.74891	2		1								
14	193.00	99.59	62.00	22.00	2	38.44197	24.02058	30.47697	2		1								
15	268.00	99.08	94.00	21.00	2	25.81452	3.148029	40.84952	2		1								
16	372.00	98.58	87.00	21.00	1	13.56166	27.52422	68.52667	1		1								
17	177.00	99.27	80.00	20.00	2	45.36159	23.60096	21.39659	3		0								
18	406.00	98.67	73.00	21.00	1	18.53909	39.50164	80.50409	1		1								
19	477.00	98.54	73.00	20.00	1	36.57199	57.53454	98.03699	1		1								
20	200.00	98.96	94.00	21.00	2	42.78307	17.67949	23.81806	2		1								
21	96.00	99.65	90.00	20.00	3	68.20718	42.75537	5.242177	3		1								

Once the model is developed and trained and validated, it is used for classification. The results are shown below and all the results are discussed in the later section of Results and Interpretation.

```
========== RESTART: C:\Users\ADMIN\Desktop\MRP\Final\Original Implement Code.py ==========
Iteration   :    1 Accuracy rate :  82.16 %
Given Test dataset :  [340.0, 99.0, 65.0, 24.0]
3 - Please Admit in Hospital
Program Finished...
>>>
```

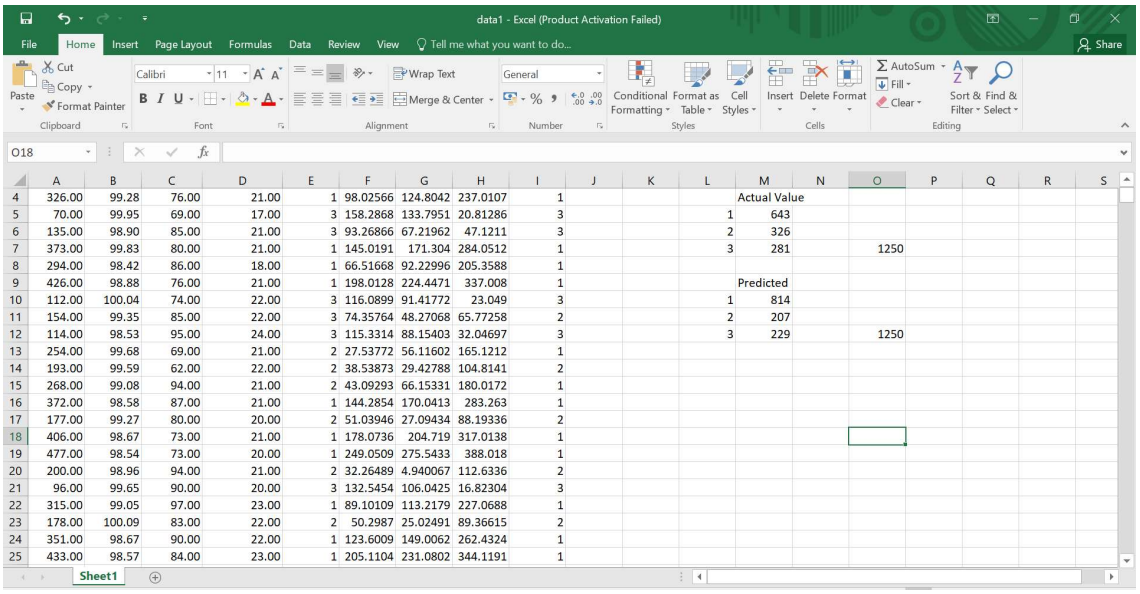
VI. RESULTS AND INTERPRETATION

Result while training

```
>>>
===== RESTART: C:\Users\ADMIN\Desktop\MRP\Final\Original Implement Code.py =====
Iteration   :    1 Accuracy rate :  82.16 %
```

```
Given Test dataset :  [100.0, 98.4, 99, 17]
1 - Be Happy
Program Finished...
>>>
========== RESTART: C:\Users\ADMIN\Desktop\MRP\Final\Original Implement Code.py ==========
Iteration   :    1 Accuracy rate :  82.16 %
Given Test dataset :  [230.0, 99.1, 80.0, 18.0]
2 - Quarantine and Take Rest
Program Finished...
>>>
========== RESTART: C:\Users\ADMIN\Desktop\MRP\Final\Original Implement Code.py ==========
Iteration   :    1 Accuracy rate :  82.16 %
Given Test dataset :  [340.0, 99.0, 65.0, 24.0]
3 - Please Admit in Hospital
Program Finished...
>>>
========== RESTART: C:\Users\ADMIN\Desktop\MRP\Final\Original Implement Code.py ==========
Iteration   :    1 Accuracy rate :  82.16 %
Given Test dataset :  [140.0, 99.0, 65.0, 24.0]
1 - Be Happy
Program Finished...
>>>
========== RESTART: C:\Users\ADMIN\Desktop\MRP\Final\Original Implement Code.py ==========
Iteration   :    1 Accuracy rate :  82.16 %
Given Test dataset :  [240.0, 99.0, 65.0, 24.0]
2 - Quarantine and Take Rest
Program Finished...
>>>
========== RESTART: C:\Users\ADMIN\Desktop\MRP\Final\Original Implement Code.py ==========
Iteration   :    1 Accuracy rate :  82.16 %
Given Test dataset :  [40.0, 99.0, 65.0, 24.0]
1 - Be Happy
Program Finished...
>>>
```

```
========== RESTART: C:\Users\ADMIN\Desktop\MRP\Final\Original Implement Code.py ==========
Iteration  :  1 Accuracy rate :  82.16 %
Given Test dataset :  [100.0, 99.0, 95.0, 24.0]
1 - Be Happy
Program Finished...
>>>
========== RESTART: C:\Users\ADMIN\Desktop\MRP\Final\Original Implement Code.py ==========
Iteration  :  1 Accuracy rate :  82.16 %
Given Test dataset :  [100.0, 99.0, 95.0, 17.0]
1 - Be Happy
Program Finished...
>>>
Confusion Matrix
```



	A	B	C	D	E	F	G	H	I	L	M	O
4	326.00	99.28	76.00	21.00	1	98.02566	124.8042	237.0107	1		Actual Value	
5	70.00	99.95	69.00	17.00	3	158.2868	133.7951	20.81286	3	1	643	
6	135.00	98.90	85.00	21.00	3	93.26866	67.21962	47.1211	3	2	326	
7	373.00	99.83	80.00	21.00	1	145.0191	171.304	284.0512	1	3	281	1250
8	294.00	98.42	86.00	18.00	1	66.51668	92.22996	205.3588	1			
9	426.00	98.88	76.00	21.00	1	198.0128	224.4471	337.008	1		Predicted	
10	112.00	100.04	74.00	22.00	3	116.0899	91.41772	23.049	3	1	814	
11	154.00	99.35	85.00	22.00	3	74.35764	48.27068	65.77258	2	2	207	
12	114.00	98.53	95.00	24.00	3	115.3314	88.15403	32.04697	3	3	229	1250
13	254.00	99.68	69.00	21.00	2	27.53772	56.11602	165.1212	1			
14	193.00	99.59	62.00	22.00	2	38.53873	29.42788	104.8141	2			
15	268.00	99.08	94.00	21.00	2	43.09293	66.15331	180.0172	1			
16	372.00	98.58	87.00	21.00	1	144.2854	170.0413	283.263	1			
17	177.00	99.27	80.00	20.00	2	51.03946	27.09434	88.19336	2			
18	406.00	98.67	73.00	21.00	1	178.0736	204.719	317.0138	1			
19	477.00	98.54	73.00	20.00	1	249.0509	275.5433	388.018	1			
20	200.00	98.96	94.00	21.00	2	32.26489	4.940067	112.6336	2			
21	96.00	99.65	90.00	20.00	3	132.5454	106.0425	16.82304	3			
22	315.00	99.05	97.00	23.00	1	89.10109	113.2179	227.0688	1			
23	178.00	100.09	83.00	22.00	2	50.2987	25.02491	89.36615	2			
24	351.00	98.67	90.00	22.00	1	123.6009	149.0062	262.4324	1			
25	433.00	98.57	84.00	23.00	1	205.1104	231.0802	344.1191	1			

```
Validation
========== RESTART: C:/Users/admin/Desktop/MRP/Final/Training.py ==========
TRAINING 1000
Iteration  :  1 Training Accuracy rate :  81.69999999999999 %
VALIDATION 250
Validation Accuracy rate :  93.2 %
Given Test dataset :  [100.0, 99.0, 95.0, 17.0]
1 - Be Happy
```

Program Finished...

==========

========== RESTART:

C:/Users/admin/Desktop/MRP/Final/Training.py

==========

==========

TRAINING 1000

Iteration : 1 Training Accuracy rate : 81.69999999999999 %

VALIDATION 250

Validation Accuracy rate : 93.2 %

Given Test dataset : [230.0, 99.1, 80.0, 18.0]

2 - Quarantine and Take Rest

Program Finished...

==========

========== RESTART:

C:/Users/admin/Desktop/MRP/Final/Training.py

==========

==========

TRAINING 1000

Iteration : 1 Training Accuracy rate : 81.69999999999999 %

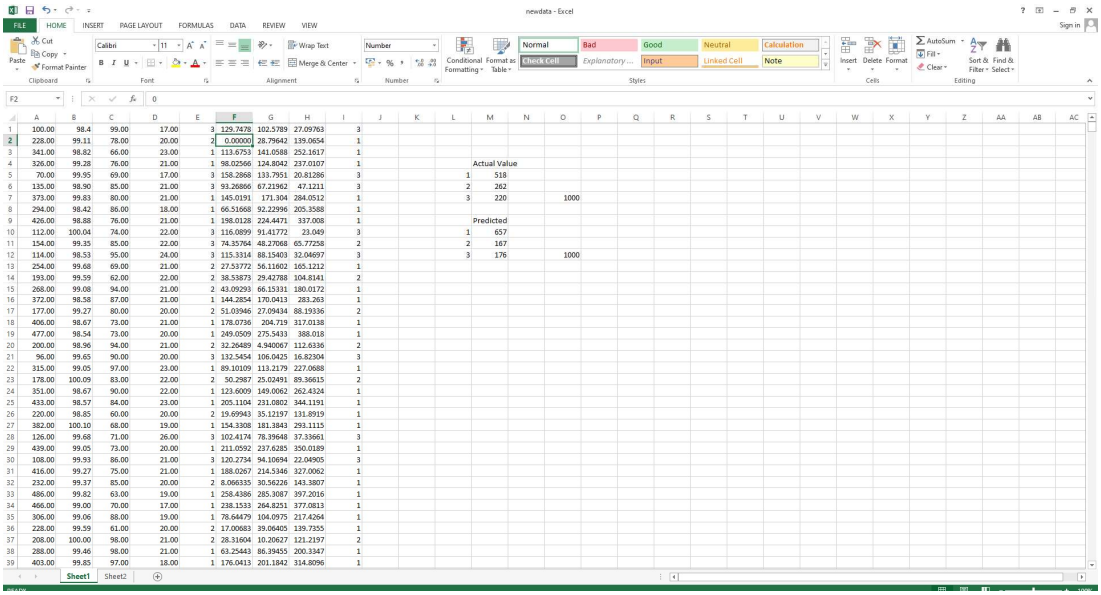
VALIDATION 250

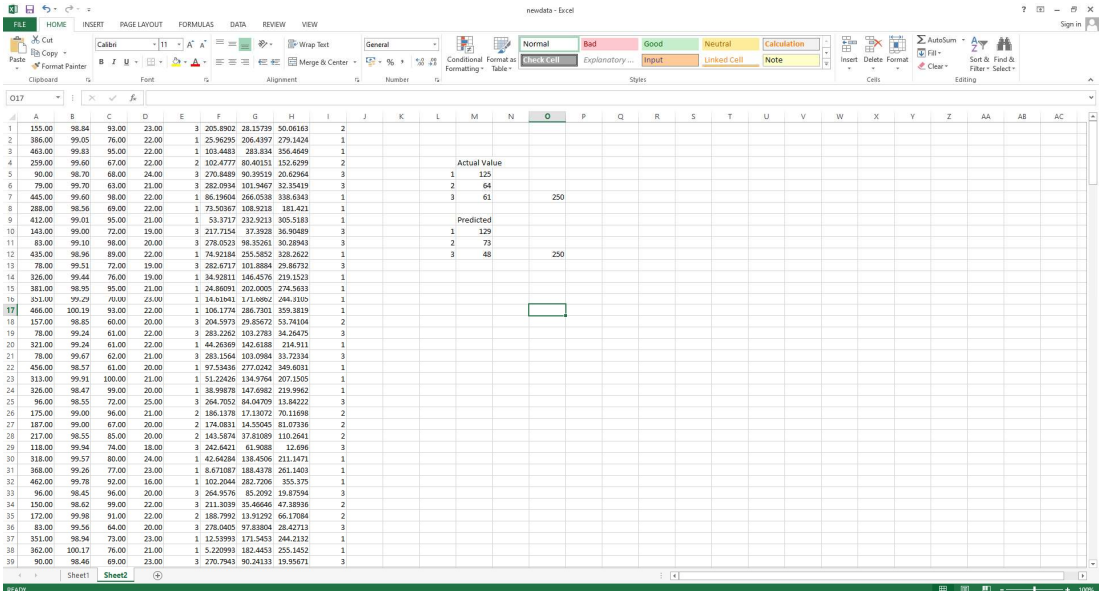
Validation Accuracy rate : 93.2 %

Given Test dataset : [100.0, 98.4, 99, 17]

1 - Be Happy

Program Finished...

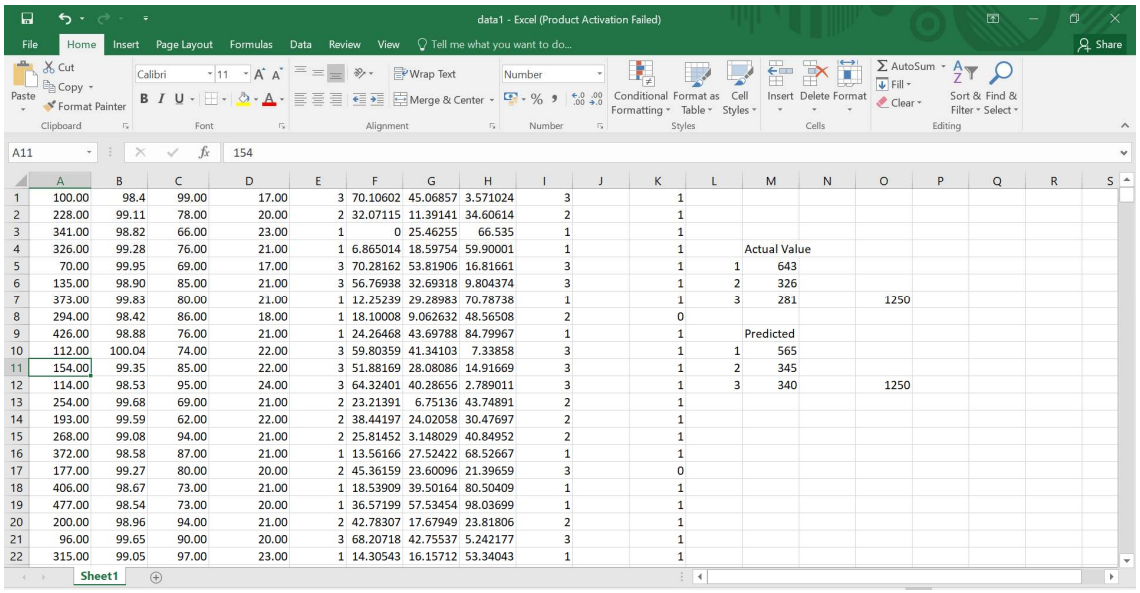




Modified Algorithm

```
>>>
====== RESTART: C:\Users\ADMIN\Desktop\MRP\Final\Modify Implement Code.py
======
Iteration   :    1 Accuracy rate :  89.03999999999999 %
Given Test dataset :  [100.0, 99.0, 95.0, 17.0]
1 - Be Happy
Program Finished...
>>>
================================================ RESTART:
C:\Users\ADMIN\Desktop\MRP\Final\Modify Implement Code.py
================================================
Iteration   :    1 Accuracy rate :  89.03999999999999 %
Given Test dataset :  [230.0, 99.1, 80.0, 18.0]
2 - Quarantine and Take Rest
Program Finished...
>>>
================================================ RESTART:
C:\Users\ADMIN\Desktop\MRP\Final\Modify Implement Code.py
================================================
Iteration   :    1 Accuracy rate :  89.03999999999999 %
Given Test dataset :  [100.0, 98.4, 99, 17]
1 - Be Happy
Program Finished...
```

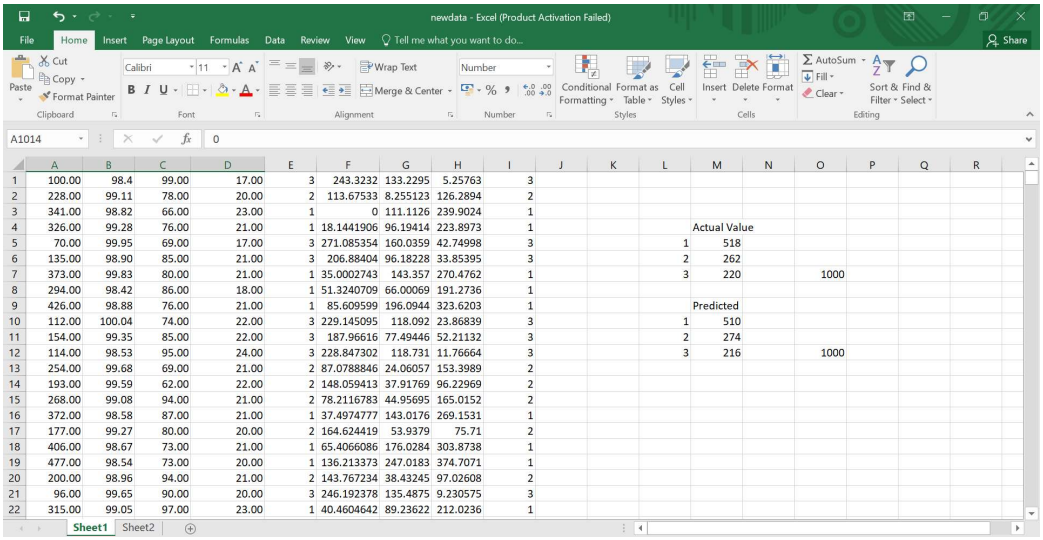
>>>



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	100.00	98.4	99.00	17.00	3	70.10602	45.06857	3.571024	3		1								
2	228.00	99.11	78.00	20.00	2	32.07115	11.39141	34.60614	2		1								
3	341.00	98.82	66.00	23.00	1	0	25.46255	66.535	1		1								
4	326.00	99.28	76.00	21.00	1	6.865014	18.59754	59.90001	1		1		Actual Value						
5	70.00	99.95	69.00	17.00	3	70.28162	53.81906	16.81661	3		1	1	643						
6	135.00	98.90	85.00	21.00	3	56.76938	32.69318	9.804374	3		1	2	326						
7	373.00	99.83	80.00	21.00	1	12.25239	29.28983	70.78738	1		1	3	281		1250				
8	294.00	98.42	86.00	18.00	1	18.10008	9.062632	48.56508	2		0								
9	426.00	98.88	76.00	21.00	1	24.26468	43.69788	84.79967	1		1		Predicted						
10	112.00	100.04	74.00	22.00	3	59.80359	41.34103	7.33858	3		1	1	565						
11	154.00	99.35	85.00	22.00	3	51.88169	28.08086	14.91669	3		1	2	345						
12	114.00	98.53	95.00	24.00	3	64.32401	40.28656	2.789011	3		1	3	340		1250				
13	254.00	99.68	69.00	21.00	2	23.21391	6.75136	43.74891	2		1								
14	193.00	99.59	62.00	22.00	2	38.44197	24.02058	30.47697	2		1								
15	268.00	99.08	94.00	21.00	2	25.81452	3.148029	40.84952	2		1								
16	372.00	98.58	87.00	21.00	1	13.56166	27.52422	68.52667	1		1								
17	177.00	99.27	80.00	20.00	2	45.36159	23.60096	21.39659	3		0								
18	406.00	98.67	73.00	21.00	1	18.53909	39.50164	80.50409	1		1								
19	477.00	98.54	73.00	20.00	1	36.57199	57.53454	98.03699	1		1								
20	200.00	98.96	94.00	21.00	2	42.78307	17.67949	23.81806	2		1								
21	96.00	99.65	90.00	20.00	3	68.20718	42.75537	5.242177	3		1								
22	315.00	99.05	97.00	23.00	1	14.30543	16.15712	53.34043	1		1								

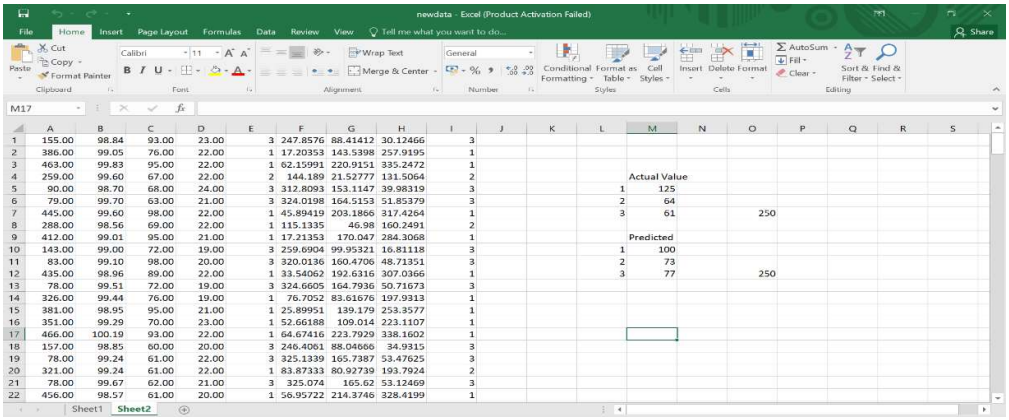
Validation
Training
>>>
========== RESTART: C:\Users\ADMIN\Desktop\MRP\Final\NewTraining.py ==========
TRAINING 1000
Iteration : 1 Training Accuracy rate : 89.3 %
VALIDATION 250
Validation Accuracy rate : 83.6 %
Given Test dataset : [100.0, 99.0, 95.0, 17.0]
1 - Be Happy
Program Finished...
>>>
========== RESTART: C:\Users\ADMIN\Desktop\MRP\Final\NewTraining.py ==========
TRAINING 1000
Iteration : 1 Training Accuracy rate : 89.3 %
VALIDATION 250
Validation Accuracy rate : 83.6 %
Given Test dataset : [230.0, 99.1, 80.0, 18.0]
2 - Quarantine and Take Rest
Program Finished...
>>>
========== RESTART: C:\Users\ADMIN\Desktop\MRP\Final\NewTraining.py ==========

TRAINING 1000
Iteration : 1 Training Accuracy rate : 89.3 %
VALIDATION 250
Validation Accuracy rate : 83.6 %
Given Test dataset : [100.0, 98.4, 99, 17]
1 - Be Happy
Program Finished...
>>>



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	100.00	98.4	99.00	17.00	3	243.3232	133.2295	5.25763	3									
2	228.00	99.11	78.00	20.00	2	113.67533	8.255123	126.2894	2									
3	341.00	98.82	66.00	23.00	1	0	111.1126	239.9024	1									
4	326.00	99.28	76.00	21.00	1	18.1441906	96.19414	223.8973	1				Actual Value					
5	70.00	99.95	69.00	17.00	3	271.085354	160.0359	42.74998	3			1	518					
6	135.00	98.90	85.00	21.00	3	206.88404	96.18228	33.85395	3			2	262					
7	373.00	99.83	80.00	21.00	1	35.0002743	143.357	270.4762	1			3	220		1000			
8	294.00	98.42	86.00	18.00	1	51.3240709	66.00069	191.2736	1									
9	426.00	98.88	76.00	21.00	1	85.609599	196.0944	323.6203	1				Predicted					
10	112.00	100.04	74.00	22.00	3	229.145095	118.092	23.86839	3			1	510					
11	154.00	99.35	85.00	22.00	3	187.96616	77.49446	52.21132	3			2	274					
12	114.00	98.53	95.00	24.00	3	228.847302	118.731	11.76664	3			3	216		1000			
13	254.00	99.68	69.00	21.00	2	87.0788846	24.06057	153.3989	2									
14	193.00	99.59	62.00	22.00	2	148.059413	37.91769	96.22969	2									
15	268.00	99.08	94.00	21.00	2	78.2116783	44.95695	165.0152	2									
16	372.00	98.58	87.00	21.00	1	37.4974777	143.0176	269.1531	1									
17	177.00	99.27	80.00	20.00	2	164.624419	53.9379	75.71	2									
18	406.00	98.67	73.00	21.00	1	65.4066086	176.0284	303.8738	1									
19	477.00	98.54	73.00	20.00	1	136.213373	247.0183	374.7071	1									
20	200.00	98.96	94.00	21.00	2	143.767234	38.43245	97.02608	2									
21	96.00	99.65	90.00	20.00	3	246.192378	135.4875	9.230575	3									
22	315.00	99.05	97.00	23.00	1	40.4604642	89.23622	212.0236	1									

Validation



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	155.00	98.84	93.00	23.00	3	247.8576	88.41412	30.12466	3										
2	386.00	99.05	76.00	22.00	1	17.20353	143.5398	257.9195	1										
3	463.00	99.83	95.00	22.00	1	62.15991	220.9151	335.2472	1										
4	259.00	99.60	67.00	22.00	2	144.189	21.52777	131.5064	2				Actual Value						
5	90.00	98.70	68.00	24.00	3	312.8093	153.1147	39.98319	3			1	125						
6	79.00	99.70	63.00	21.00	3	324.0198	164.5153	51.85379	3			2	64						
7	445.00	99.60	98.00	22.00	1	45.89419	203.1866	317.4264	1			3	61		250				
8	288.00	98.56	69.00	22.00	1	115.1335	46.98	160.2491	2										
9	412.00	99.01	95.00	21.00	1	17.21353	170.047	284.3068	1				Predicted						
10	143.00	99.00	72.00	19.00	3	259.6904	99.95321	16.81118	3			1	100						
11	83.00	99.10	98.00	20.00	3	320.0136	160.4706	48.71351	3			2	73						
12	435.00	98.96	89.00	22.00	1	33.54062	192.6316	307.0366	1			3	77		250				
13	78.00	99.51	72.00	19.00	3	324.6605	164.7936	50.71673	3										
14	326.00	99.44	76.00	19.00	1	76.7052	83.61676	197.9313	1										
15	381.00	98.95	95.00	21.00	1	25.89951	139.179	253.3577	1										
16	351.00	99.29	70.00	23.00	1	52.66188	109.014	223.1107	1										
17	466.00	100.19	93.00	22.00	1	64.67416	223.7929	338.1602	1										
18	157.00	98.85	60.00	20.00	3	246.4061	88.04666	34.9315	3										
19	78.00	99.24	61.00	22.00	3	325.1339	165.7387	53.47625	3										
20	321.00	99.24	61.00	22.00	1	83.87333	80.92739	193.7924	2										
21	78.00	99.67	62.00	21.00	3	325.074	165.62	53.12469	3										
22	456.00	98.57	61.00	20.00	1	56.95722	214.3746	328.4199	1										

VII. CONCLUSION

Covid 19 has historic event that penetrated into almost every parts of our life and changed the entire life style of every human being. A huge responsibility is laid on the healthcare sector in curing and avoiding Covid 19. The number of patients were increasing day by day and identifying the severity of the patients and treating them by the doctors were time consuming. Further, doctors and healthcare personnel were wasting their effort in identifying and classifying patients. The help of IT and computers are the need of the hour. Machine learning approach that uses a clustering algorithm is to be developed. This project aims at analysing

various clustering algorithm and develop a new algorithm that classifies the data more accurately. And this is achieved in this project.

VIII. REFERENCES

[1] Chaudhary, Divya & Vasuja, Richa. (2019). "A Review on Various Algorithms used in Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology", pp 915-920. 10.32628/CSEIT1952248.

[2] Halat Ahmed Hussein, & Adnan Mohsin Abdulazeez. (2021)." COVID-19 Pandemic Datasets Based on Machine Learning Clustering Algorithms: A Review", PalArch's Journal of Archaeology of Egypt / Egyptology, 18(4), pp 2672-2700. Retrieved from https://archives.palarch.nl/index.php/jae/article/view/6703

[3] Ahmed, Mohiuddin, Raihan Seraj, and Syed Mohammed Shamsul Islam. 2020. "The *k-means* Algorithm: A Comprehensive Survey and Performance Evaluation" *Electronics* 9, no. 8: 1295. https://doi.org/10.3390/electronics9081295

[4] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI.* 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x

[5] Stonebraker, Michael, and Ihab F. Ilyas. "Data Integration: The Current Status and the Way Forward." *IEEE Data Eng. Bull.* 41, no. 2 (2018): pp 3-9

[6] Lee, Ga & Alzamil, Lubna & Doskenov, Bakhtiyar & Termehchy, Arash. (2021). A Survey on Data Cleaning Methods for Improved Machine Learning Model Performance

[7] Fakhitah Ridzuan, Wan Mohd Nazmee Wan Zainon, A Review on Data Cleansing Methods for Big Data, Procedia Computer Science, Volume 161, 2019, pp 731-738, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.11.177

[8] Fan C, Chen M, Wang X, Wang J and Huang B (2021) A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Front. Energy Res.* 9:652801. doi: 10.3389/fenrg.2021.652801

[9] García, S., Ramírez-Gallego, S., Luengo, J. *et al.* Big data preprocessing: methods and prospects. *Big Data Anal* 1, 9 (2016). https://doi.org/10.1186/s41044-016-0014-0

[10] Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Science & Business Media (2009)

[11] Wolfgang Maass, Veda C. Storey, Pairing conceptual modeling with machine learning, Data & Knowledge Engineering, Volume 134, 2021, 101909, ISSN 0169-023X, https://doi.org/10.1016/j.datak.2021.101909

[12] Provost F., Fawcett T. Data Science for Business: What You Need To Know About Data Mining and Data-Analytic Thinking O'Reilly Media (2013)

[13] E. Breck, N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data validation for machine learning," in Proceedings of the 2nd SysML Conference. Online, 2019

[14] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger, "Automating large-scale data quality verification," Proc. VLDB Endow., vol. 11, no. 12, pp. 1781–1794,2018

[15] S. Krishnan, M. J. Franklin, K. Goldberg, J. Wang, and E. Wu, "Activeclean: An interactive data cleaning framework for modern machine learning," in Proceedings of the 2016 International Conference on Management of Data. New York: ACM, 2016, pp. 2117–2120

[16] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger, "Automating large-scale data quality verification," Proc. VLDB Endow., vol. 11, no. 12, pp. 1781–1794, 2018

[17] L. Ehrlinger, E. Rusz, and W. Woß, "A survey of data ¨ quality measurement and monitoring tools," in arXiv preprint arXiv:1907.08138. arXiv, 2019, pp. 1–30

[18] L. Ehrlinger, V. Haunschmid, D. Palazzini, and C. Lettner, "A DaQL to monitor data quality in machine learning applications," in Database and Expert Systems Applications, S. Hartmann, J. Kung, S. Chakravarthy, G. Anderst-Kotsis, A. M. Tjoa, and ¨ I. Khalil, Eds. Springer, 2019, pp. 227–237