An invited Lecture on Big Data Analytics in Value Added Course " Python Programming"

Mamta Sagar
Department of Bioinformatics
University Institute of Engineering & Technology, CSJM University, Kanpur

- Irrespective of the size of the enterprise whether it is big or small, data continues to be a precious and irreplaceable asset.
- Data is present in homogeneous sources as well as in heterogeneous sources.
- The need of the hour is to understand, manage, process, and take the data for analysis to draw valuable insights.
- Digital data can be structured, semi-structured or unstructured data.

1. Unstructured data:

This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program. About 80% data of an organization is in this format; for example, memos, chat rooms, PowerPoint presentations, images, videos, letters. researches, white papers, body of an email, etc.

2. Semi-structured data:

- Semi-structured data is also referred to as self-describing structure. This is the data which does not conform to a data model but has some structure. However, it is not in a form which can be used easily by a computer program. About 10% data of an organization is in this format; for example, HTML, XML, JSON, email data etc.

Structured data:

- When data follows a pre-defined schema/structure we say it is structured data. This is the data which is in an organized form (e.g., in rows and columns) and be easily used by a computer program. Relationships exist between entities of data, such as classes and their objects. About 10% data of an organization is in this format. Data stored in databases is an example of structured data.

Big data

- The term "Big Data" refers to the heterogeneous mass of digital data produced by companies and individuals whose characteristics (large volume, different forms, speed of processing) require specific and increasingly sophisticated computer storage and analysis tools. Concept of Big Data, its concepts, challenges and applications, as well as the importance of Big Data Analytics are discussed here.

- Being a complex polymorphic object, its definition varies according to the communities that are interested in it as a user or provider of services. Invented by the giants of the web, the Big Data presents itself as a solution designed to provide everyone a real-time access to giant databases.
- Big Data is a very difficult concept to define precisely, since the very notion of big in terms of volume of data varies from one area to another. It is not defined by a set of technologies, on the contrary, it defines a category of techniques and technologies.
- This is an emerging field, and as we seek to learn how to implement this new paradigm and harness the value, the definition is changing. [2]

Characteristics of Big Data

- 1) Characteristics of Big Data The term Big Data refers to gigantic larger datasets (volume); more diversified, including structured, semi-structured, and unstructured (variety) data, and arriving faster (velocity) than before. These are the 3V.

- Volume: represents the amount of data generated, stored and operated within the system. The increase in volume is explained by the increase in the amount of data generated and stored, but also by the need to exploit it.
- Variety: represents the multiplication of the types of data managed by an information system. This multiplication leads to a complexity of links and link types between these data. The variety also relates to the possible uses associated with a raw data.

- -Velocity: represents the frequency at which data is generated, captured, and shared. The data arrive by stream and must be analyzed in real time.
- To this classical characterization, two other "V"s are important: - Veracity: level of quality, accuracy and uncertainty of data and data sources.
- -Value: the value and potential derived from data.

5V

b) 5V:

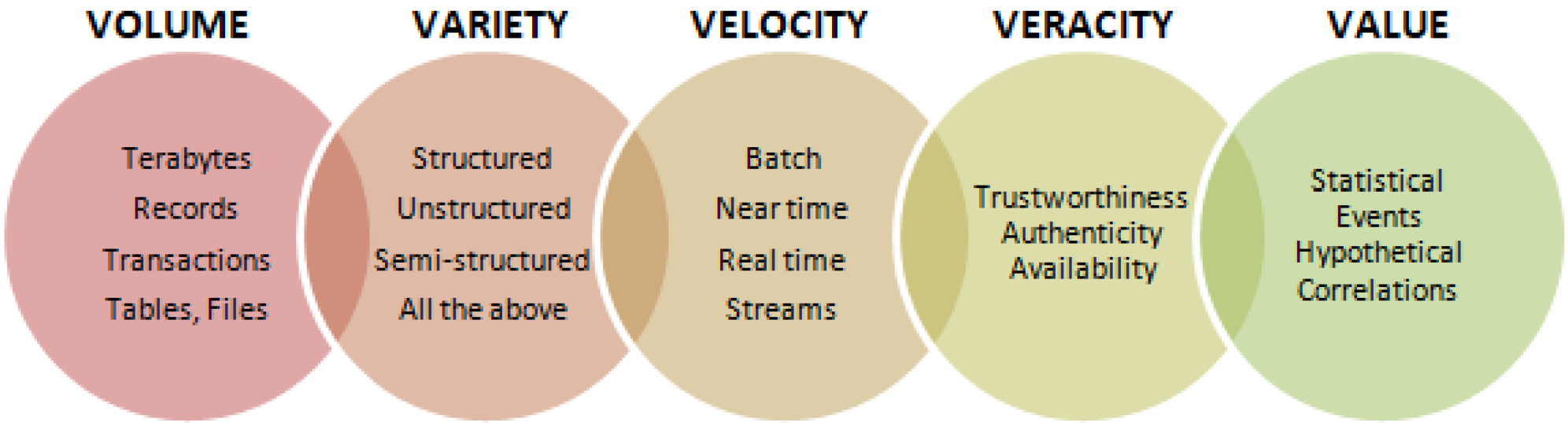


Fig. 2. *5V Concept*

WHAT IS BIG DATA ANALYTICS ?

- Big Data generally refers to data that exceeds the typical storage, processing, and computing capacity of conventional databases and data analysis techniques. As a resource, Big Data requires tools and methods that can be applied to analyze and extract patterns from large-scale data. [analytics-Najafabadi et al.]

Big Data Analytics refers to the process of collecting, organizing, analyzing large data sets to discover different patterns and other useful information.

Big data analytics is a set of technologies and techniques that require new forms of integration to disclose large hidden values from large datasets that are different from the usual ones, more complex, and of a large enormous scale.

It mainly focuses on solving new problems or old problems in better and effective ways. [4]

A. Types of Big Data Analytics

- a) Descriptive Analytics
- It consists of asking the question: What is happening? It is a preliminary stage of data processing that creates a set of historical data. Data mining methods organize data and help uncover patterns that offer insight. Descriptive analytics provides future probabilities and trends and gives an idea about what might happen in the future.

Diagnostic Analytics

- b) Diagnostic Analytics It consists of asking the question: Why did it happen? Diagnostic analytics looks for the root cause of a problem. It is used to determine why something happened. This type attempts to find and understand the causes of events and behaviors.

Predictive Analytics

- Predictive Analytics It consists of asking the question: What is likely to happen? It uses past data in order to predict the future. It is all about forecasting. Predictive analytics uses many techniques like data mining and artificial intelligence to analyze current data and make scenarios of what might happen.

Prescriptive Analytics

- d) Prescriptive Analytics It consists of asking the question: What should be done? It is dedicated to finding the right action to be taken. Descriptive analytics provides a historical data, and predictive analytics helps forecast what might happen. Prescriptive analytics uses these parameters to find the best solution.

HADOOP FOR BIG DATA APPLICATIONS

- Big Data are collections of information that would have been considered gigantic, impossible to store and process, a decade ago. The processing of such large quantities of data imposes particular methods. A classic database management system is unable to process as much information. Hadoop is an open source software product (or, more accurately, „software library framework“) that is collaboratively produced and freely distributed by the Apache Foundation – effectively, it is a developer‟s toolkit designed to simplify the building of Big Data solutions. [5]

Hadoop

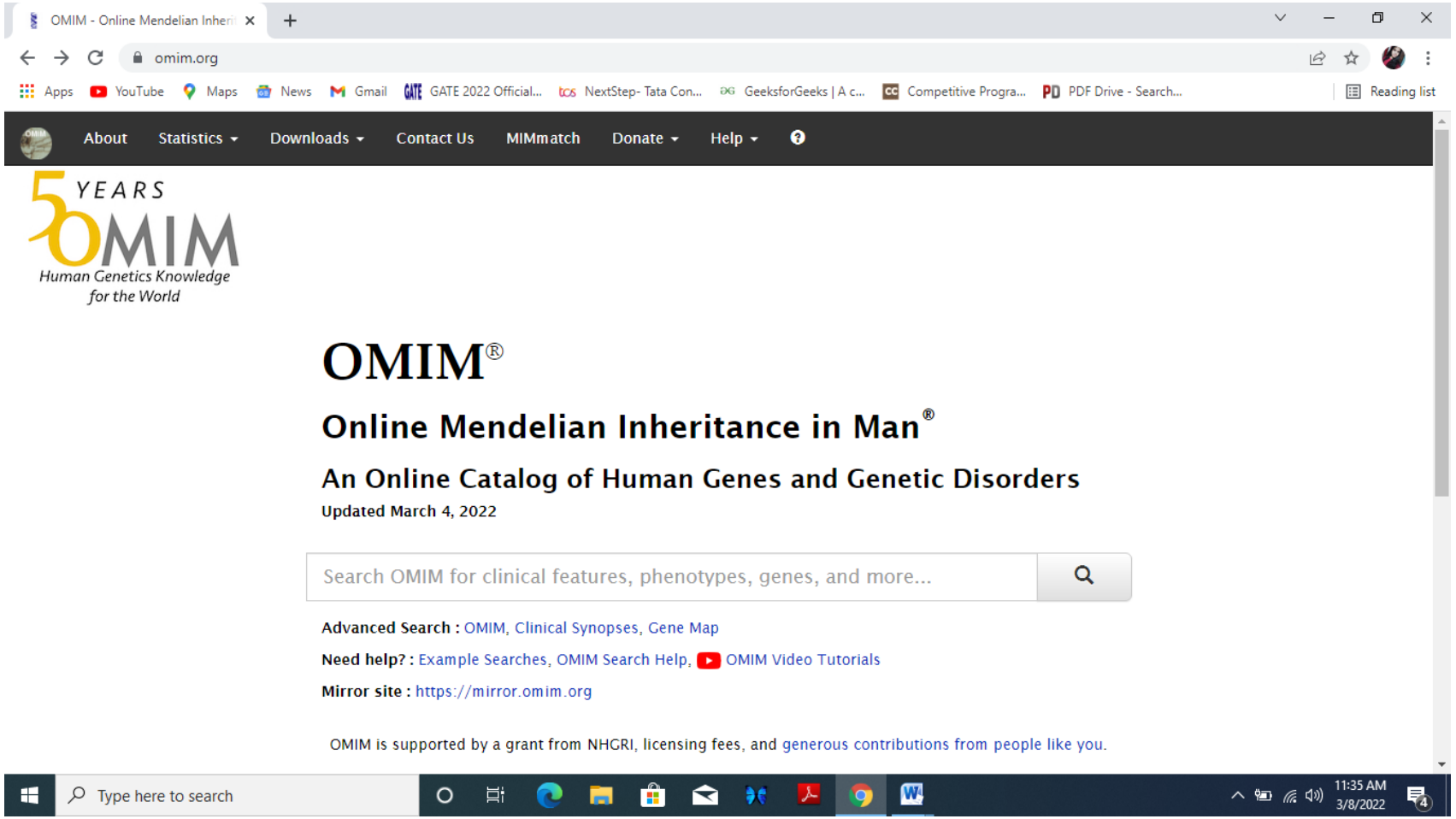
- Hadoop is a distributed data processing and management system. It contains many components, including: HDFS, YARN,Map Reduce. HDFS is a distributed file system that provides high-performance access to data across Hadoop clusters. [6]
- MapReduce is a core component of the Apache Hadoop software framework.]

- Hadoop relies on two servers: JobTracker: there is only one JobTracker per Hadoop cluster. It receives Map/Reduce tasks to run and organizes their execution on the cluster.When you submit your code to be executed on the Hadoop cluster, it is the JobTracker‟s responsibility to build an execution plan.

Materials

1. RAPID MINOR TOOL
2. DATA from OMIM Database.
3. Power Query

OMIM Database



OMIM Entry - # 601626 - LEUKE
omim.org/entry/601626?search=leukemia&highlight=%28leukaemia%7Cleukemia%29
Apps YouTube Maps News Gmail GATE GATE 2022 Official... NextStep- Tata Con... GeeksforGeeks | A c... Competitive Progra... PDF Drive - Search... Reading list

#601626
Table of Contents

Other entities represented in this entry:

LEUKEMIA, ACUTE MYELOID, SUSCEPTIBILITY TO, INCLUDED

Phenotype-Gene Relationships

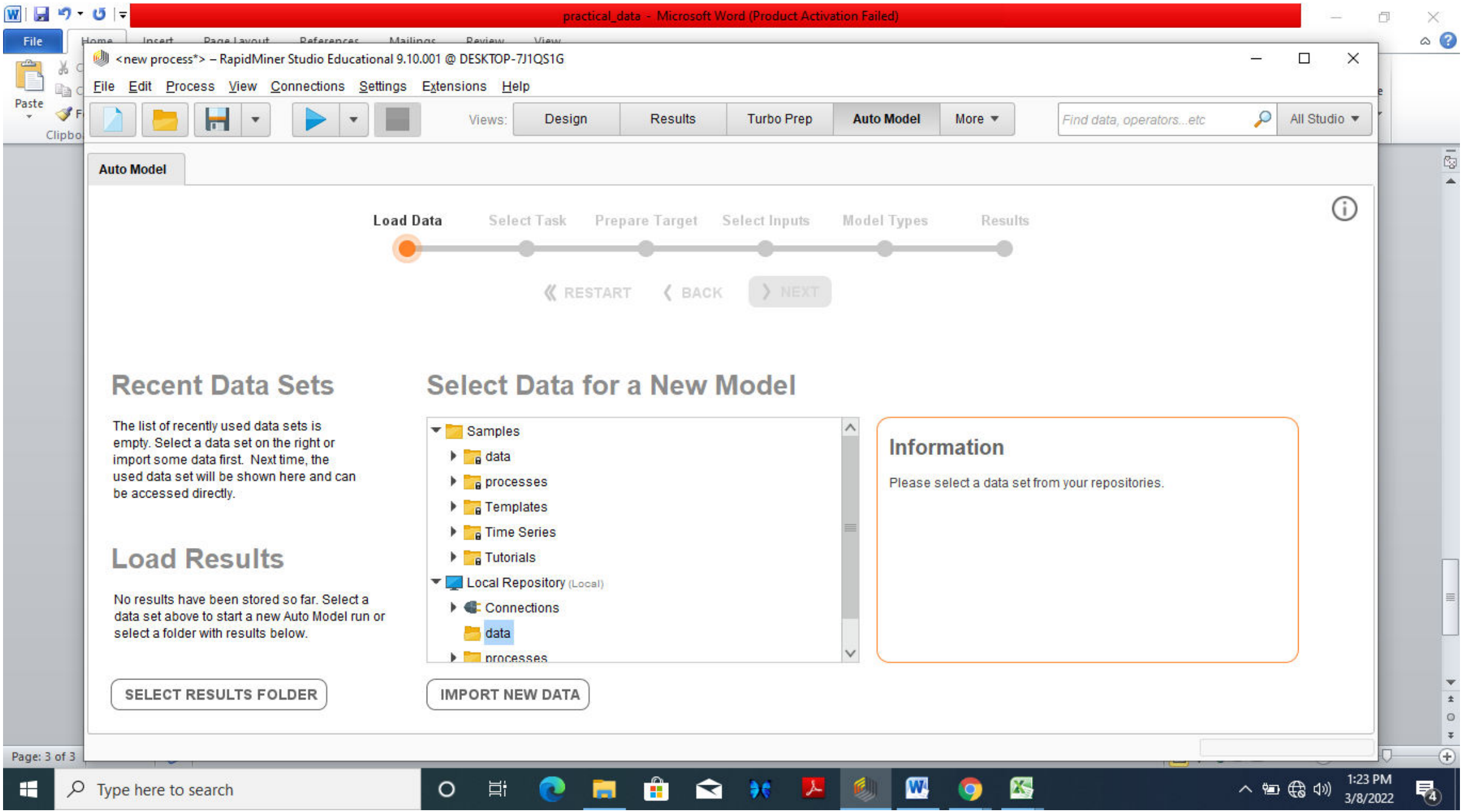
Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key	Gene/Locus	Gene/Locus MIM number
2p23.3	Acute myeloid leukemia, somatic	601626		3	DNMT3A	602769
3q21.3	{Leukemia, acute myeloid, susceptibility to}	601626	AD, SMu	3	GATA2	137295
3q27.3-q28	Leukemia, acute myeloid	601626	AD, SMu	3	LPP	600700
4q12	{Leukemia, acute myeloid}	601626	AD, SMu	3	CHIC2	604332
4q12	Leukemia, acute myeloid, somatic	601626		3	KIT	164920
5p15.33	{Leukemia, acute myeloid}	601626	AD, SMu	3	TERT	187270
5q35.1	Leukemia, acute myeloid, somatic	601626		3	NPM1	164040
9p24.1	Leukemia, acute myeloid, somatic	601626		3	JAK2	147796
9q34.13	Leukemia, acute myeloid, somatic	601626		3	NUP214	114350
10p12.31	Leukemia, acute myeloid	601626	AD, SMu	3	AF10	602409
11q14.2	Leukemia, acute myeloid, somatic	601626		3	PICALM	603025
12p13.2	Leukemia, acute myeloid, somatic	601626		3	ETV6	600618
12p12.1	Leukemia, acute myeloid, somatic	601626		3	KRAS	190070
13q12.2	Leukemia, acute myeloid, somatic	601626		3	FLT3	136351
13q12.2	Leukemia, acute myeloid, reduced survival in, somatic	601626		3	FLT3	136351
	Eo subtype, somatic	601626		1	CBFB	121360

External Links
Protein
Clinical Resources
Clinical Trials
EuroGentest
Gene Reviews
Genetic Alliance
MedlinePlus
Genetics
GTR
GARD
Orphanet
Animal Models
Cell Lines

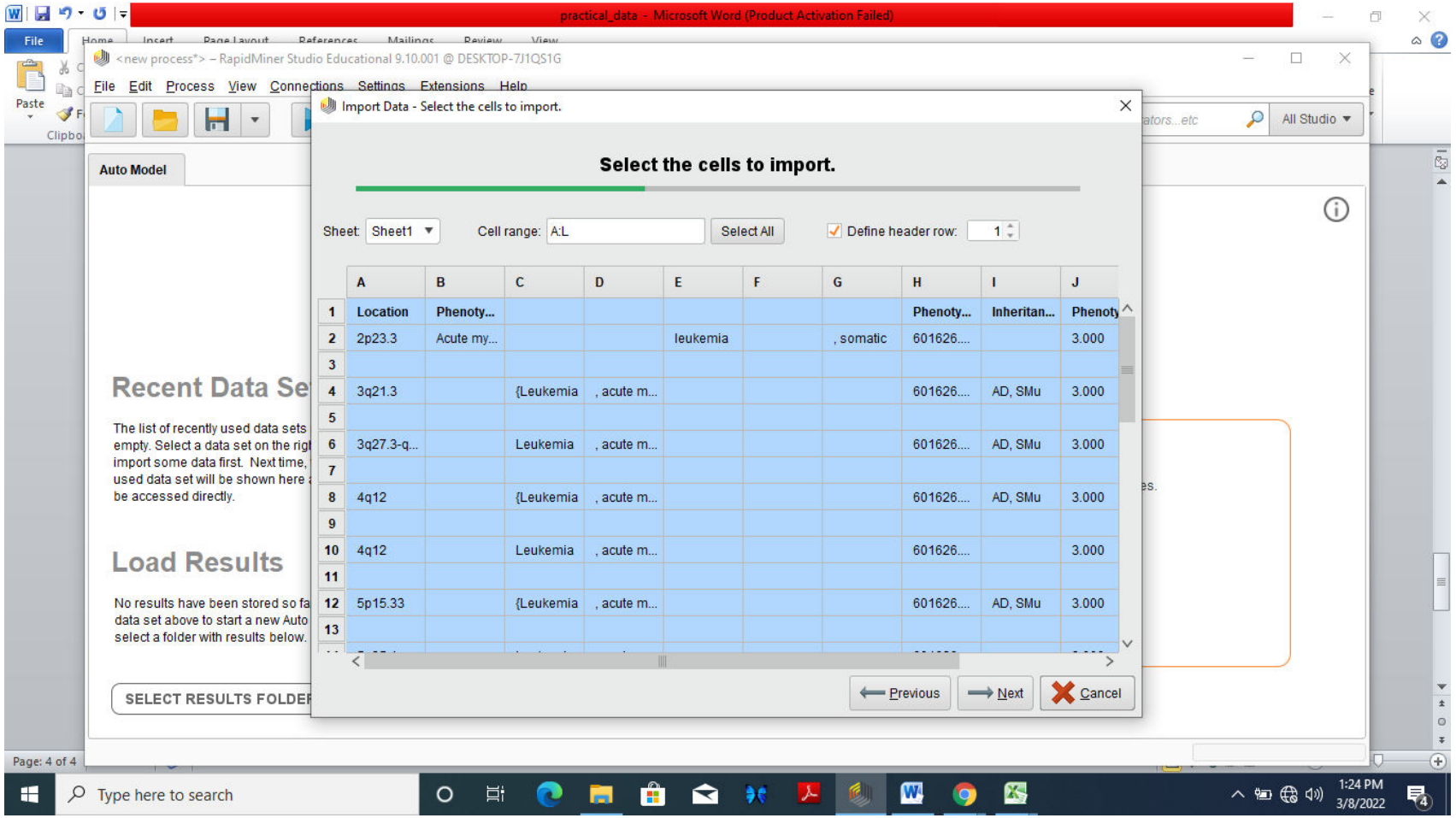
https://www.omim.org/entry/601626?search=leukemia&highlight=%28leukaemia%7Cleukemia%29#
Type here to search
11:30 AM 3/8/2022

RAPID MINOR
Procedures Of Clustering

1. **Load Data :** Import gene data as excel file
2. **Select Task :** Choose clustering
3. **Prepare Target :** Cleaning
4. **Select Input :** choose selected row/column
5. **Model Type :** Automodel
6. **Result :** K-mean cluster


practical_data - Microsoft Word (Product Activation Failed)
<new process*> – RapidMiner Studio Educational 9.10.001 @ DESKTOP-7J1QS1G
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model More
Find data, operators...etc
All Studio
Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
NEXT
Recent Data Sets
The list of recently used data sets is empty. Select a data set on the right or import some data first. Next time, the used data set will be shown here and can be accessed directly.
Load Results
No results have been stored so far. Select a data set above to start a new Auto Model run or select a folder with results below.
SELECT RESULTS FOLDER
Select Data for a New Model
Samples
data
processes
Templates
Time Series
Tutorials
Local Repository (Local)
Connections
data
processes
IMPORT NEW DATA
Information
Please select a data set from your repositories.
Page: 3 of 3
Type here to search
1:23 PM
3/8/2022

1.Load Data



	A	B	C	D	E	F	G	H	I	J
1	Location	Phenoty...						Phenoty...	Inheritan...	Phenoty
2	2p23.3	Acute my...			leukemia		, somatic	601626....		3.000
3										
4	3q21.3		{Leukemia	, acute m...				601626....	AD, SMu	3.000
5										
6	3q27.3-q...		Leukemia	, acute m...				601626....	AD, SMu	3.000
7										
8	4q12		{Leukemia	, acute m...				601626....	AD, SMu	3.000
9										
10	4q12		Leukemia	, acute m...				601626....		3.000
11										
12	5p15.33		{Leukemia	, acute m...				601626....	AD, SMu	3.000
13										

practical_data - Microsoft Word (Product Activation Failed)

<new process*> – RapidMiner Studio Educational 9.10.001 @ DESKTOP-7J1QS1G

File Edit Process View Connections Settings Extensions Help

Views: Design | Results | Turbo Prep | Auto Model | More

Find data, operators...etc | All Studio

Auto Model

Load Data — Select Task — Prepare Target — Select Inputs — Model Types — Results

« RESTART ‹ BACK › NEXT

Predict
Want to predict the values of a column?

Clusters
Want to identify groups in your data?

Outliers
Want to detect outliers in your data?

Location *Category*	Phenotype *Category*	C *Category*	D *Category*	E *Category*	F *Category*	G *Category*	Phenotype M... *Number*	Inheritance *Category*	Phenotype m... *Number*	Ge *Cat*
2p23.3	Acute myeloid	?	?	leukemia	?	, somatic	601626	?	3	D
?	?	?	?	?	?	?	?	?	?	?
3q21.3	?	{Leukemia	, acute myeloid,...	?	?	?	601626	AD, SMu	3	G
?	?	?	?	?	?	?	?	?	?	?
3q27.3-q28	?	Leukemia	, acute myeloid	?	?	?	601626	AD, SMu	3	Ll
?	?	?	?	?	?	?	?	?	?	?

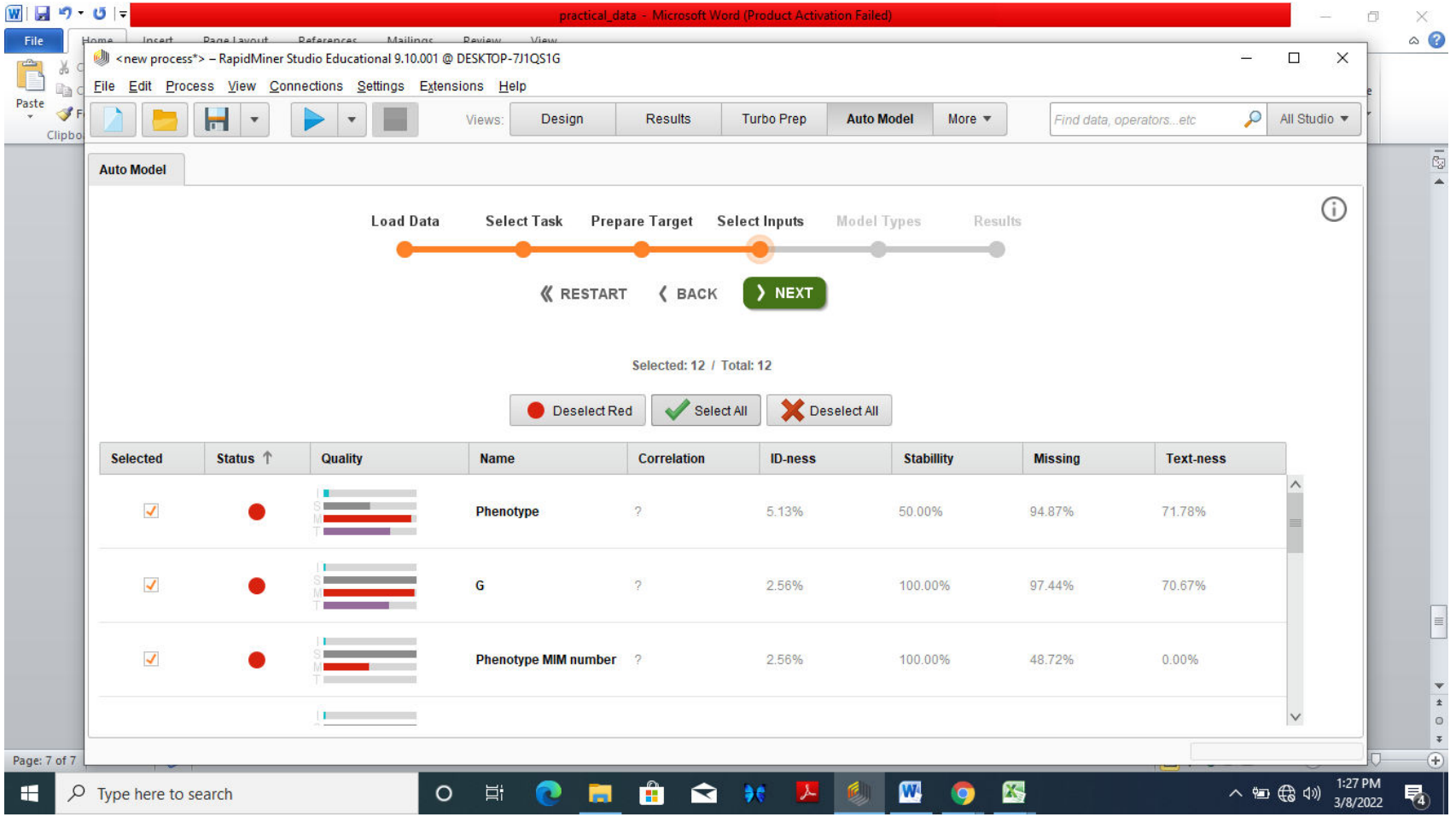
39 rows - 12 columns (9 nominal, 3 numerical)

Page: 6 of 6

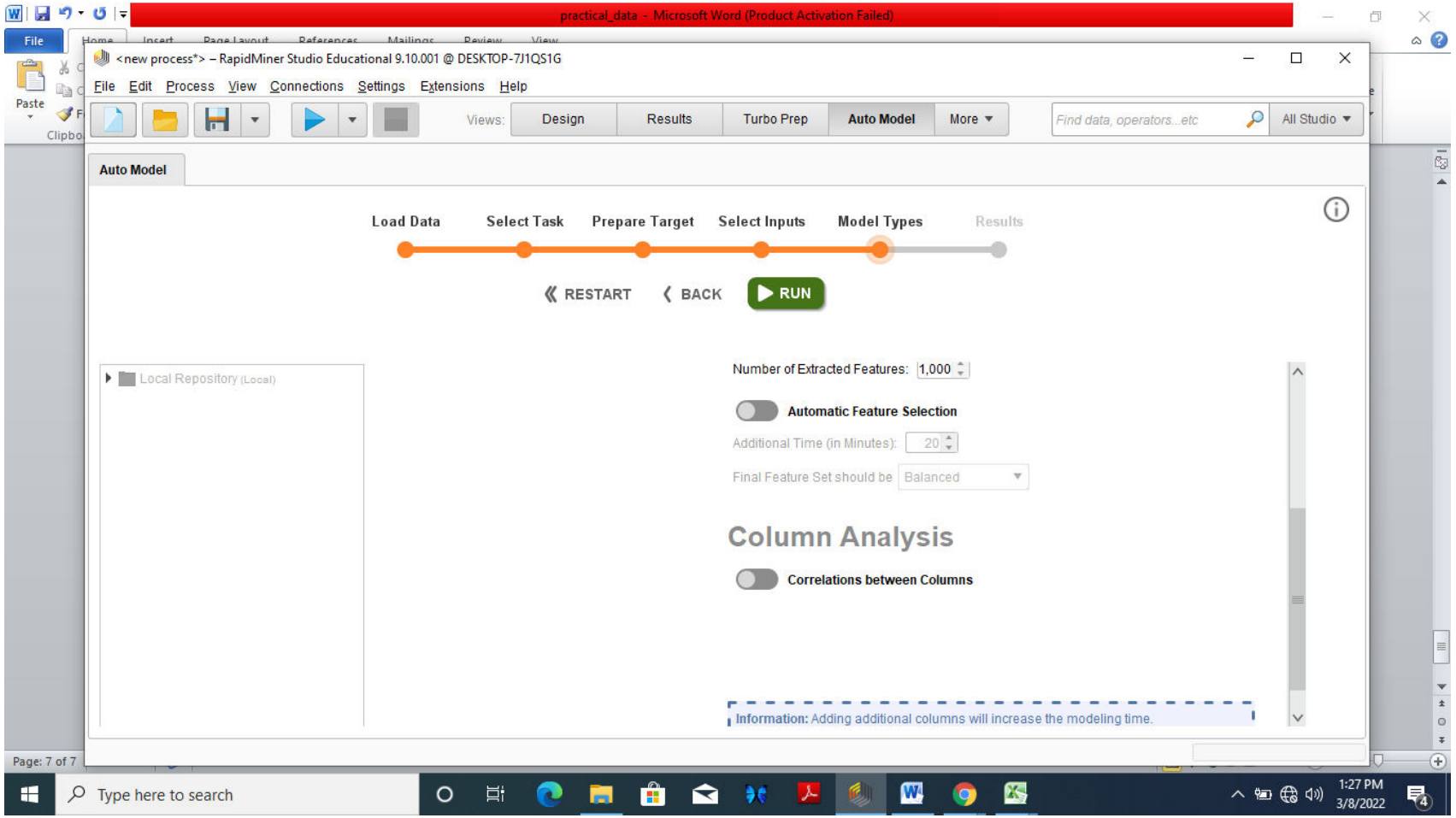
Type here to search

1:26 PM
3/8/2022

3. Cleaning & 4. Select Input



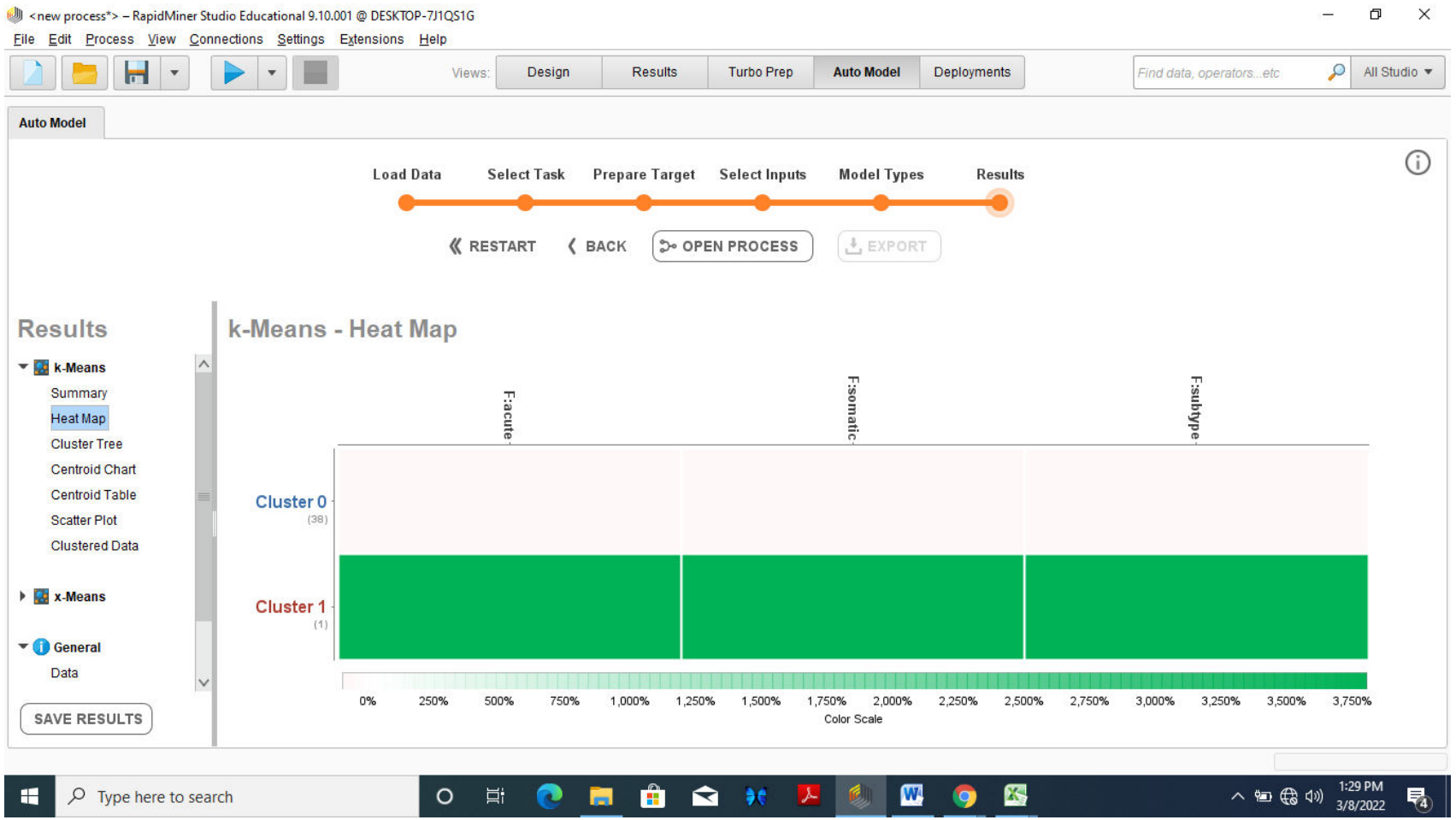
5.Model Type


practical_data - Microsoft Word (Product Activation Failed)
<new process*> – RapidMiner Studio Educational 9.10.001 @ DESKTOP-7J1QS1G
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model More
Find data, operators...etc
All Studio
Auto Model
Load Data
Select Task
Prepare Target
Select Inputs
Model Types
Results
RESTART
BACK
RUN
Local Repository (Local)
Number of Extracted Features: 1,000
Automatic Feature Selection
Additional Time (in Minutes): 20
Final Feature Set should be Balanced
Column Analysis
Correlations between Columns
Information: Adding additional columns will increase the modeling time.
Page: 7 of 7
Type here to search
1:27 PM
3/8/2022

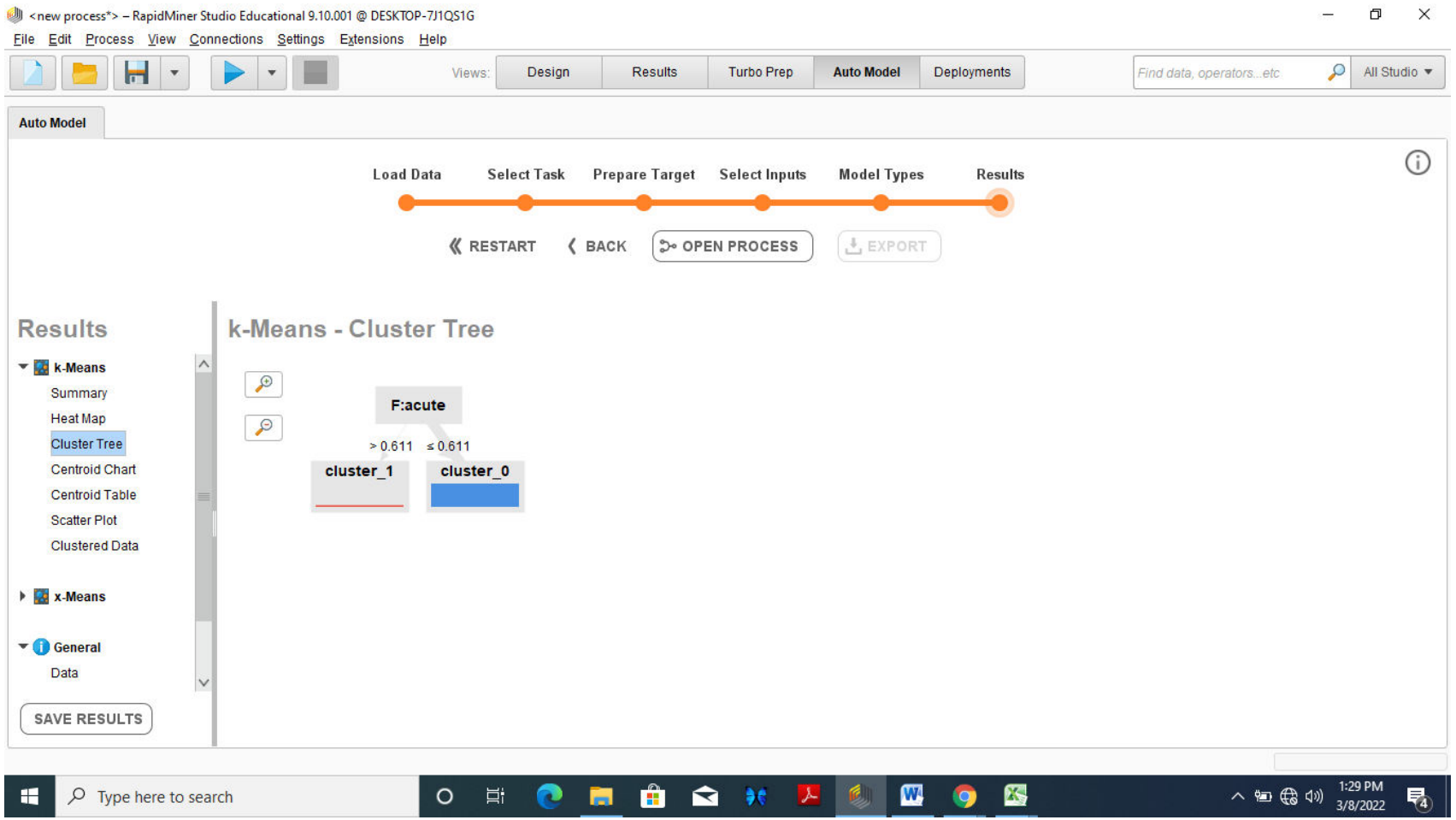
6.Result

(a) K-mean summary

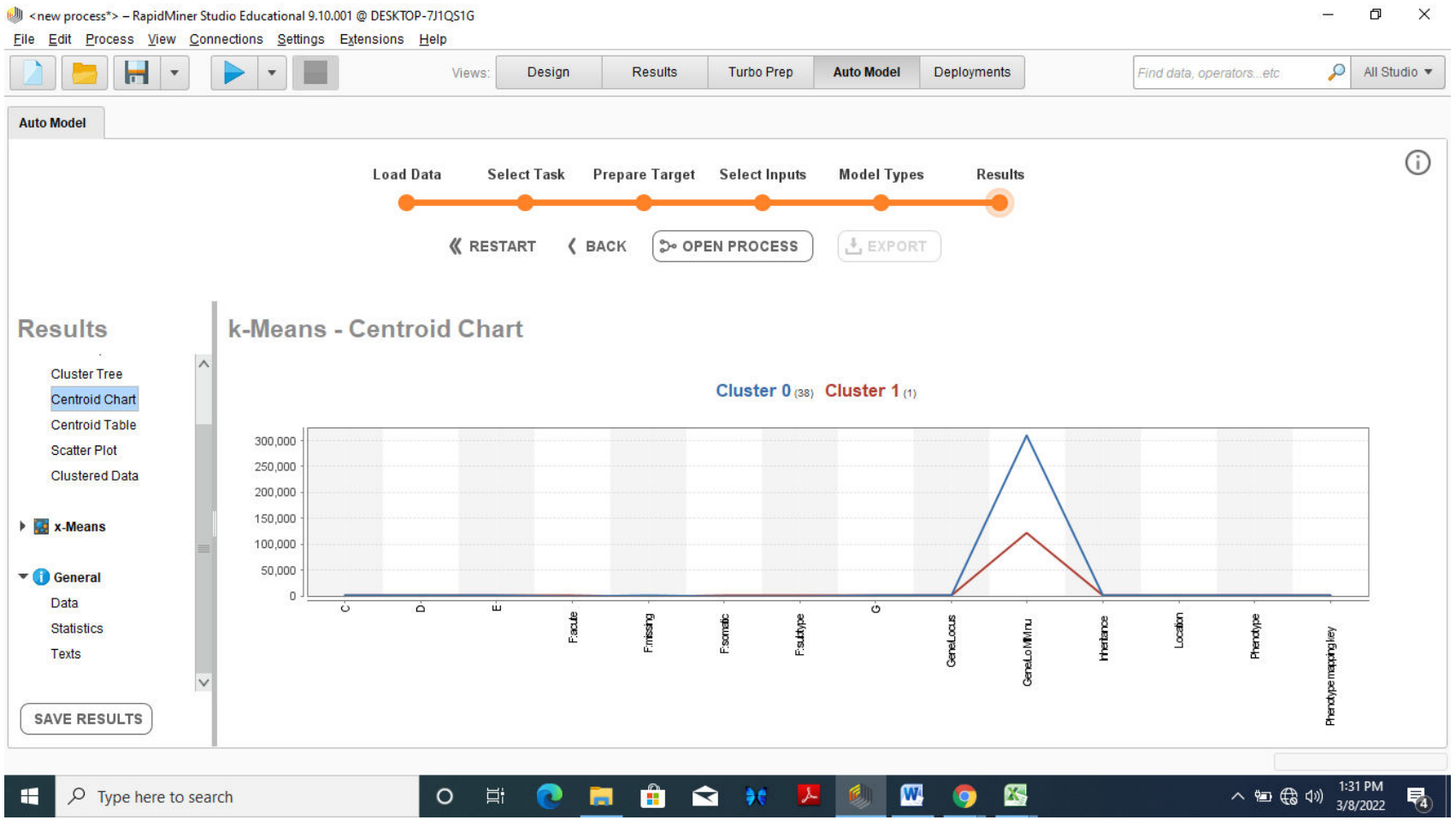
(b) K-mean HeatMap



(c) k-mean cluster tree



(d) k-mean centroid chart



Conclusion

- Rapid Minor tool can be used for executing automatic models from raw data by non-programmers and researchers etc.
- As in this practical , selected data is too short to process , still it’s get processed and most optimal results are concluded .

Healthcare Healthcare organizations are using big data for everything from improving profitability to helping save lives. Healthcare companies, hospitals, and researchers collect massive amounts of data. But all of this data isn’t useful in isolation. It becomes important when the data is analyzed to highlight trends and threats in patterns and create predictive models.

- Genomic research Big data can play in a significant role in genomic research. Using big data, researchers can identify disease genes and biomarkers to help patients pinpoint health issues they may face in the future. The results can even allow healthcare organizations to design personalized treatments. Challenges The volume of genome data is enormous, and running complex algorithms on the data is complicated and can require long processing times.

- Patient experience and outcomes Healthcare organizations seek to provide better treatment and improved quality of care—without increasing costs. Big data helps them improve the patient experience in the most cost-efficient manner.
- With big data, healthcare organizations can create a 360-degree view of patient care as the patient moves through various treatments and departments. Challenges Improving the patient experience requires a large volume of patient data, some of which could be multi-structured data, such as doctor notes or images. Additionally, to analyze patient journeys, path and graph analyses are often needed.

- Claims fraud For every healthcare claim, there can be hundreds of associated reports in a variety of different formats. This makes it extremely difficult to verify the accuracy of insurance incentive programs and find the patterns that indicate fraudulent activity. Big data helps healthcare organizations detect potential fraud by flagging certain behaviors for further examination. Challenges Claims fraud analytics is a complex process that involves integrating different data sets, analyzing the claims data, and identifying complex fraud patterns.

- Healthcare billing analytics Big data can improve the bottom line. By analyzing billing and claims data, organizations can discover lost revenue opportunities and places where payment cash flows can be improved.
- This use case requires integrating billing data from various payers, analyzing a large volume of that data, and then identifying activity patterns in the billing data. Challenges Sifting through large volumes of data can be complicated, especially when it comes to integrating different data sources

Oil and gas For the past few years, the oil and gas industry has been leveraging big data to find new ways to innovate. The industry has long made use of data sensors to track and monitor the performance of oil wells, machinery, and operations. Oil and gas companies have been able to harness this data to monitor well activity, create models of the Earth to find new oil sources, and perform many other value-added tasks.

- Predictive equipment maintenance Oil and gas companies often lack visibility into the condition of their equipment, especially in remote offshore and deep-water locations. Big data can help by providing insight so companies can predict the remaining optimal life of their systems and components, ensuring that their assets operate at optimum production efficiency.

- Challenges Machine, log, and sensor data from different types of equipment comes in varying formats. Integrating all of this data can be difficult. Moreover, the data needs to be analyzed quickly and put into operation to effectively prevent downtime.

- Oil exploration and discovery Exploring for oil and gas can be expensive. But companies can make use of the vast amount of data generated in the drilling and production process to make informed decisions about new drilling sites. Data generated from seismic monitors can be used to find new oil and gas sources by identifying traces that were previously overlooked. Challenges To discover potential new oil deposits, companies will need to integrate and analyze an enormous volume of unstructured data.

Oil production optimization

- Oil production optimization Unstructured sensor and historical data can be used to optimize oil well production. By creating predictive models, companies can measure well production to understand usage rates. With deeper data analysis, engineers can determine why actual well outputs aren’t tallying with their predictions. Challenges This use case involves analyzing a large volume of data. Complex algorithms are also needed to identify the curve shape associated with that data to identify trends.

References

- Top big data analytics use cases, Oracle
- Big Data and Big Data Analytics: Concepts, Types and Technologies Author(s) : 1Youssra Riahi, 2 Sara Riahi , International Journal of Research and Engineering ISSN: 2348-7860 (O) | 2348-7852 (P) | Vol. 5 No. 9 | September-October 2018 | PP. 524-528
- The Big Data Revolution, Issues and Applications, Azzeddine Riahi, Sara Riahi- IJARCSSE, Volume 5, Issue 8
- Deep learning applications and challenges in big data analytics- Najafabadi et al. Journal of Big Data (2015) 2:1 DOI 10.1186/s40537-014-0007-7

- BIG DATA ANALYTICS: CHALLENGES AND APPLICATIONS FOR TEXT, AUDIO, VIDEO, AND SOCIAL MEDIA DATA-International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), Vol.5, No.1, February 2016
- Big Data- The definitive guide to the revolution in business analytics-Fujitsu
- [7] http://searchcloudcomputing.techtarget.com/definition/MapReduce [8] http://www.informit.com/articles/article.aspx?p=20 08905 [9] http://www.informit.com/articles/article.aspx?p=20 08905