Data Mining Tool

Usman Ghani Mughal
SP17-BCS-087

Submitted to : Dr. Hikmat Ullah Khan
Data Warehousing and Data Mining
Introduction to Data Science

Introduction

- This Tool is a Data Science software develop using python.
- Provides Implementation of Algorithms on one click.
- Provides Preprocessing, Apply algorithm and shows results.
- At this time it can be used for Learning purpose.

Aim and Objectives

- Aim of this project was to learn and implement different algorithms from Scratch.
- Provide an Easy use for non-technical person.
- Learn how to develop end to end Product.

Algorithms

- Algorithms divided into 3 categories.
  - Association Rule mining
  - Classification
  - Clustering

Association Rule mining

- Apriori
- FP-Growth

Classification

- Naïve Bayesian classification
- Decision Tree
- Support Vector Machine
- Random Forest

Clustering

- K-Means Clustering
- Mean-Shift Clustering
- Hierarchical Clustering

Dataset's

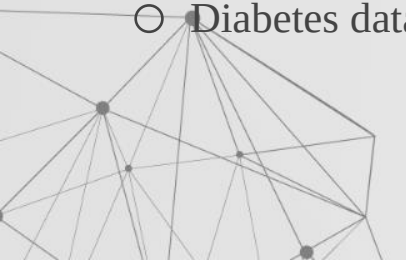
Some Data Sets which are used while testing and implementations are provided with Tool

Association Datasets

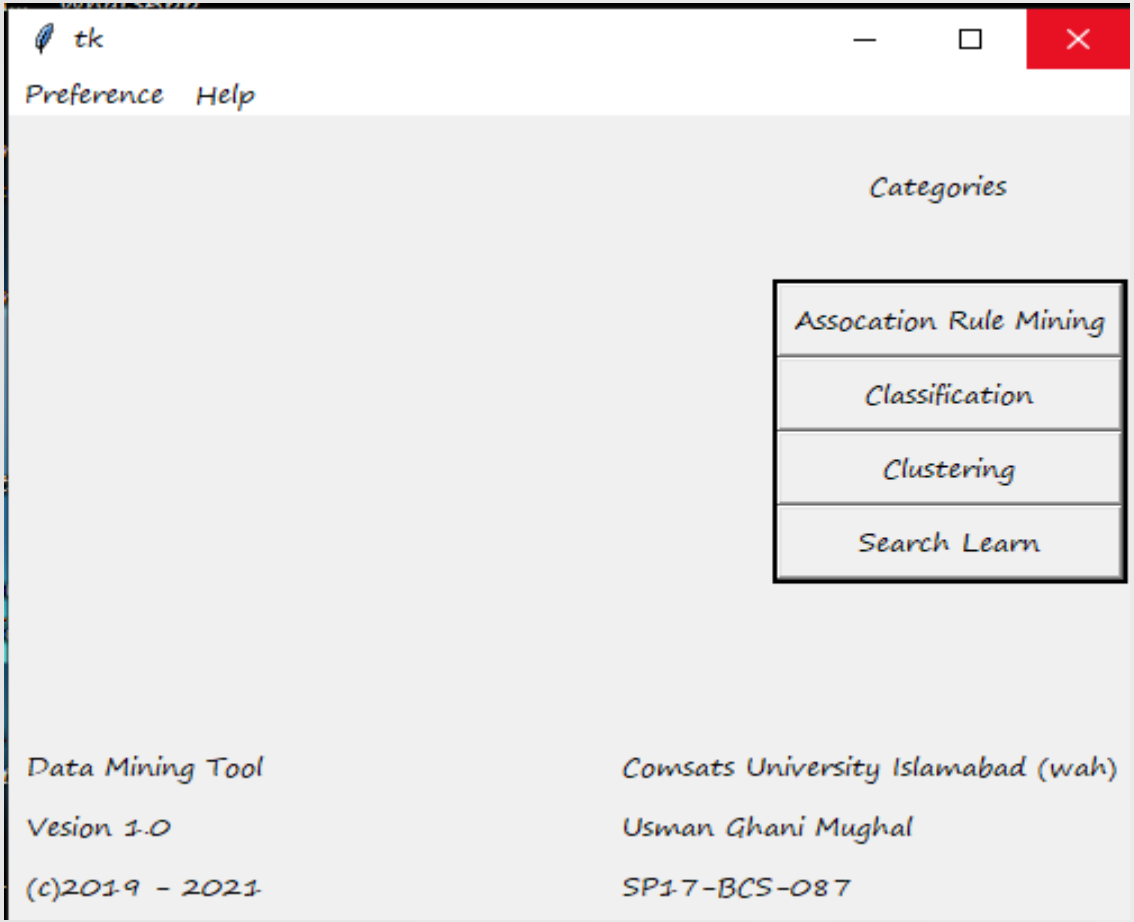
- My Facebook posts data set Generated by my self
- Store/mart tractions data set download from google data set

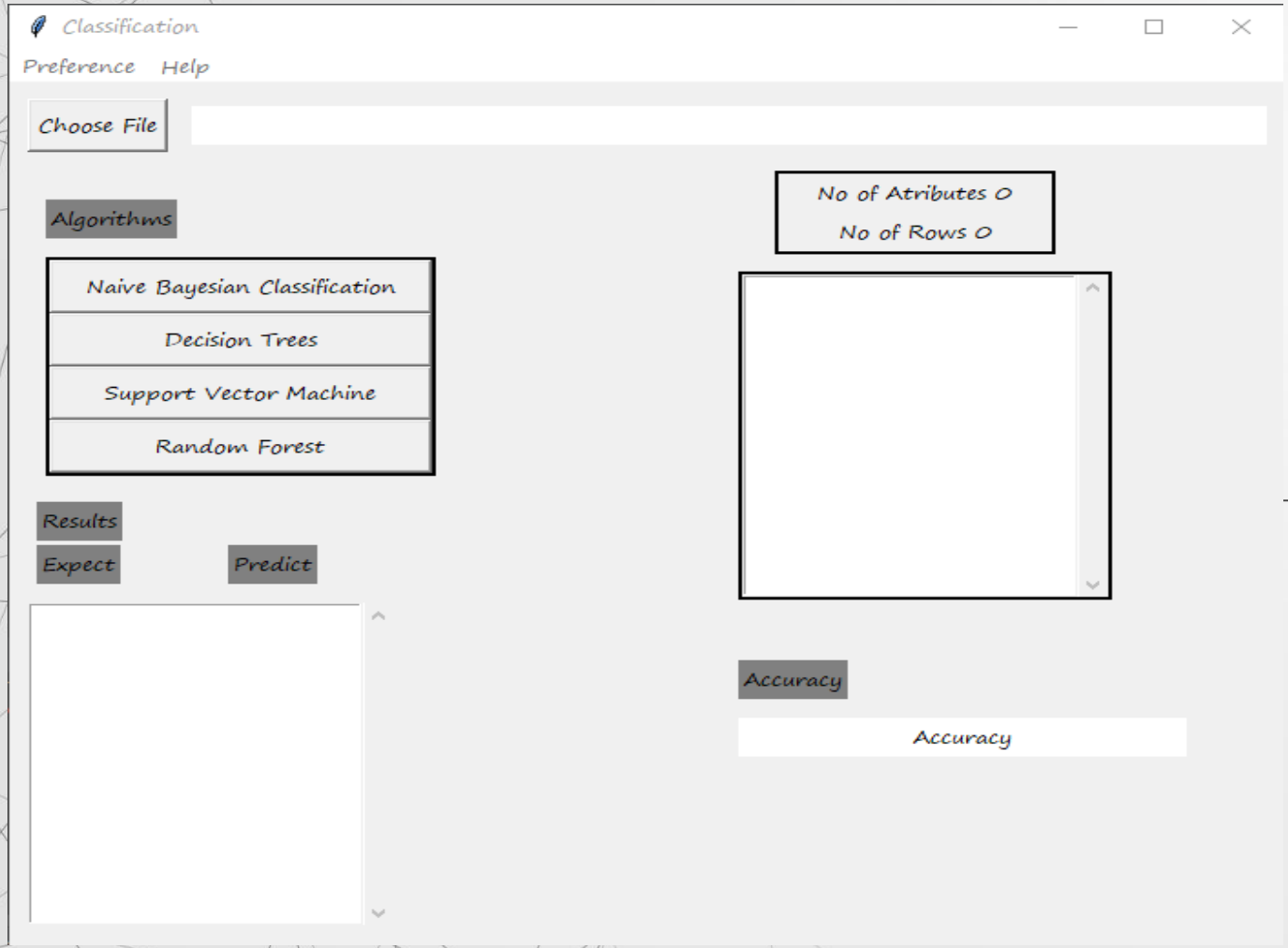
Classification Datasets

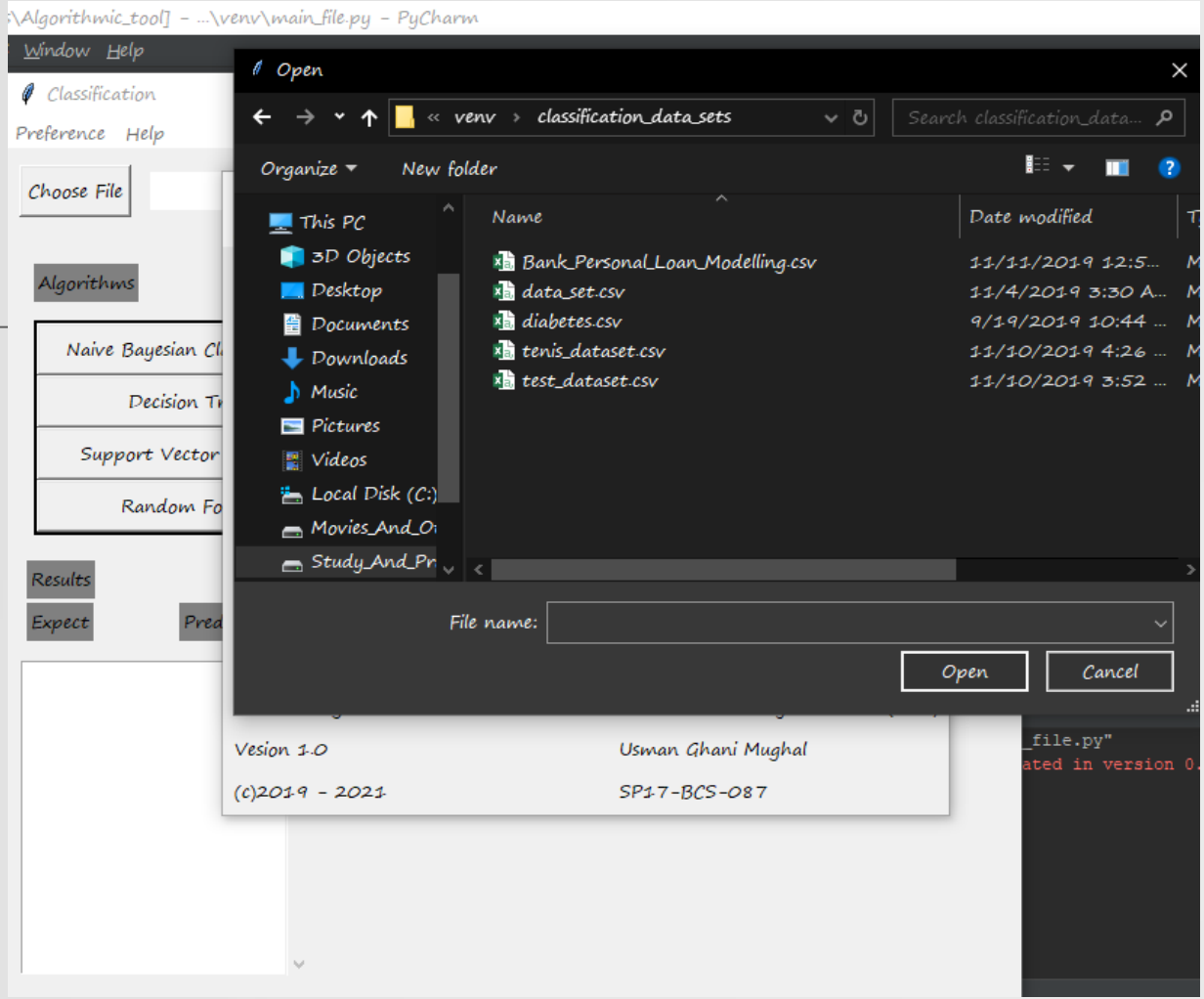
- Leg Before Wicket(LBW) Generated by my self
- Bank Personal Loan Modeling downloaded from Github
- Diabetes data set downloaded from UCI Repository

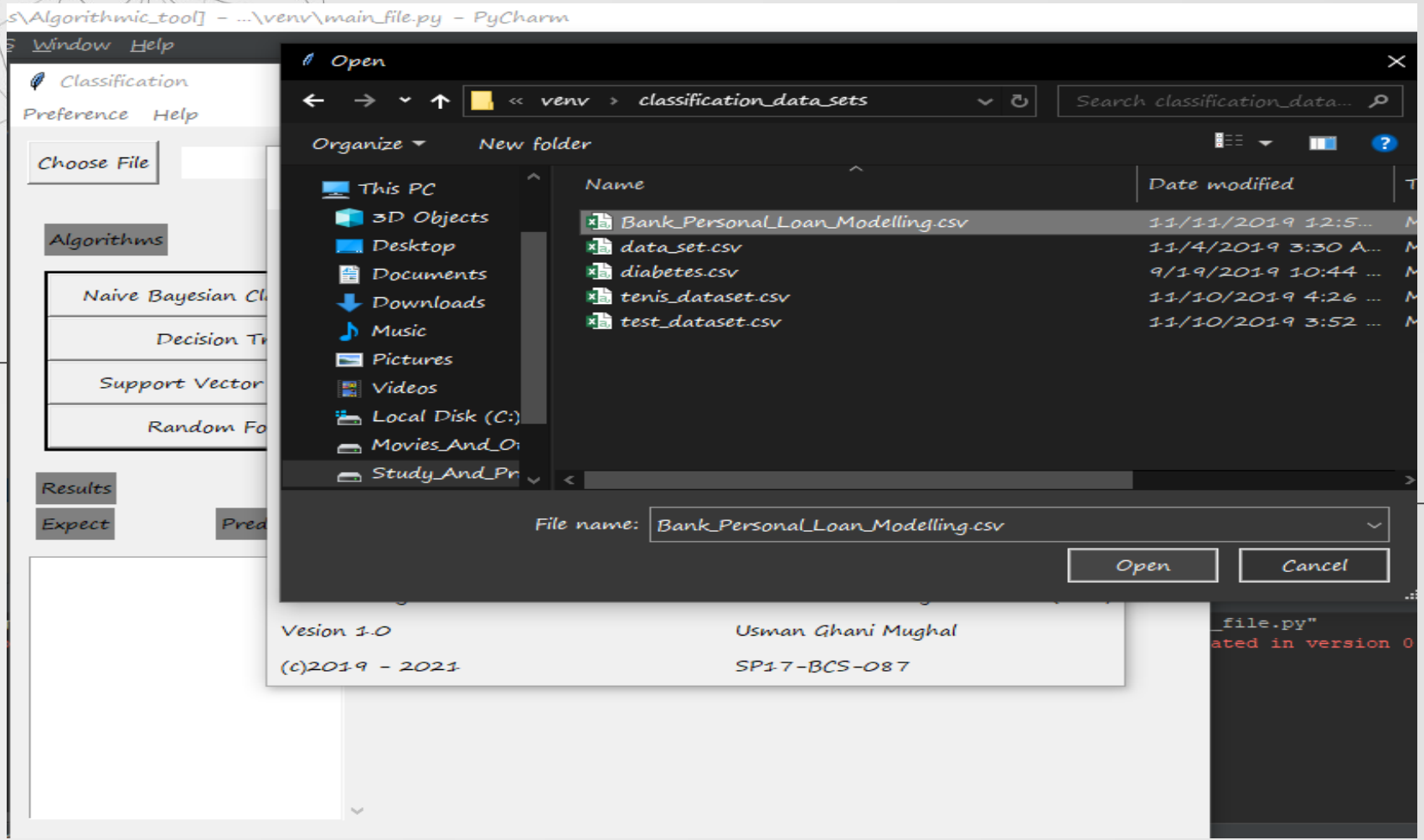


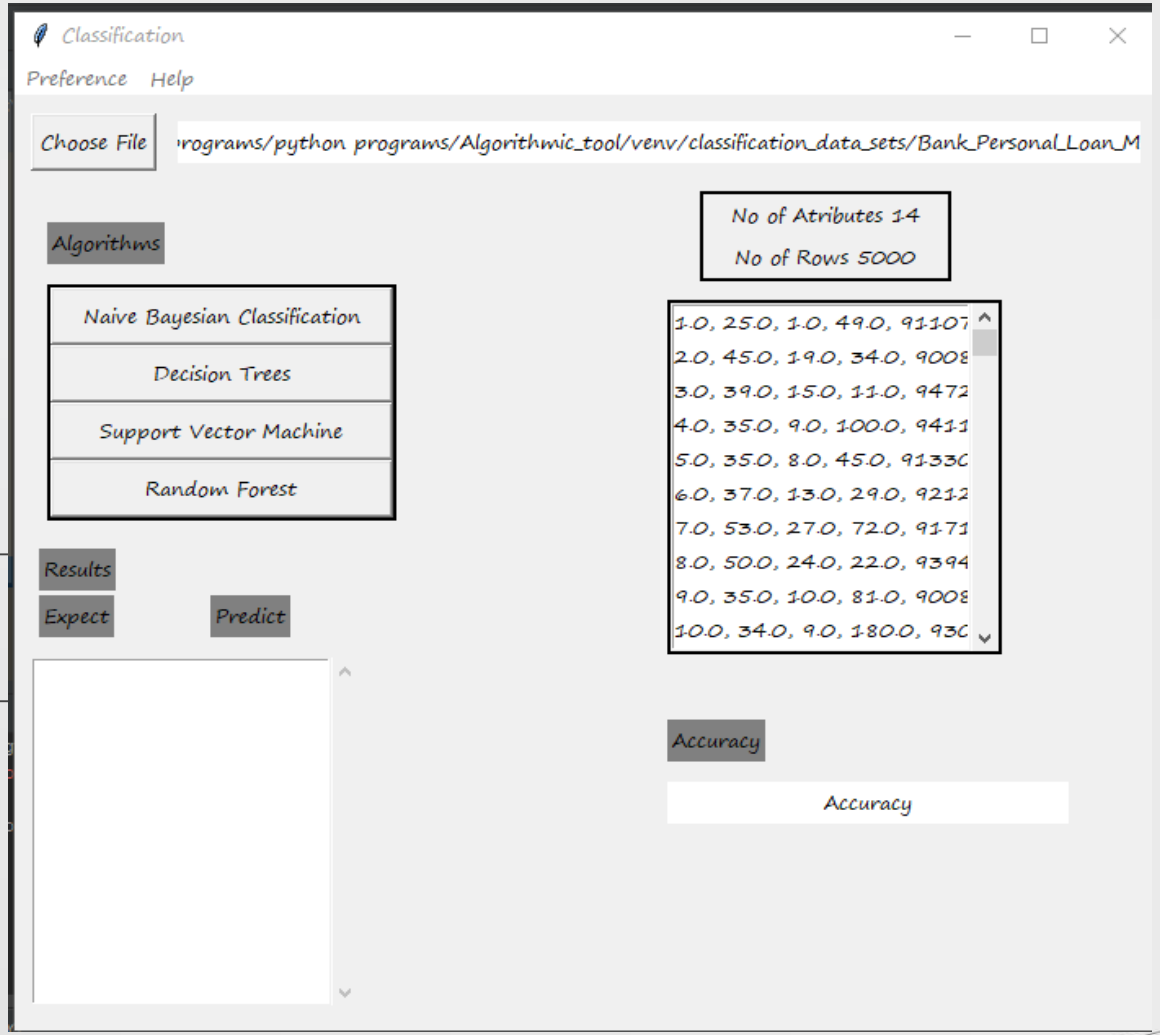
DEMO

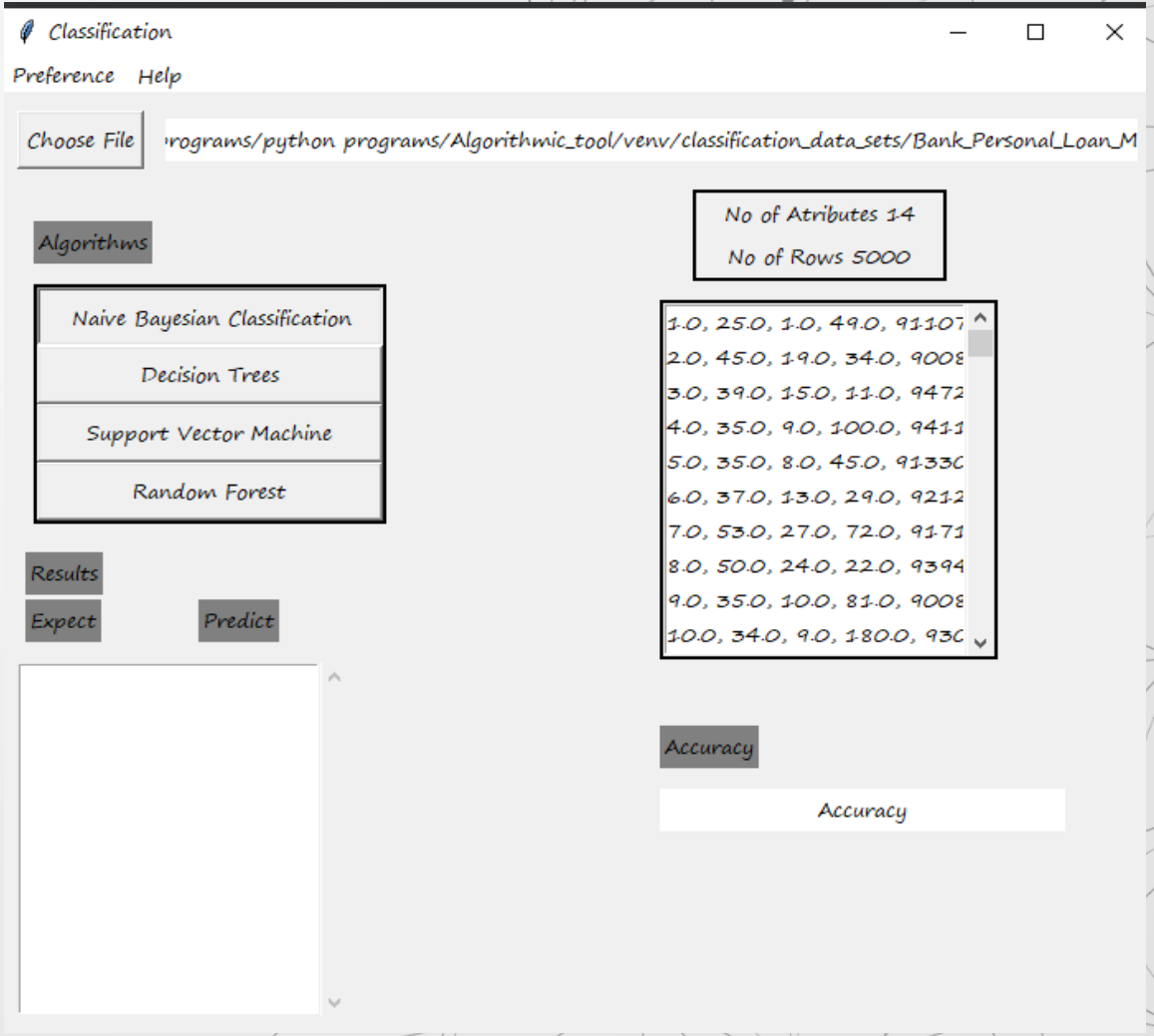

tk
Preference
Help
Categories
Assocation Rule Mining
Classification
Clustering
Search Learn
Data Mining Tool
Vesion 1.0
(c)2019 - 2021
Comsats University Islamabad (wah)
Usman Ghani Mughal
SP17-BCS-087

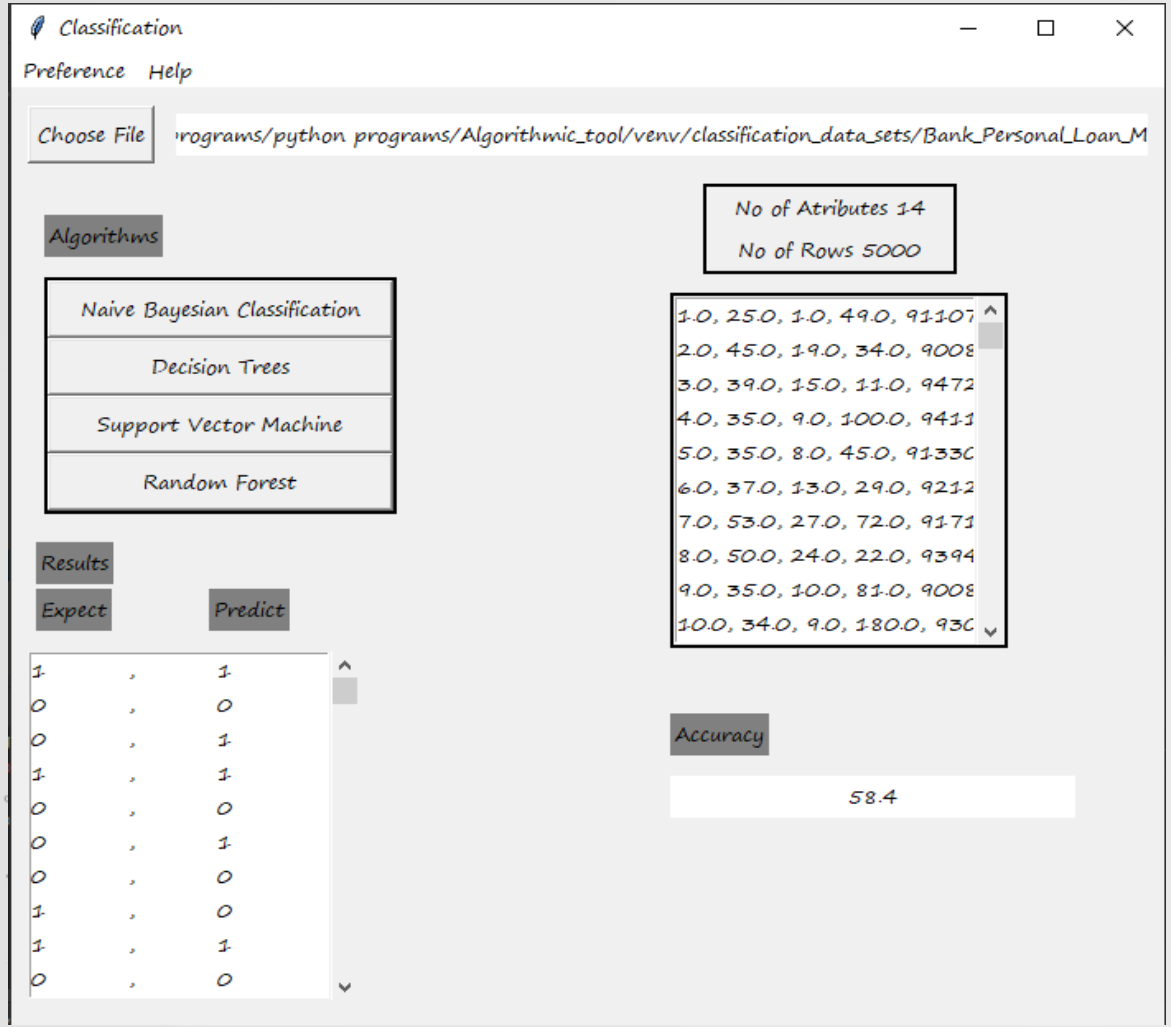

Classification
Preference
Help
Choose File
No of Atributes 0
No of Rows 0
Algorithms
Naive Bayesian Classification
Decision Trees
Support Vector Machine
Random Forest
Results
Expect
Predict
Accuracy
Accuracy

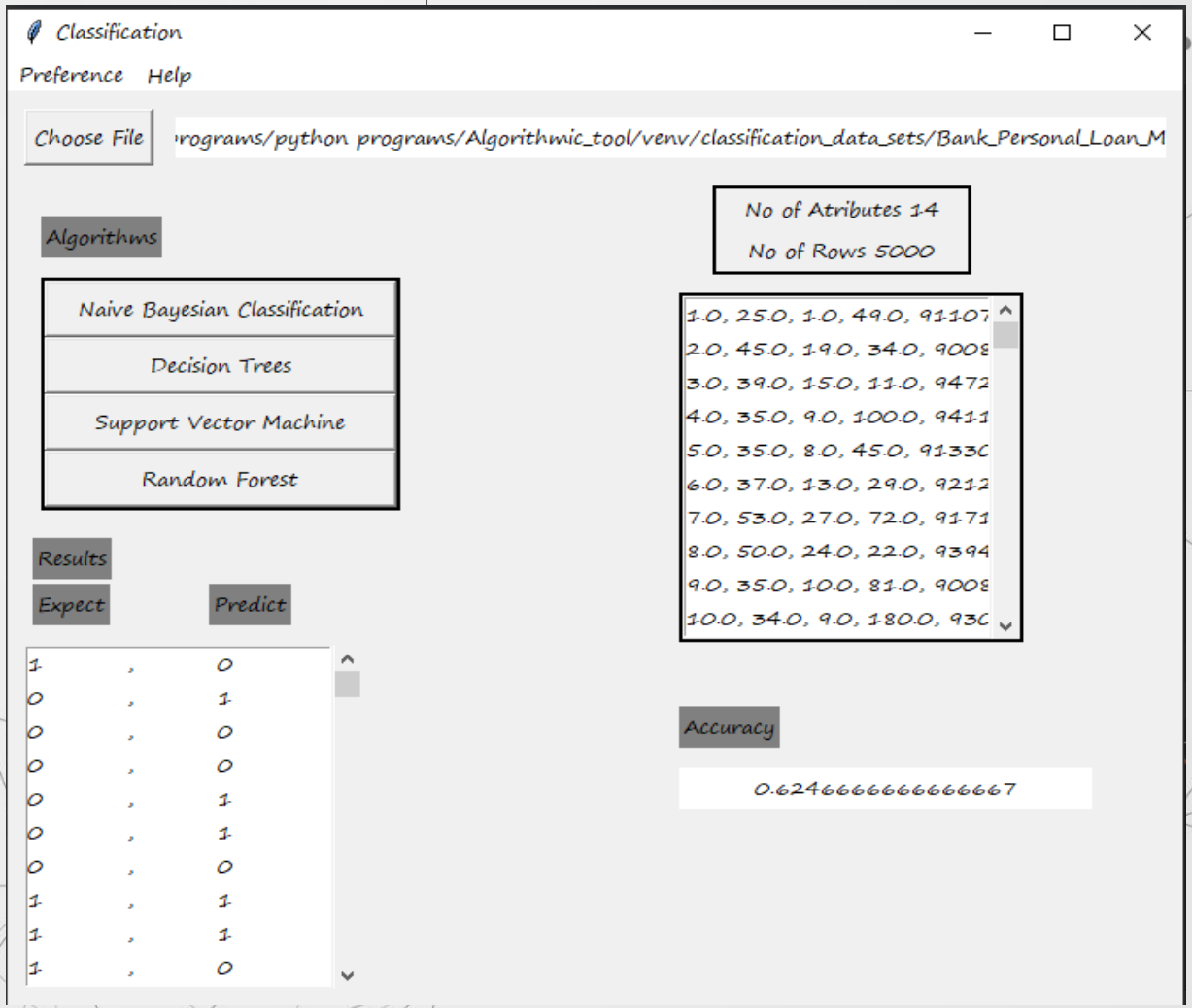

Algorithmic_tool] – ...\venv\main_file.py – PyCharm
Window Help
Classification
Preference Help
Choose File
Algorithms
Naive Bayesian Cl
Decision Tr
Support Vector
Random Fo
Results
Expect
Pred
Open
venv › classification_data_sets
Search classification_data...
Organize
New folder
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
Local Disk (C:)
Movies_And_O
Study_And_Pr
Name
Date modified
Bank_Personal_Loan_Modelling.csv
11/11/2019 12:5...
data_set.csv
11/4/2019 3:30 A...
diabetes.csv
9/19/2019 10:44 ...
tenis_dataset.csv
11/10/2019 4:26 ...
test_dataset.csv
11/10/2019 3:52 ...
File name:
Open
Cancel
Vesion 1.0
Usman Ghani Mughal
(c)2019 - 2021
SP17-BCS-087
_file.py"
ated in version 0.


s\Algorithmic_tool] – ...\venv\main_file.py – PyCharm
Window Help
Classification
Preference Help
Choose File
Algorithms
Naive Bayesian Cl
Decision Tr
Support Vector
Random Fo
Results
Expect
Pred
Open
« venv › classification_data_sets
Search classification_data...
Organize
New folder
This PC
3D Objects
Desktop
Documents
Downloads
Music
Pictures
Videos
Local Disk (C:)
Movies_And_O
Study_And_Pr
Name
Date modified
Bank_Personal_Loan_Modelling.csv 11/11/2019 12:5...
data_set.csv 11/4/2019 3:30 A...
diabetes.csv 9/19/2019 10:44 ...
tenis_dataset.csv 11/10/2019 4:26 ...
test_dataset.csv 11/10/2019 3:52 ...
File name: Bank_Personal_Loan_Modelling.csv
Open
Cancel
Vesion 1.0
Usman Ghani Mughal
(c)2019 - 2021
SP17-BCS-087
_file.py"
ated in version 0

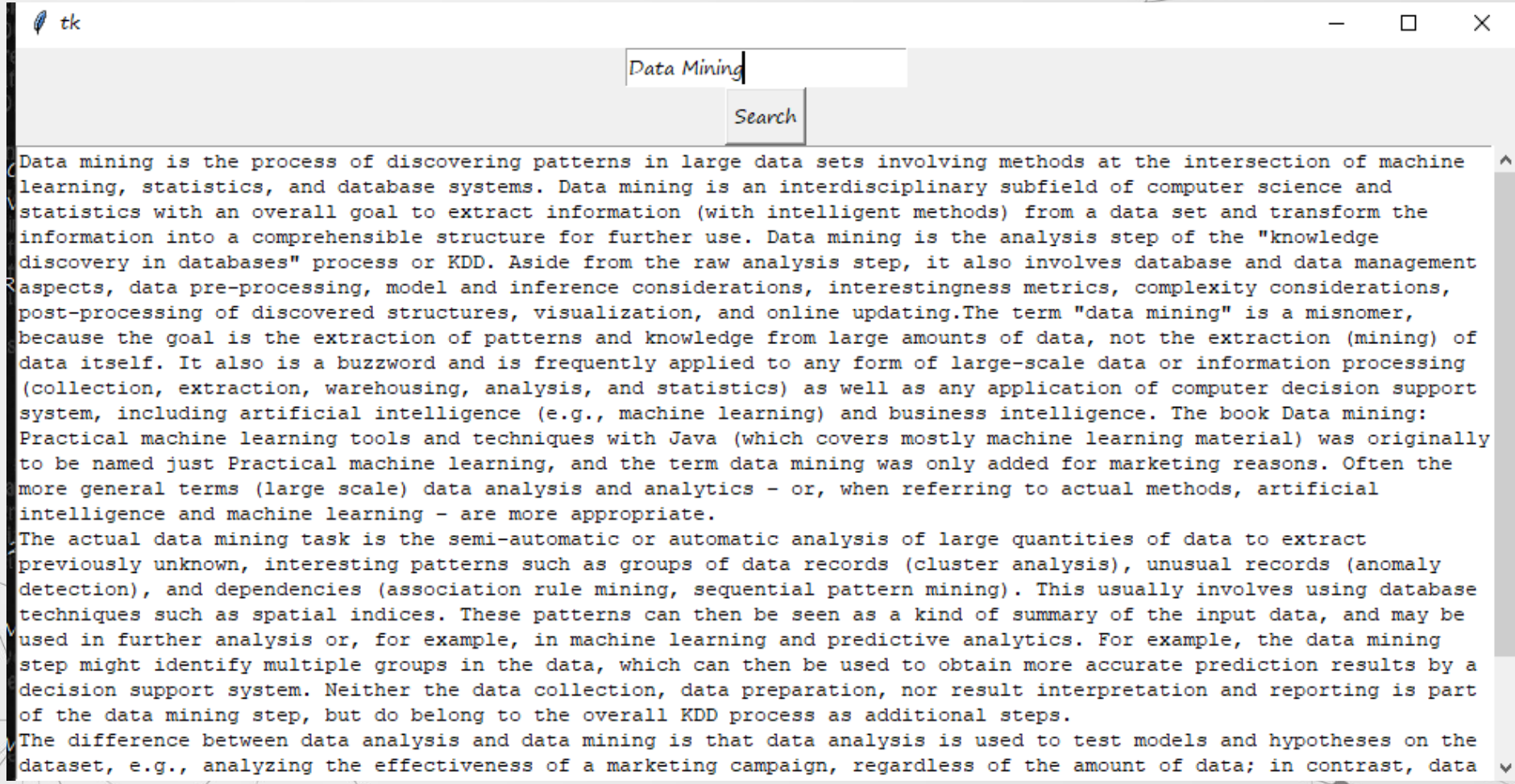

Classification
Preference
Help
Choose File
rograms/python programs/Algorithmic_tool/venv/classification_data_sets/Bank_Personal_Loan_M
No of Atributes 14
No of Rows 5000
Algorithms
Naive Bayesian Classification
Decision Trees
Support Vector Machine
Random Forest
1.0, 25.0, 1.0, 49.0, 91107
2.0, 45.0, 19.0, 34.0, 9008
3.0, 39.0, 15.0, 11.0, 9472
4.0, 35.0, 9.0, 100.0, 9411
5.0, 35.0, 8.0, 45.0, 91330
6.0, 37.0, 13.0, 29.0, 9212
7.0, 53.0, 27.0, 72.0, 9171
8.0, 50.0, 24.0, 22.0, 9394
9.0, 35.0, 10.0, 81.0, 9008
10.0, 34.0, 9.0, 180.0, 930
Results
Expect
Predict
Accuracy
Accuracy


Classification
Preference
Help
Choose File
rograms/python programs/Algorithmic_tool/venv/classification_data_sets/Bank_Personal_Loan_M
No of Atributes 14
No of Rows 5000
Algorithms
Naive Bayesian Classification
Decision Trees
Support Vector Machine
Random Forest
1.0, 25.0, 1.0, 49.0, 91107
2.0, 45.0, 19.0, 34.0, 9008
3.0, 39.0, 15.0, 11.0, 9472
4.0, 35.0, 9.0, 100.0, 9411
5.0, 35.0, 8.0, 45.0, 91330
6.0, 37.0, 13.0, 29.0, 9212
7.0, 53.0, 27.0, 72.0, 9171
8.0, 50.0, 24.0, 22.0, 9394
9.0, 35.0, 10.0, 81.0, 9008
10.0, 34.0, 9.0, 180.0, 930
Results
Expect
Predict
Accuracy
Accuracy


Classification
Preference
Help
Choose File
rograms/python programs/Algorithmic_tool/venv/classification_data_sets/Bank_Personal_Loan_M
No of Atributes 14
No of Rows 5000
Algorithms
Naive Bayesian Classification
Decision Trees
Support Vector Machine
Random Forest
1.0, 25.0, 1.0, 49.0, 91107
2.0, 45.0, 19.0, 34.0, 9008
3.0, 39.0, 15.0, 11.0, 9472
4.0, 35.0, 9.0, 100.0, 9411
5.0, 35.0, 8.0, 45.0, 91330
6.0, 37.0, 13.0, 29.0, 9212
7.0, 53.0, 27.0, 72.0, 9171
8.0, 50.0, 24.0, 22.0, 9394
9.0, 35.0, 10.0, 81.0, 9008
10.0, 34.0, 9.0, 180.0, 930
Results
Expect
Predict
1 , 1
0 , 0
0 , 1
1 , 1
0 , 0
0 , 1
0 , 0
1 , 0
1 , 1
0 , 0
Accuracy
58.4


Classification
Preference
Help
Choose File
rograms/python programs/Algorithmic_tool/venv/classification_data_sets/Bank_Personal_Loan_M
No of Atributes 14
No of Rows 5000
Algorithms
Naive Bayesian Classification
Decision Trees
Support Vector Machine
Random Forest
1.0, 25.0, 1.0, 49.0, 91107
2.0, 45.0, 19.0, 34.0, 9008
3.0, 39.0, 15.0, 11.0, 9472
4.0, 35.0, 9.0, 100.0, 9411
5.0, 35.0, 8.0, 45.0, 91330
6.0, 37.0, 13.0, 29.0, 9212
7.0, 53.0, 27.0, 72.0, 9171
8.0, 50.0, 24.0, 22.0, 9394
9.0, 35.0, 10.0, 81.0, 9008
10.0, 34.0, 9.0, 180.0, 930
Results
Expect
Predict
1 , 0
0 , 1
0 , 0
0 , 0
0 , 1
0 , 1
0 , 0
1 , 1
1 , 1
1 , 0
Accuracy
0.6246666666666667


tk
Data Mining
Search


tk
Data Mining
Search
Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.The term "data mining" is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (mining) of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence (e.g., machine learning) and business intelligence. The book Data mining: Practical machine learning tools and techniques with Java (which covers mostly machine learning material) was originally to be named just Practical machine learning, and the term data mining was only added for marketing reasons. Often the more general terms (large scale) data analysis and analytics – or, when referring to actual methods, artificial intelligence and machine learning – are more appropriate.
The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step, but do belong to the overall KDD process as additional steps.
The difference between data analysis and data mining is that data analysis is used to test models and hypotheses on the dataset, e.g., analyzing the effectiveness of a marketing campaign, regardless of the amount of data; in contrast, data

Conclusion

From my perspective the only way to understand the deep concepts and logic of Algorithms is to write a code of it from Scratch

This Product is only a prototype, A lot of work is left on it yet.

Thanks.

References

- https://archive.ics.uci.edu/ml/datasets/diabetes
- https://toolbox.google.com/datasetsearch

- https://www.youtube.com/watch?v=2znbn8_efVc&list=PLjC8JXsSUrri0XWbCGffJ5to1P40hebu2
- https://stackoverflow.com/users/10096169/usman-ghani-mughal
- https://github.com/