Development of Statistical and Machine Learning App

Manual Description

Ankita Narvekar
Arvind Parameswaran
Jyoti Narnolia
Nidhi Shivgan
Varsha GanapathyRao

Table of Contents

Description of the App

The app is suitable for Statistical and Machine Learning analytics. It allows the user to do descriptive and predictive analytics on a real dataset.

This app contains four stages:

1. Descriptive Analysis
2. Discrete Models
3. Continuous Models
4. Classification Models

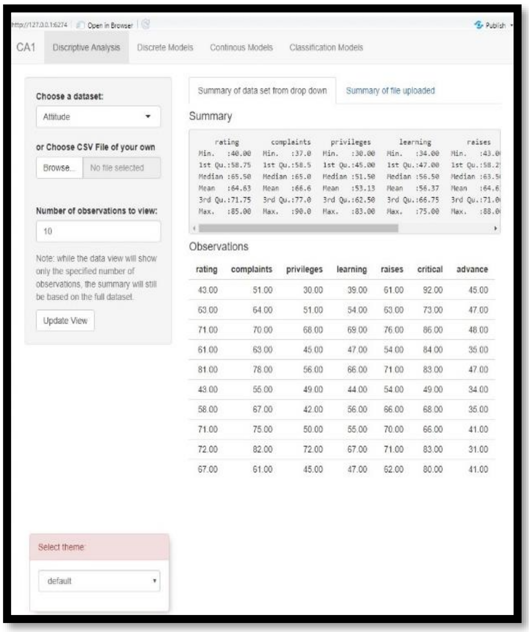
1. Descriptive Analysis

Descriptive statistics are used to summarize data in a way that provides insight into the information contained in the data. This might include examining the mean or median of numeric data or the frequency of observations for nominal data.

In the app, the user has the benefit of choosing from the existing datasets (from - Attitude, Cars, Islands, Rock, Seatbelts) as well as user can choose his own'.CSV' file from his system.

After uploading the file from the system, the user needs to wait until the upload is completed and a message is shown just below the uploaded file.

The user can change the number of observations to view as per his requirements. After changing or uploading the settings, press 'Update View' to update the existing summary. (Refer images)



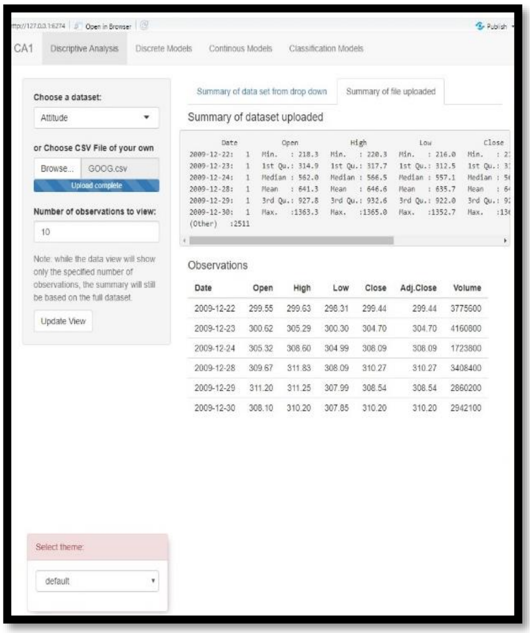
Note: while the data view will show only the specified number of observations, the summary will still be based on the full dataset.

Summary

```
     rating          complaints      privileges        learning          raises
 Min.   :40.00   Min.   :37.0   Min.   :30.00   Min.   :34.00   Min.   :43.0
 1st Qu.:58.75   1st Qu.:58.5   1st Qu.:45.00   1st Qu.:47.00   1st Qu.:58.2
 Median :65.50   Median :65.0   Median :51.50   Median :56.50   Median :63.5
 Mean   :64.63   Mean   :66.6   Mean   :53.13   Mean   :56.37   Mean   :64.6
 3rd Qu.:71.75   3rd Qu.:77.0   3rd Qu.:62.50   3rd Qu.:66.75   3rd Qu.:71.0
 Max.   :85.00   Max.   :90.0   Max.   :83.00   Max.   :75.00   Max.   :88.0
```

Observations

rating	complaints	privileges	learning	raises	critical	advance
43.00	51.00	30.00	39.00	61.00	92.00	45.00
63.00	64.00	51.00	54.00	63.00	73.00	47.00
71.00	70.00	68.00	69.00	76.00	86.00	48.00
61.00	63.00	45.00	47.00	54.00	84.00	35.00
81.00	78.00	56.00	66.00	71.00	83.00	47.00
43.00	55.00	49.00	44.00	54.00	49.00	34.00
58.00	67.00	42.00	56.00	66.00	68.00	35.00
71.00	75.00	50.00	55.00	70.00	66.00	41.00
72.00	82.00	72.00	67.00	71.00	83.00	31.00
67.00	61.00	45.00	47.00	62.00	80.00	41.00



Note: while the data view will show only the specified number of observations, the summary will still be based on the full dataset.

Summary of dataset uploaded

```
         Date          Open             High              Low             Close
 2009-12-22:   1   Min.   : 218.3   Min.   : 220.3   Min.   : 216.0   Min.   : 2
 2009-12-23:   1   1st Qu.: 314.9   1st Qu.: 317.7   1st Qu.: 312.5   1st Qu.: 3
 2009-12-24:   1   Median : 562.0   Median : 566.5   Median : 557.1   Median : 5
 2009-12-28:   1   Mean   : 641.3   Mean   : 646.6   Mean   : 635.7   Mean   : 6
 2009-12-29:   1   3rd Qu.: 927.8   3rd Qu.: 932.6   3rd Qu.: 922.0   3rd Qu.: 9
 2009-12-30:   1   Max.   :1363.3   Max.   :1365.0   Max.   :1352.7   Max.   :13
 (Other)   :2511
```

Observations

Date	Open	High	Low	Close	Adj.Close	Volume
2009-12-22	299.55	299.63	298.31	299.44	299.44	3775600
2009-12-23	300.62	305.29	300.30	304.70	304.70	4160800
2009-12-24	305.32	308.60	304.99	308.09	308.09	1723800
2009-12-28	309.67	311.83	308.09	310.27	310.27	3408400
2009-12-29	311.20	311.25	307.99	308.54	308.54	2860200
2009-12-30	308.10	310.20	307.85	310.20	310.20	2942100

2. Discrete Models

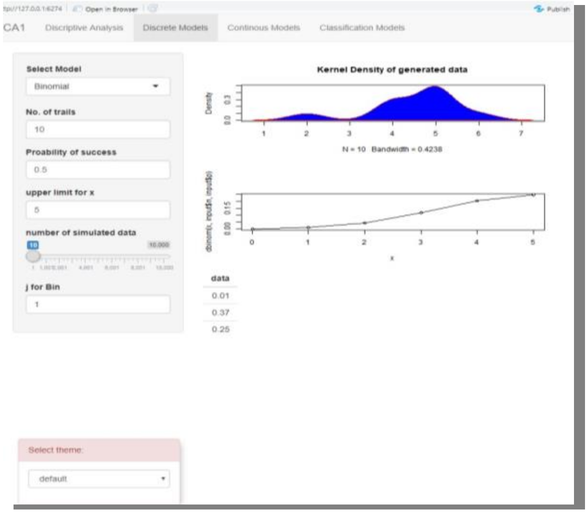
Discrete modelling is the discrete analogue of continuous modelling. In discrete modelling, formulae are fit to discrete data—data that could potentially take on only a countable set of values, such as the integers, and which are not infinitely divisible.

In the app, Discrete Models consists of three models, namely, Binomial, Geometric and Poisson.

In Binomial Model, the user can select the number of trails, the probability of success, upper limit for *x*, number of simulated data (default value - 10) and *j* value. After choosing all the data, the final graph is to be considered for final output. The Kernel Density will be updated as soon as any of the above specified data is altered.

In Geometric Model, the user can alter the value of *p* parameter, upper limit of *x*, number of simulated data (default value - 10) and *j* value. After choosing all the data, the final graph is to be considered for final output. The Kernel Density will be updated as soon as any of the above specified data is altered.

In Poisson Model, the user is given an option to alter the values of the parameter *lambda*, upper limit of *x*, number of simulated data (default value - 10) and *j* value. After choosing all the data, the final graph is to be considered for final output. The Kernel Density will be updated as soon as any of the above specified data is altered.



Select Model
Binomial

No. of trails
10

Proability of success
0.5

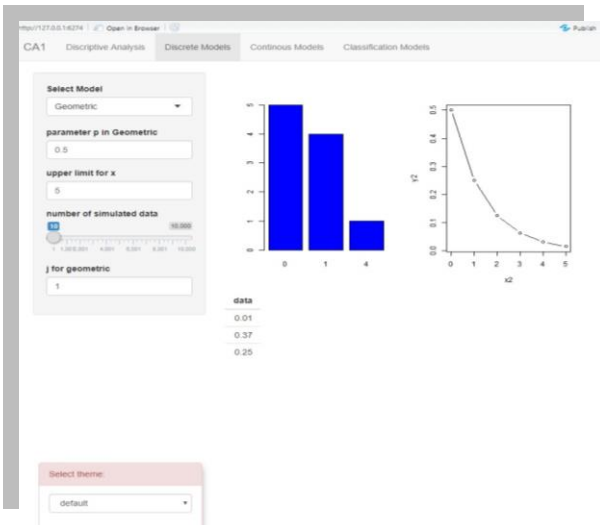
upper limit for x
5

number of simulated data
10

j for Bin
1

data
0.01
0.37
0.25

Select theme:
default



Select Model
Geometric

parameter p in Geometric
0.5

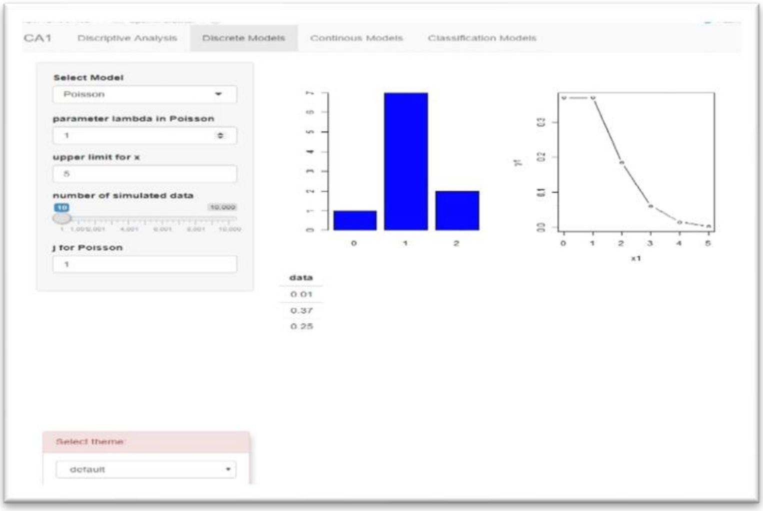
upper limit for x
5

number of simulated data
10

j for geometric
1

data
0.01
0.37
0.25

Select theme:
default



Select Model
Poisson

parameter lambda in Poisson
1

upper limit for x
5

number of simulated data
10

j for Poisson
1

data
0.01
0.37
0.25

Select theme:
default

3. Continuous Models

If a random variable is a continuous variable, its probability distribution is called a **continuous probability distribution.**

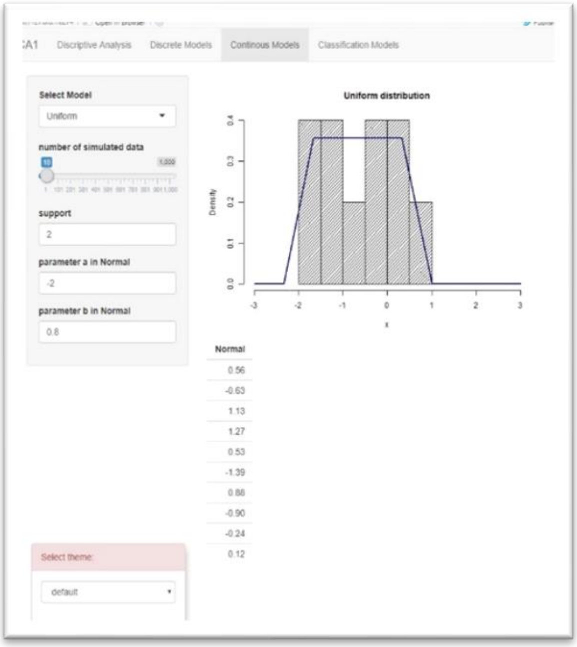
In the app, Continuous Models consists of three models, namely, Normal, Exponential and Uniform.

In Uniform Model, the user needs to feed the value of simulated data, support variable, parameter ‘*a*’ and parameter ‘*b*’.

In Exponential Model, the user needs to feed the value of simulated data, support variable, parameter lambda and *j* value.

In Normal Model, the user needs to feed the value of simulated data, support variable, parameter ‘*mu*’, parameter ‘*sigma’, and* j value.

After updating the values of the input, the graph is plotted on the right hand corner space of the screen.



Select Model: Uniform

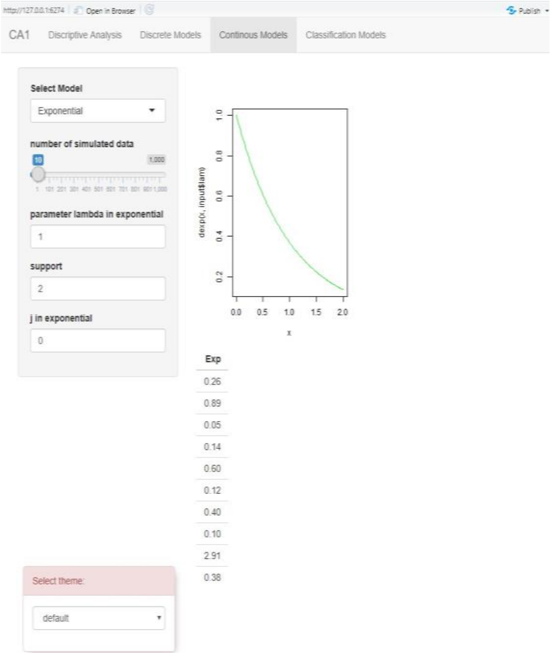
number of simulated data: 10

support: 2

parameter a in Normal: -2

parameter b in Normal: 0.8

Normal
0.56
-0.63
1.13
1.27
0.53
-1.39
0.88
-0.90
-0.24
0.12



Select Model: Exponential

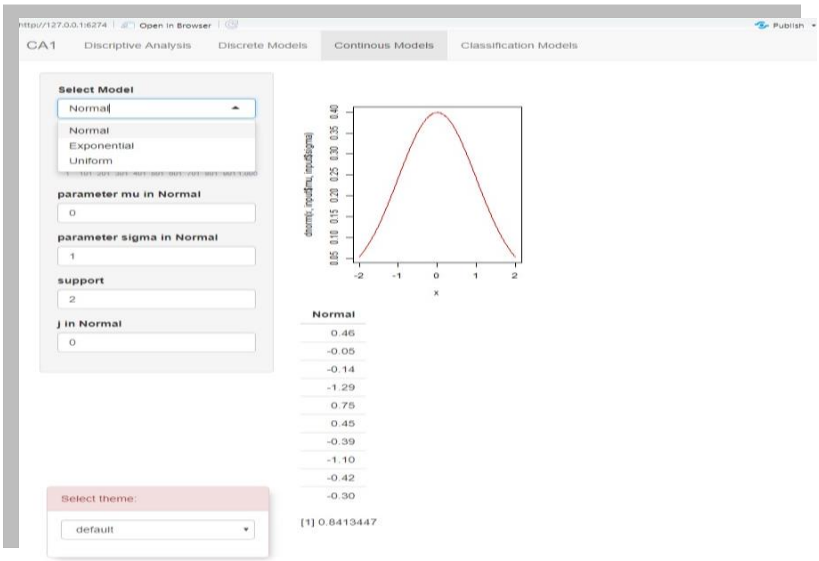
number of simulated data: 10

parameter lambda in exponential: 1

support: 2

j in exponential: 0

Exp
0.26
0.89
0.05
0.14
0.60
0.12
0.40
0.10
2.91
0.38



Select Model: Normal

parameter mu in Normal: 0

parameter sigma in Normal: 1

support: 2

j in Normal: 0

Normal
0.46
-0.05
-0.14
-1.29
0.75
0.45
-0.39
-1.10
-0.42
-0.30

```
[1] 0.8413447
```

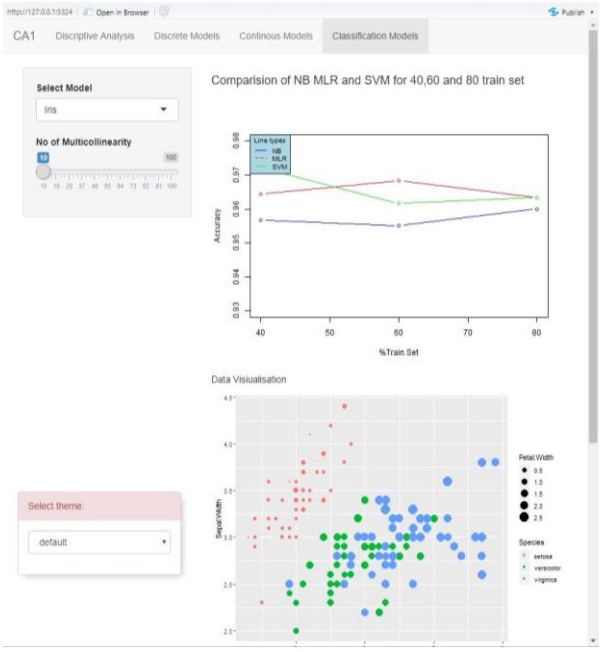
4. Classification Models

A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data. In short Classification either predicts categorical class labels or classifies data (construct a model) based on the training set and the values (class labels) in classifying attributes and uses it in classifying new data. There are a number of classification models. Here in our app we are using classification on the basis of Naïve Bayes, Logistic Regression and Support Vector Method.

In the app developed, we have used an existing in-built dataset – 'Iris' and an existing dataset (which is needed to be uploaded) – 'Supermarket'. In both the datasets, the user can change the number of multi-collinearities as per requirement and can see the respective output and see a comparison between all the three attributes; i.e.; Naïve Bayes, Logistic Regression and Support Vector Method.

*Description of **'Iris'** dataset:*

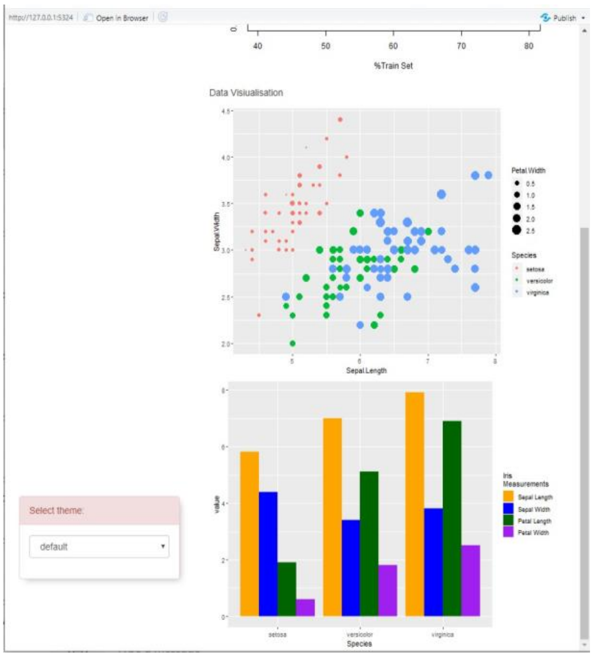
In Iris dataset, the comparison of NB, LR and SVM for 40, 60 and 80 train set is plotted in graph showing the relationship between percentage of train set and accuracy. Here, we can see as to which method have the maximum accuracy and at what percentage. Also, the Data Visualisation exemplifies the species, their size as per Petal Width and their sepal. It is distributed on the basis of Sepal length and Sepal Width, but also shows the species and the width of the petals of a particular species and hence covers four-dimensions in one graph. On the basis of this, a new graph is plotted which shows the value of all the four characteristics of the species.



Select Model: Iris

No of Multicollinearity: 10

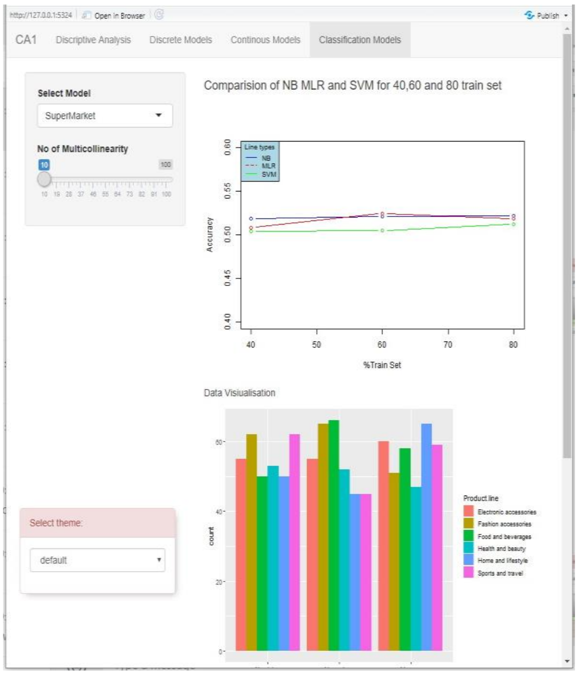
Select theme: default



Select theme: default

*Description of **'Supermarket'** dataset:*

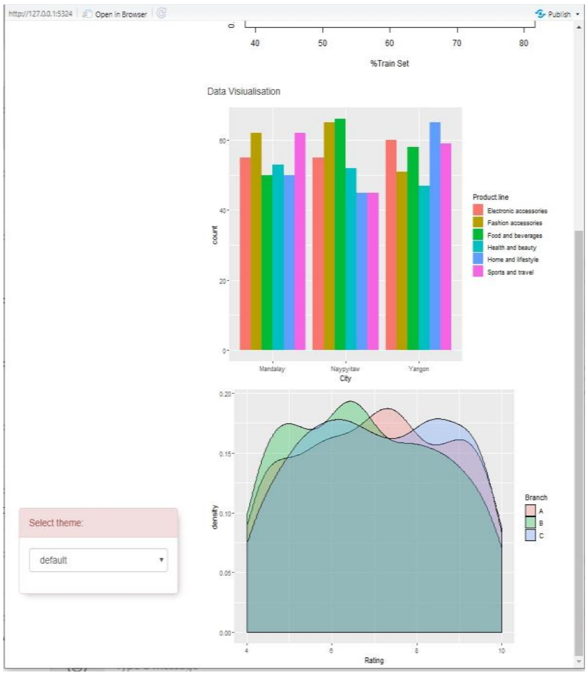
For Supermarket dataset, the user needs to download the dataset attached in his system and put its path in the program and he can then run the app. In this dataset the comparison of NB, LR and SVM for 40, 60 and 80 train set are plotted in graph showing the relationship between percentage of train set and accuracy (similarly, as it is done in 'Iris'). Here, we can see as to which method have the maximum accuracy and at what percentage. In this dataset, the Data Visualisation exemplifies the relation between city and count and represents six different product lines. It basically shows the product used by number of people in a particular city. The final graph visualises the density of rating verse each branch.



Select Model: SuperMarket

No of Multicollinearity: 10

Select theme: default



Select theme: default

Contribution towards Project:

Nidhi: Descriptive Analysis

Ankita: Discrete Models

Jyoti and Arvind: Continuous models

Classification Models and consolidating all tabs to shiny app : Varsha Ganapathy Rao