CA-02 DATA MINING

DEEPAKVANNSHIKUMAR TUNIYA

10512387

Distinguish between rocks and mines using Sonar Signal Prediction.

INTRODUCTION:

- We predict the best accurate model for this project by using sonar signals to discriminate them testing which bounces off by metal cylinders such as mines and bounce off by roughly rock cylinders.
- We apply CRISP-DM methodology for the modelling of our project which stands as CRoss Industry Standard Project for Data Mining.
- The model is robust and powerful.
- CRISP-DM explains the life cycle of the Data Mining Project.

Business Problem Framing :

- Sonar signals are used by Navy to know if there are any obstacles baring their path.
- In this objective we need to discriminate between the metal structures such as sea mines and rock structures on the sea bed.
- The sonar signals are tested for the Cold War on the sandy ocean floor.
- Distinguish between mines and rocks is very important as we need to know that the sonar signal is received from which of the geological material on the sea floor .

- Key stakeholders are Navy Ships used for Cold War.

Analytics solution is suitable for this problem. Supervised learning includes variety of data mining and machine learning statistical techniques that analyze records for both regression and classification of the labeled and unlabeled data in the CRISP-DM methodology throughout. Support Vector Machine is the proposed algorithm.

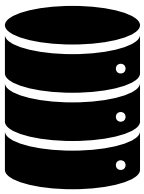
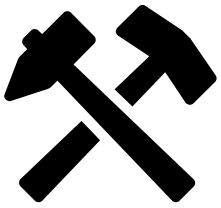
- It is difficult to classify the Support Vectors.
- The process is directly affected to find optimum location of the boundaries.
- The decision function is fully specified on a very small subset of training sample SVM.

- Constrained optimization is appreciated. In optimization we can find the support vectors which are maximally separated and Constrained as the support vectors should be away.

We need to create a model which is able to predict the sonar signals of rocks and mines so that in each signal it can identify whether the object is rock or mine.

Data Understanding:

Resources used for the data mining project:


DATA SET -Sonar Signals **TOOL – Rapidminer** **PROGRAMMING – R**

Data Source:

- Data is in CSV format.
- It is a sample dataset with 208 (examples) rows and 61 (attributes) columns .
- The dataset is Multivariate.
- It has 60 columns with each column having set of ranges from 0.0 to 1.0. The range represents energy with particular frequency band, integrated over certain period of time.
- If the object is rock it is labelled as ‘R’ for each record.

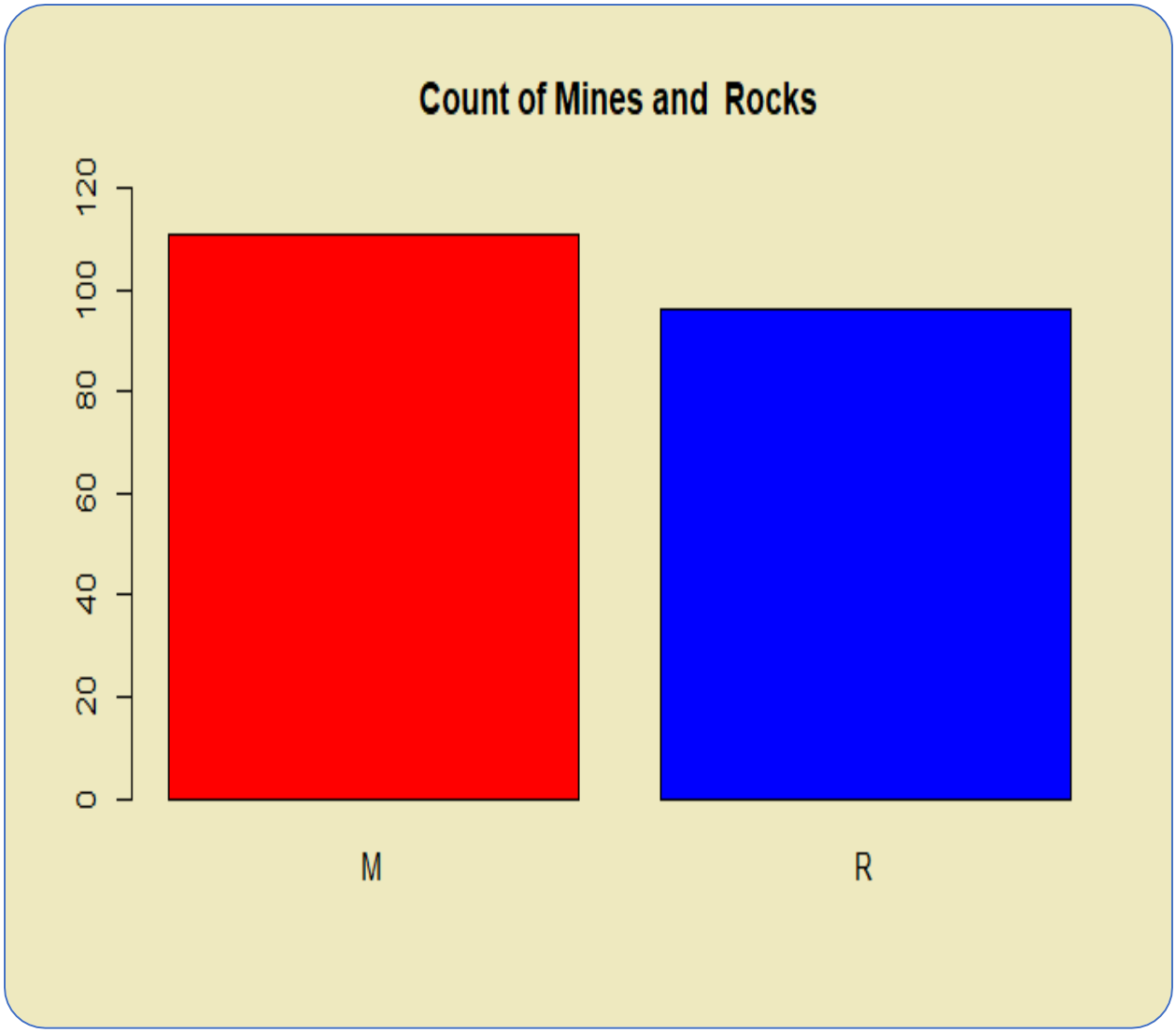
- If the object is mine it is labelled as ‘M’ for each record.

Data Need:

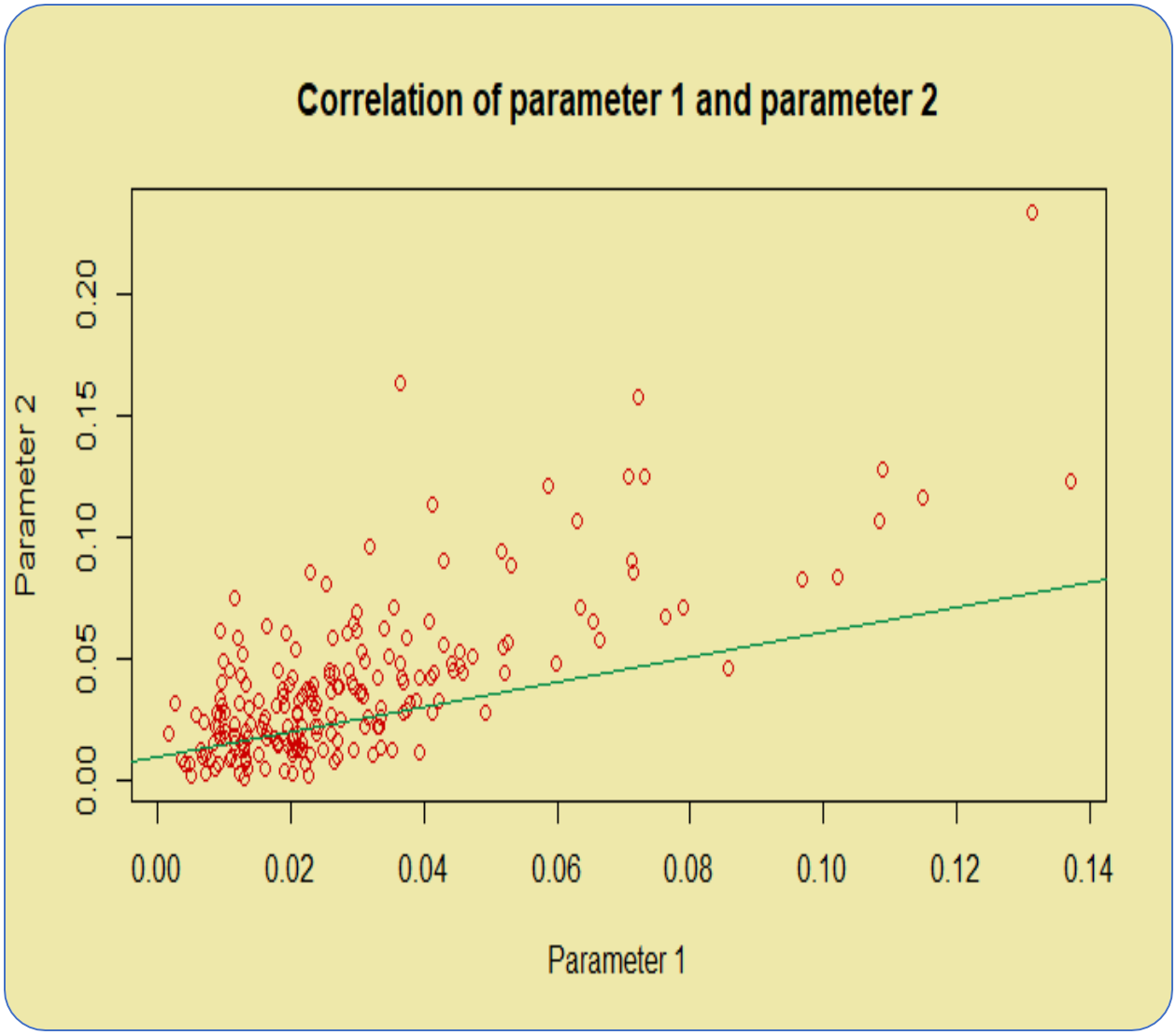
- Using this dataset it helps to plan and predict whether the obstacle received by the sonar signal is rock or mine.
- With the machine learning algorithm and supervised learning techniques we can differentiate each record with accurate prediction.

- After understanding of data it is important ro acquire data.
- When the data is acquired we can search for the business we need to apply on it.
- We can work on the acquired data and apply the business solutions and provide the business insights which we implement on the dataset

- After acquiring the data we now explore the data.
- The graphical presentation of the data which is done in R Studio.
- The graph shows count of rocks and mines which will be used as prediction column in the performance of the model.
- There are 111 Mines and 97 Rocks in the given column.

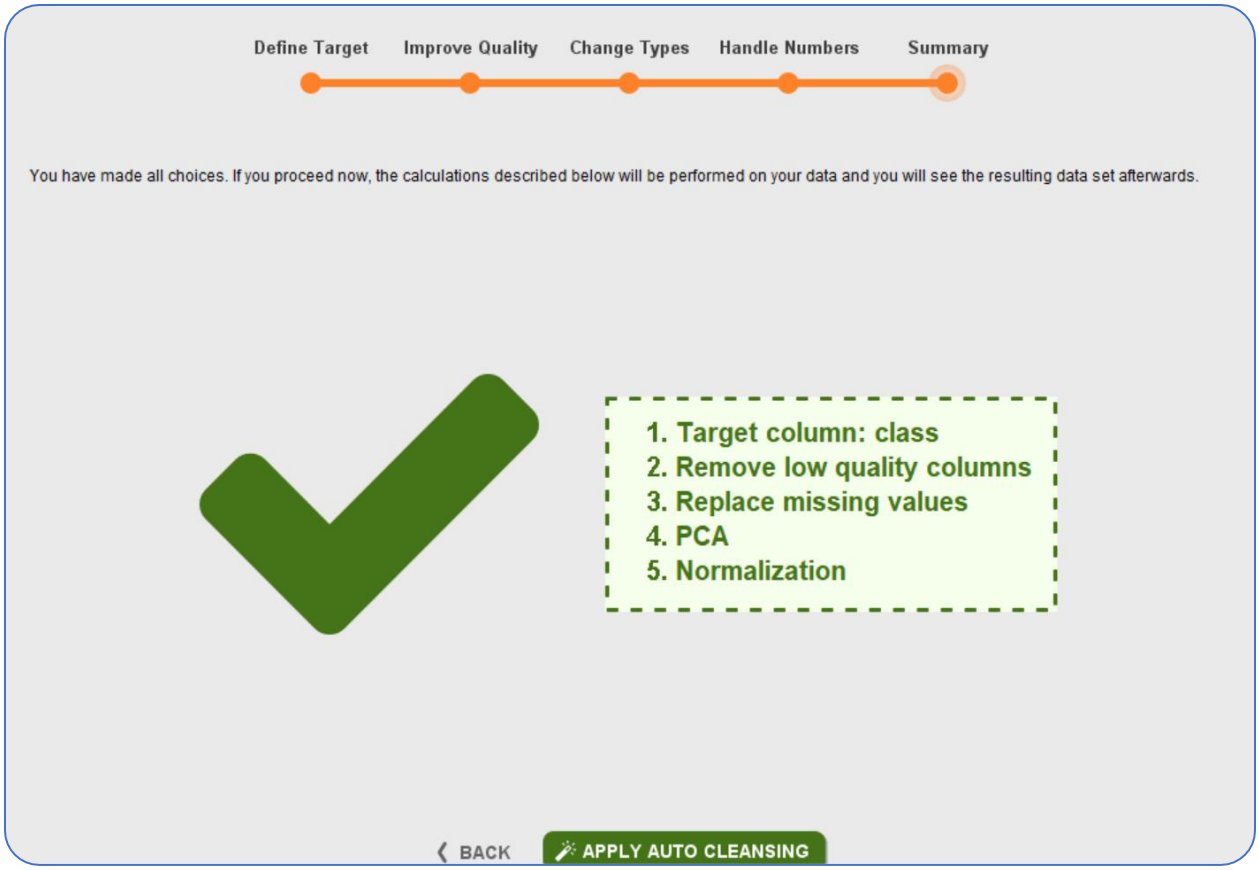


- This graph is a scatter plot presentation of column 1 and column 2 of the dataset.
- The graph plots all the data of parameter 1 and parameter 2 which is separated by the correlation.



- This is a glimpse of our dataset.
- It shows 60 attributes which range from 0.0 to 1.0.
- And last column which is column 61 it shows rocks and mines as ‘R’ and ‘M’.

0.2825	0.4256	0.2641	0.1386	0.1051	0.1343	0.0383	0.0324	0.0232	0.0027	0.0065	0.0159	0.0072	0.0167	0.018	0.0084	0.009	0.0032	R
0.1401	0.1628	0.0621	0.0203	0.053	0.0742	0.0409	0.0061	0.0125	0.0084	0.0089	0.0048	0.0094	0.0191	0.014	0.0049	0.0052	0.0044	R
0.2129	0.2222	0.2111	0.0176	0.1348	0.0744	0.013	0.0106	0.0033	0.0232	0.0166	0.0095	0.018	0.0244	0.0316	0.0164	0.0095	0.0078	R
0.321	0.3202	0.4295	0.3654	0.2655	0.1576	0.0681	0.0294	0.0241	0.0121	0.0036	0.015	0.0085	0.0073	0.005	0.0044	0.004	0.0117	R
0.1847	0.0841	0.0692	0.0528	0.0357	0.0085	0.023	0.0046	0.0156	0.0031	0.0054	0.0105	0.011	0.0015	0.0072	0.0048	0.0107	0.0094	R
0.0723	0.1238	0.1192	0.1089	0.0623	0.0494	0.0264	0.0081	0.0104	0.0045	0.0014	0.0038	0.0013	0.0089	0.0057	0.0027	0.0051	0.0062	R
0.1957	0.1749	0.1304	0.0597	0.1124	0.1047	0.0507	0.0159	0.0195	0.0201	0.0248	0.0131	0.007	0.0138	0.0092	0.0143	0.0036	0.0103	R
0.1879	0.1437	0.2146	0.236	0.1125	0.0254	0.0285	0.0178	0.0052	0.0081	0.012	0.0045	0.0121	0.0097	0.0085	0.0047	0.0048	0.0053	R
0.2121	0.1099	0.0985	0.1271	0.1459	0.1164	0.0777	0.0439	0.0061	0.0145	0.0128	0.0145	0.0058	0.0049	0.0065	0.0093	0.0059	0.0022	R
0.1597	0.1384	0.0372	0.0688	0.0867	0.0513	0.0092	0.0198	0.0118	0.009	0.0223	0.0179	0.0084	0.0068	0.0032	0.0035	0.0056	0.004	R
0.0793	0.1269	0.1533	0.069	0.0402	0.0534	0.0228	0.0073	0.0062	0.0062	0.012	0.0052	0.0056	0.0093	0.0042	0.0003	0.0053	0.0036	R
0.0516	0.0337	0.0894	0.0861	0.0872	0.0445	0.0134	0.0217	0.0188	0.0133	0.0265	0.0224	0.0074	0.0118	0.0026	0.0092	0.0009	0.0044	R
0.1391	0.0819	0.0678	0.0663	0.1202	0.0692	0.0152	0.0266	0.0174	0.0176	0.0127	0.0088	0.0098	0.0019	0.0059	0.0058	0.0059	0.0032	R
0.2195	0.3051	0.1937	0.157	0.0479	0.0538	0.0146	0.0068	0.0187	0.0059	0.0095	0.0194	0.008	0.0152	0.0158	0.0053	0.0189	0.0102	R
0.2351	0.2298	0.1155	0.0724	0.0621	0.0318	0.045	0.0167	0.0078	0.0083	0.0057	0.0174	0.0188	0.0054	0.0114	0.0196	0.0147	0.0062	R
0.0526	0.1867	0.1553	0.1633	0.1252	0.0748	0.0452	0.0064	0.0154	0.0031	0.0153	0.0071	0.0212	0.0076	0.0152	0.0049	0.02	0.0073	R
0.3431	0.1803	0.2378	0.3424	0.2303	0.0689	0.0216	0.0469	0.0426	0.0346	0.0158	0.0154	0.0109	0.0048	0.0095	0.0015	0.0073	0.0067	R
0.2709	0.1419	0.126	0.1288	0.079	0.0829	0.052	0.0216	0.036	0.0331	0.0131	0.012	0.0108	0.0024	0.0045	0.0037	0.0112	0.0075	R
0.2207	0.1778	0.1353	0.1373	0.0749	0.0472	0.0325	0.0179	0.0045	0.0084	0.001	0.0018	0.0068	0.0039	0.012	0.0132	0.007	0.0088	R
0.1266	0.1381	0.1136	0.0516	0.0073	0.0278	0.0372	0.0121	0.0153	0.0092	0.0035	0.0098	0.0121	0.0006	0.0181	0.0094	0.0116	0.0063	R



- The unorganized data in this stage is organized by performing transformation to the data.
- It is time consuming and tedious task,
- Process of preparation for the data:

1. Data Cleaning
2. Data Construction
3. Data integration and formation.

- Data Cleaning Results:

1. Scaling of the independent attributes.
1. Remove low quality columns.
2. There are no missing values.

Methodology:

CRISP_DM method is used for the problem solving. Reasons for taking this method:

- Adoption of the method is easy
- It is result oriented.
- The method is best analytical approach.

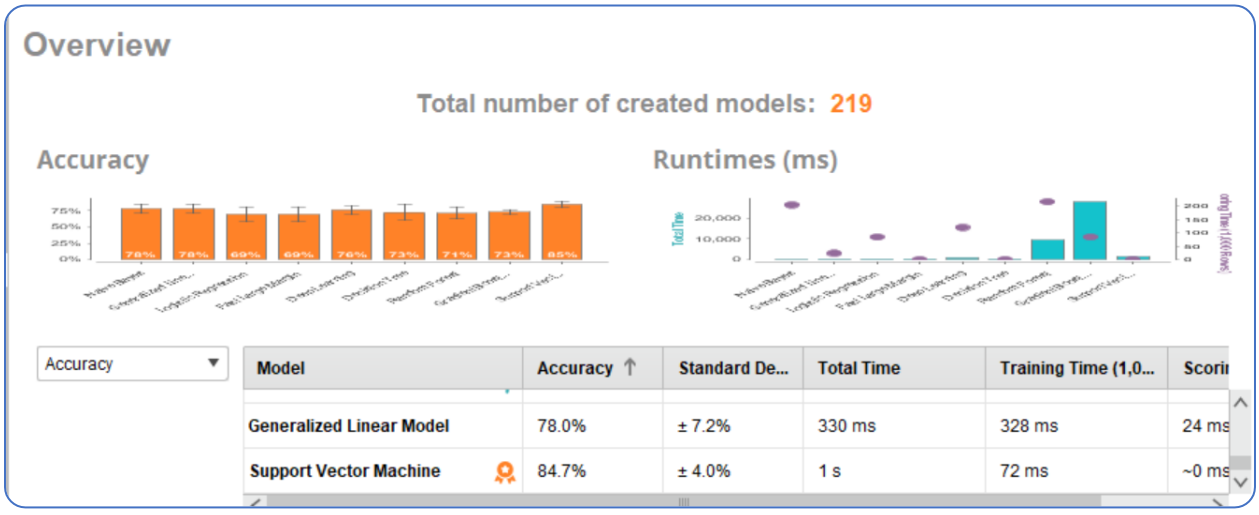
We apply Auto Model technique to view the performance of every model.

Model Building:

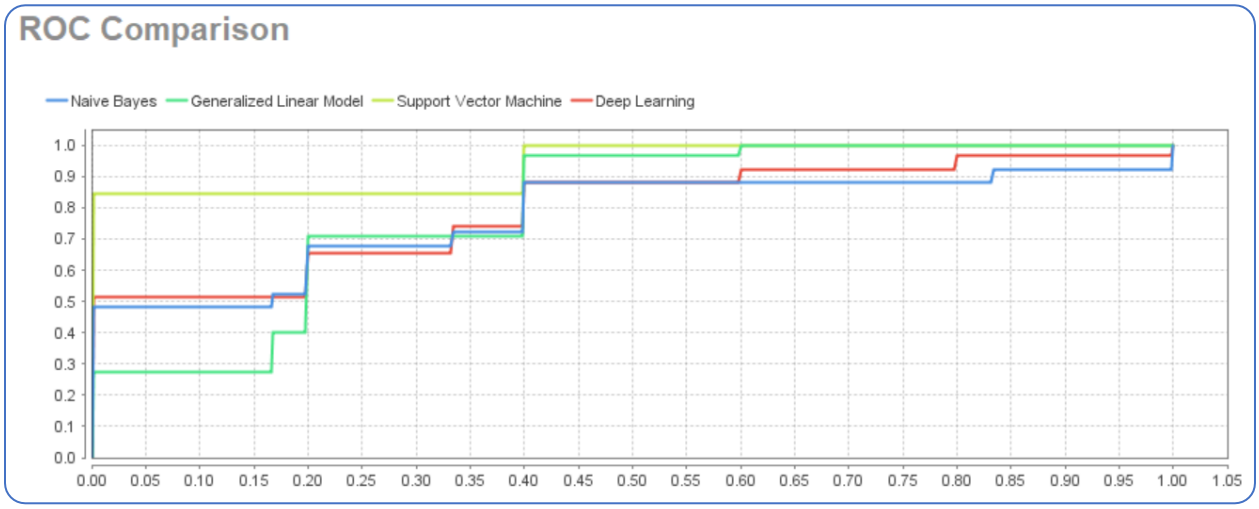
As the data is cleaned and prepared we can now move to building of the model.

- The 4 models we will be building are:

1. Support Vector Machine (SVM)
2. Generalized Linear Model (GLM)
3. Naïve Bayes
4. Deep Learning.



After performing Auto Model technique we see that Support Vector Machine (SVM) is the best performing model.



- Relative Operating Characteristics (ROC) curve shows the comparison of performance and quality of the various models.

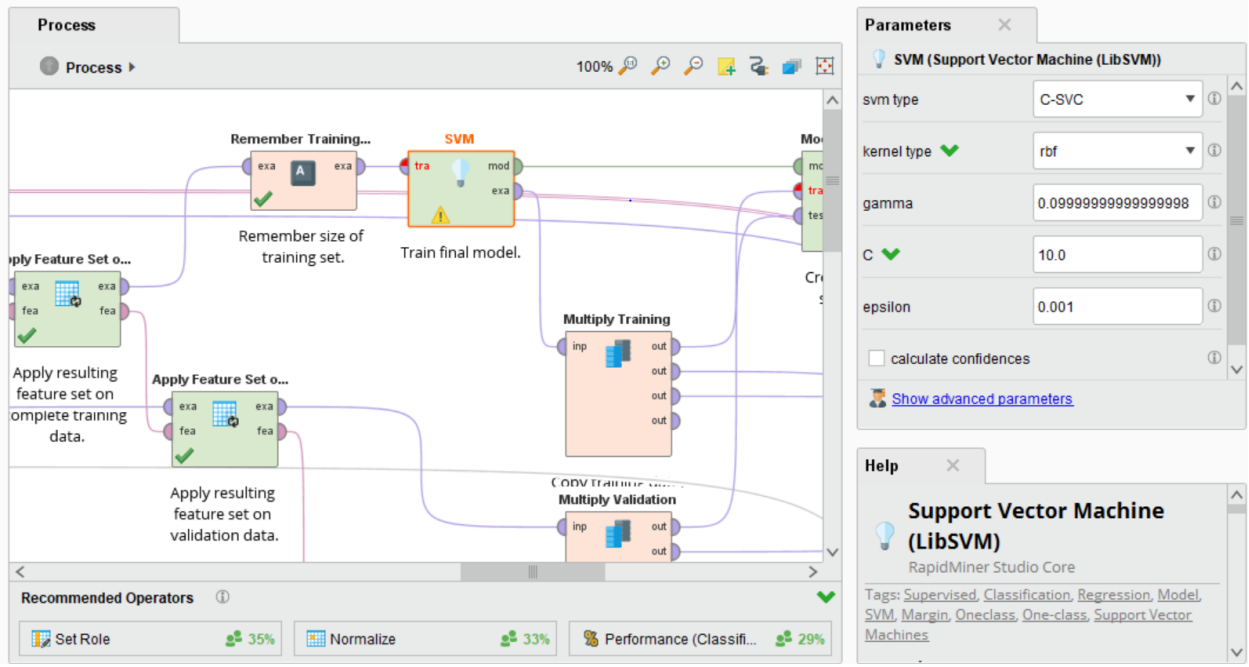
accuracy: 84.70% +/- 3.95% (micro average: 84.75%)

	true Rock	true Mine	class precision
pred. Rock	27	9	75.00%
pred. Mine	0	23	100.00%
class recall	100.00%	71.88%	

The above image shows performance accuracy of Support Vector Machine (SVM).

The accuracy of SVM is best result among the considered models.

Based on performance results has 84.70% accuracy where the true prediction of Mines is 100% accurate and true prediction of Rock is 75%.



SVM has different math functions sets which are called as kernel.

There are different types of kernels such as:

Linear,non-linear,polynomial,sigmoid and radial basis functions(RBF).

Radial basis function is most used kernel.

In our model also Kernel type used is RBF and gamma is 0.99.

It has finite and localized response on the whole axis.

class	prediction(cl...	confidence(...	confidence(...
Mine	Rock	0.515	0.485
Mine	Rock	0.516	0.484
Mine	Mine	0.477	0.523
Mine	Mine	0.319	0.681
Mine	Mine	0.327	0.673
Mine	Mine	0.402	0.598
Mine	Mine	0.413	0.587
Mine	Mine	0.388	0.612

This shows prediction of the Support Vector Machine (SVM) where it show if the prediction is same as the class and defines the comnfidence level which are eventually low.

The model is applied on the real world use cases problem solving.

Deployment:

- A prediction report has been generated on the real world problem.
- We achieve a clear understanding of the objectives achieved for the stakeholders.
- The model is monitored and maintained.
- In the regular intervals we check the accuracy of the model with new facts according to the plan.
- If there is change or addition of new parameters we can change the model.

Conclusion:

- We have experienced different stage of CRISP-DM.
- We see how the accuracy of the models and tasks performed on the CRISP-DM.
- By applying SVM we can see the results and post analysis above that we have achieved our business solution and the model is satisfactory.
- The model generates is suitable with and improves step by step.
- We successfully achieved the objectives defined in the CRISP-DM methodology.

Bibliography:

- https://www.kaggle.com/adx891/sonar-data-set
- https://elearning.dbs.ie/pluginfile.php/840730/mod_resource/content/0/CA01%20Part%20A%20-%20The%20CRISP-DM%20Model%20-%20the%20New%20Blueprint%20for%20Data%20Mining%20-%20Colin%20Shearer%20-%20Fall%202000.pdf