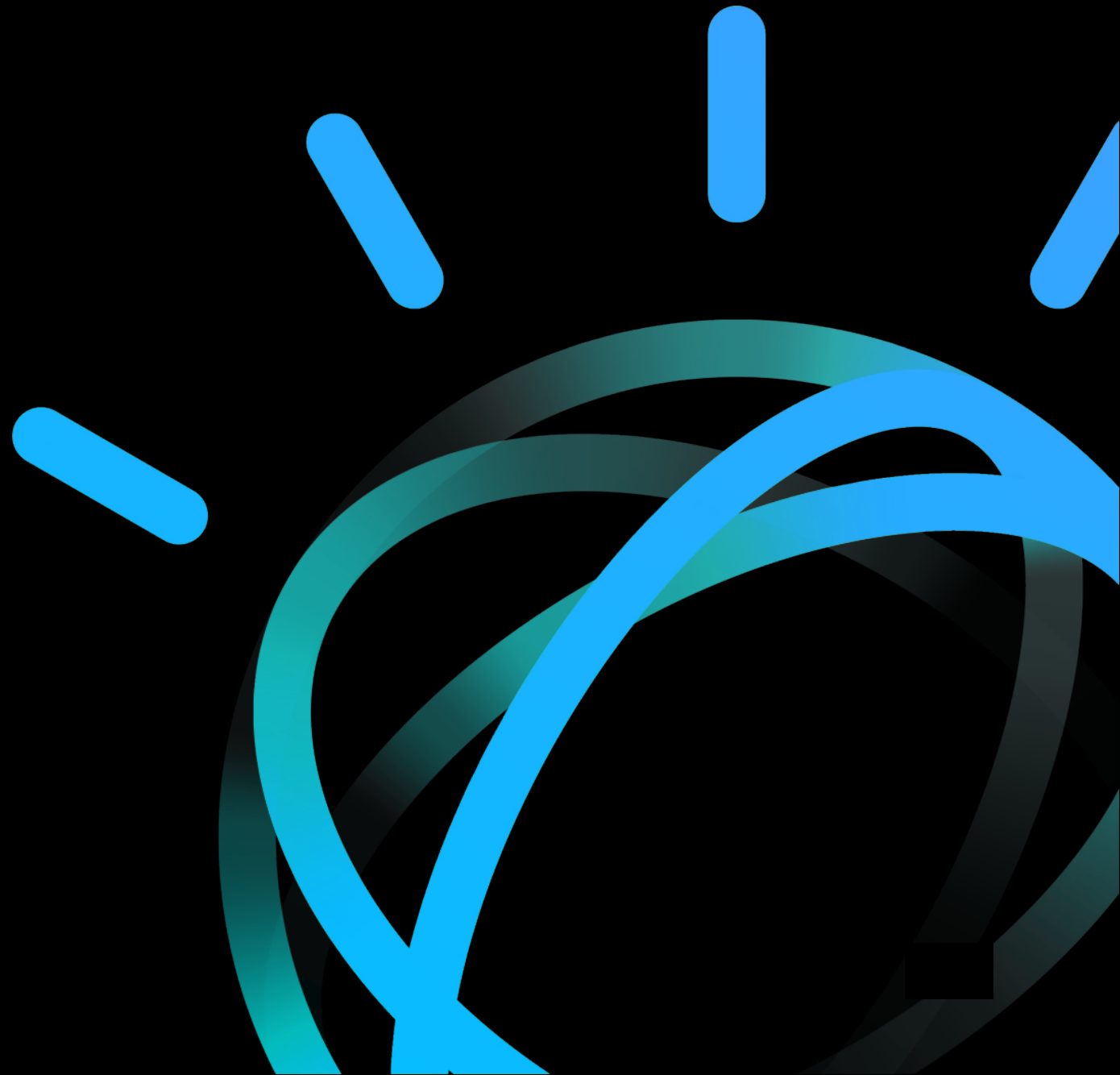
Capstone Project



Murat KONCUK

Agenda

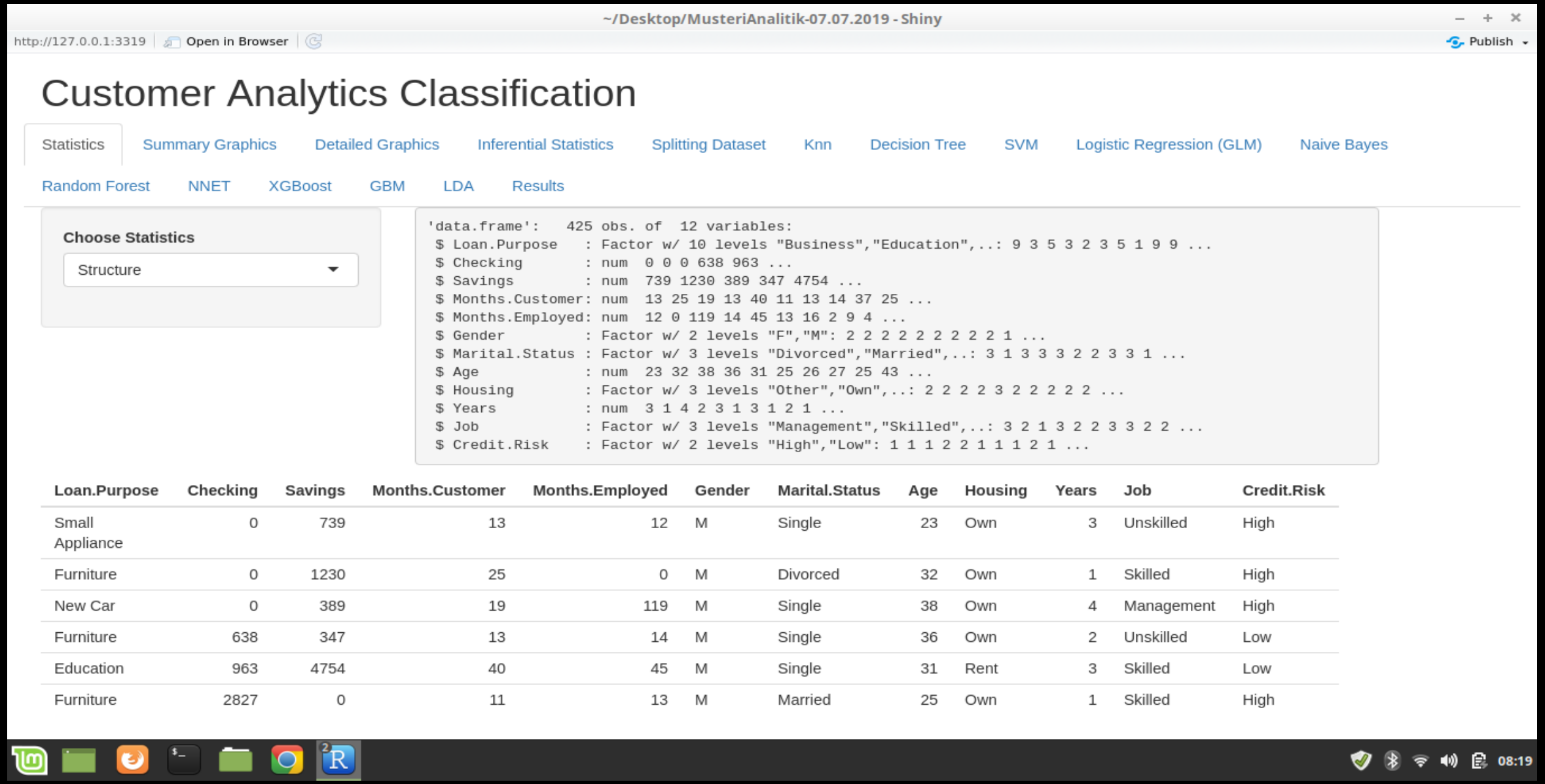
- Business Understanding
- Data Understanding (Exploration & Cleaning)
- Modelling
- Model Evaluation
- Result

Data Science Lifecycle

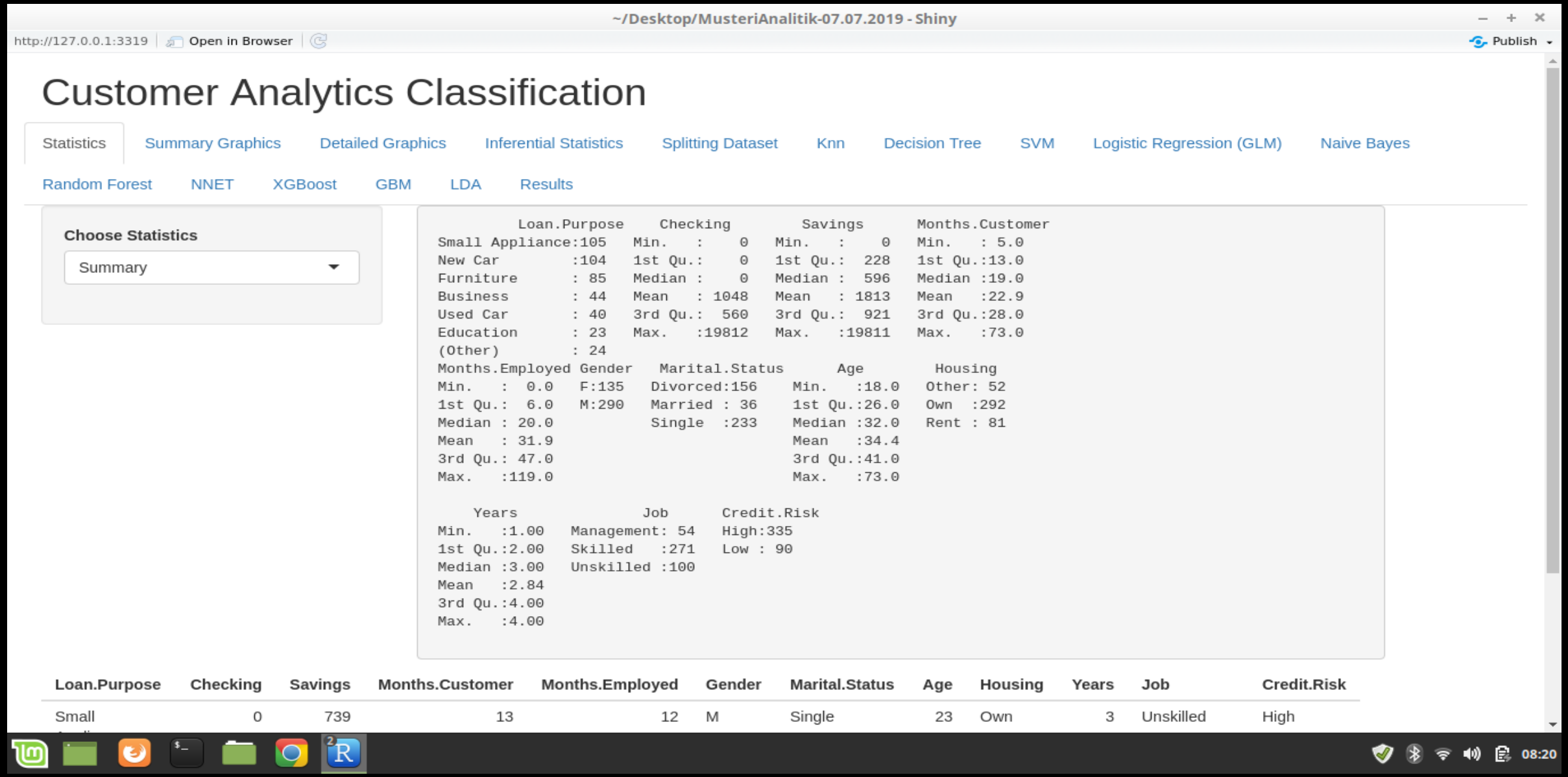
Business Understanding

	A	B	C	D	E	F	G	H	I	J	K	L
1	Loan Purpose	Checking	Savings	Months Customer	Months Employed	Gender	Marital Status	Age	Housing	Years	Job	Credit Risk
2	9	$0	$739	13	12	1	1	23	1	3	1	1
3	3	$0	$1.230	25	0	1	2	32	1	1	3	1
4	5	$0	$389	19	119	1	1	38	1	4	2	1
5	3	$638	$347	13	14	1	1	36	1	2	1	0
6	2	$963	$4.754	40	45	1	1	31	2	3	3	0
7	3	$2.827	$0	11	13	1	3	25	1	1	3	1
8	5	$0	$229	13	16	1	3	26	1	3	1	1
9	1	$0	$533	14	2	1	1	27	1	1	1	1
10	9	$6.509	$493	37	9	1	1	25	1	2	3	0
11	9	$966	$0	25	4	0	2	43	1	1	3	1
12	1	$0	$989	49	0	1	1	32	2	2	2	1
13	5	$0	$3.305	11	15	1	1	34	2	2	1	1
14	1	$322	$578	10	14	1	3	26	1	1	3	1
15	5	$0	$821	25	63	1	1	44	1	1	3	1
16	5	$396	$228	13	26	1	1	46	1	3	1	0
17	10	$0	$129	31	8	1	2	39	1	4	2	1
18	3	$652	$732	49	4	0	2	25	1	2	3	0
19	5	$708	$683	13	33	1	1	31	1	2	3	1
20	7	$207	$0	28	116	1	1	47	1	4	3	1
21	2	$287	$12.348	7	2	0	2	23	2	2	3	0
22	3	$0	$17.545	34	16	0	2	22	1	4	3	1
23	3	$101	$3.871	13	5	0	2	26	2	4	3	0
24	3	$0	$0	25	23	1	3	19	1	4	3	1
25	3	$0	$485	37	23	0	2	27	1	2	2	1
26	5	$0	$10.723	11	15	1	1	39	2	2	1	1
27	1	$141	$245	22	33	1	1	26	1	3	3	1
28	10	$0	$0	19	58	1	1	50	3	4	3	1
29	10	$2.484	$0	49	46	1	1	34	3	1	3	1
30	9	$237	$236	37	24	1	1	23	2	4	3	1

Data Understanding (Statistic)



Data Understanding (Statistic)



Data Understanding (Visualize)

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish

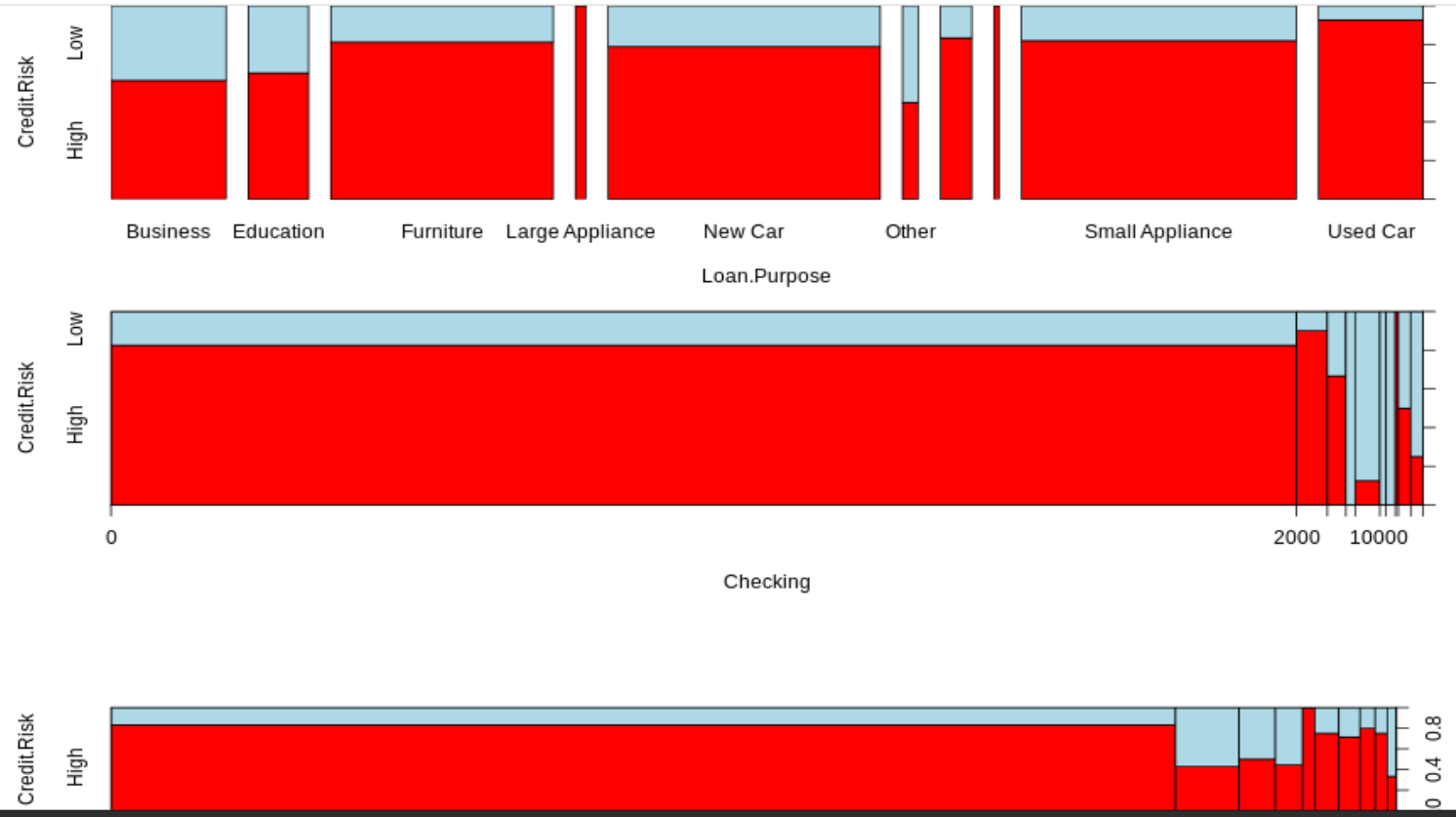
Customer Analytics Classification

Statistics | Summary Graphics | Detailed Graphics | Inferential Statistics | Splitting Dataset | Knn | Decision Tree | SVM | Logistic Regression (GLM) | Naive Bayes

Random Forest | NNET | XGBoost | GBM | LDA | Results

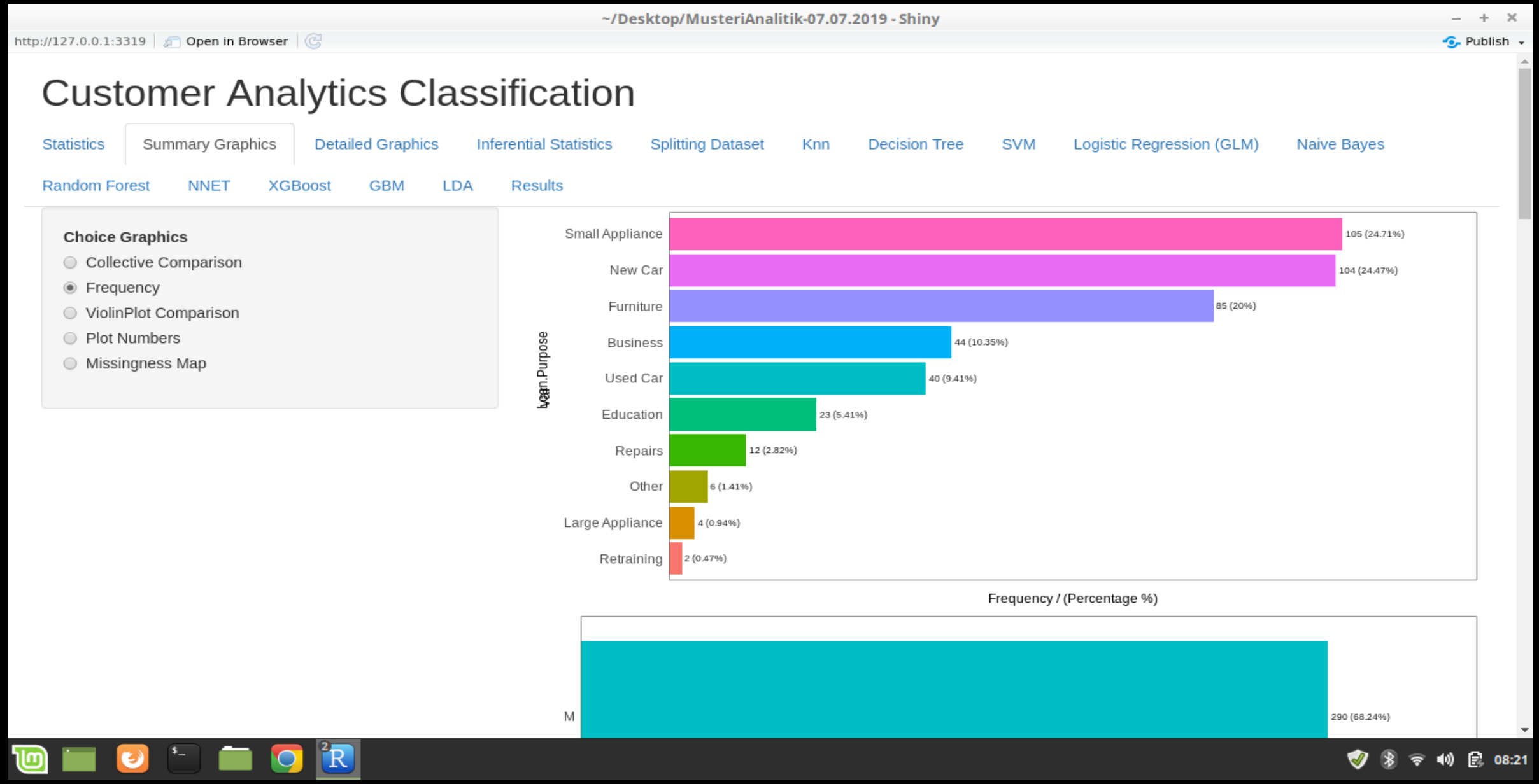
Choice Graphics

- (•) Collective Comparison
- ( ) Frequency
- ( ) ViolinPlot Comparison
- ( ) Plot Numbers
- ( ) Missingness Map

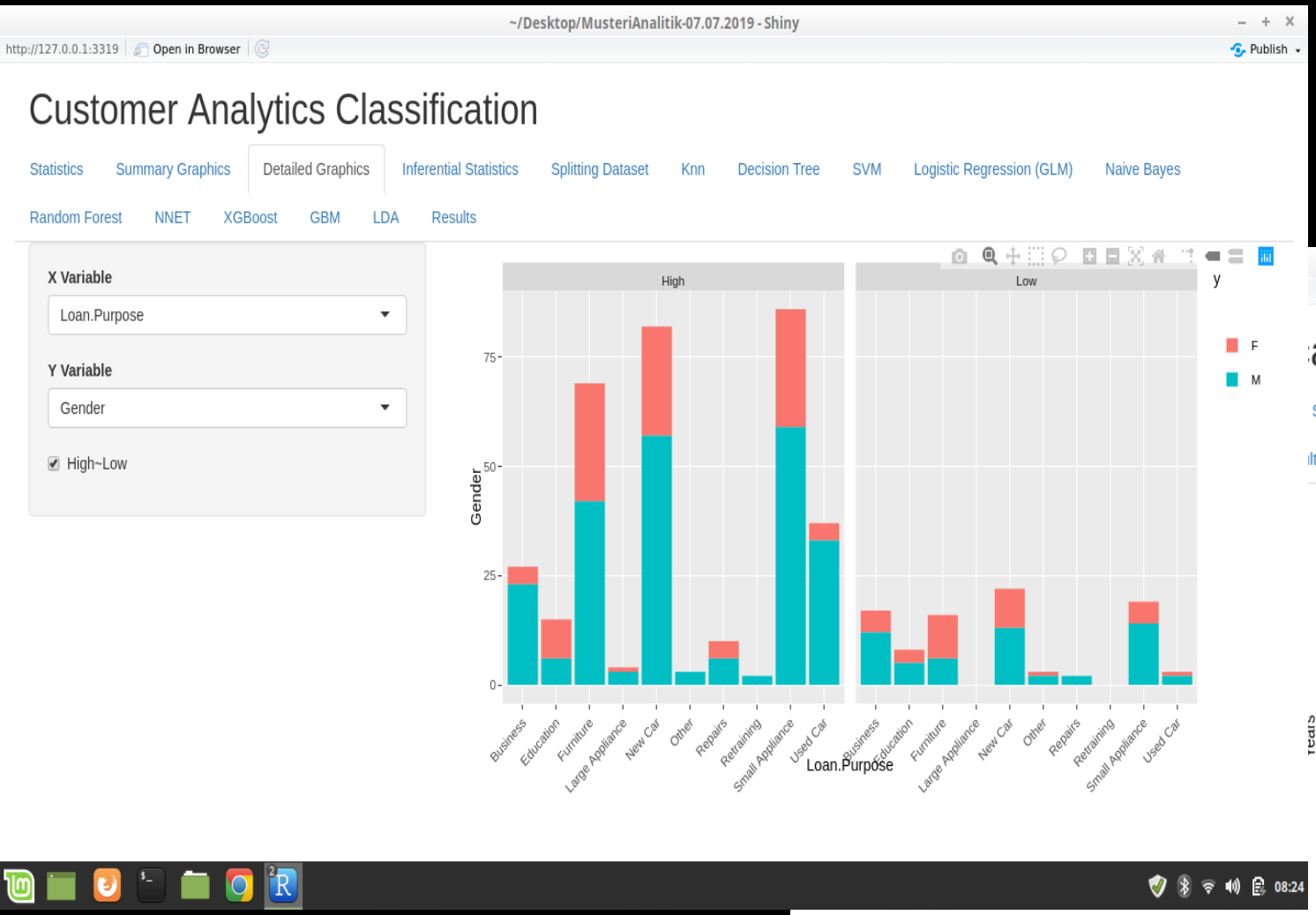


08:21

Data Understanding (Visualize)



Data Understanding (Exploration-Cleaning)



Data Understanding (Exploration-Cleaning)

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish

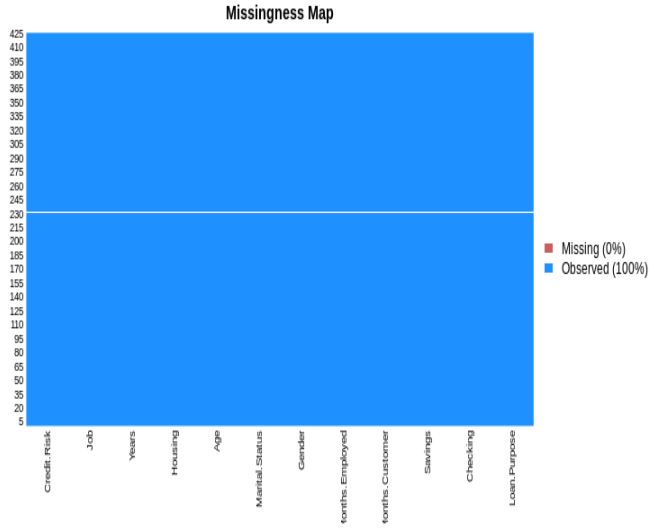
Customer Analytics Classification

Statistics | Summary Graphics | Detailed Graphics | Inferential Statistics | Splitting Dataset | Knn | Decision Tree | SVM | Logistic Regression (GLM) | Naive Bayes

Random Forest | NNET | XGBoost | GBM | LDA | Results

Choice Graphics

- Collective Comparison
- Frequency
- ViolinPlot Comparison
- Plot Numbers
- Missingness Map



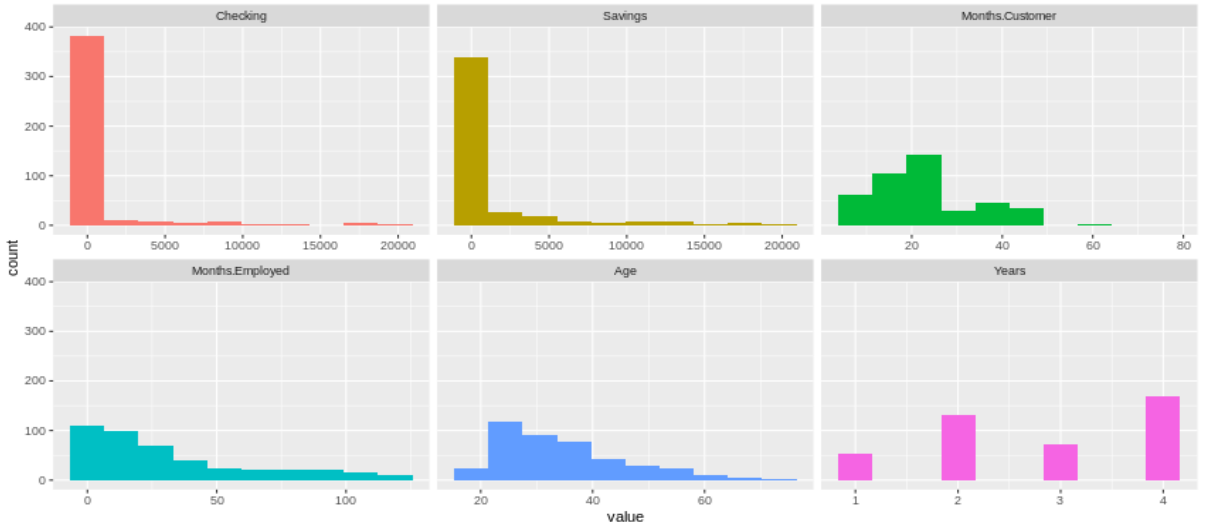
~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

Publish

ssification

Inferential Statistics | Splitting Dataset | Knn | Decision Tree | SVM | Logistic Regression (GLM) | Naive Bayes

Results



Data Understanding (Exploration-Cleaning-Python)

```
In [10]: cmap = 'coolwarm'
         corr=df.corr()
         corr.style.background_gradient(cmap, axis=1)
```

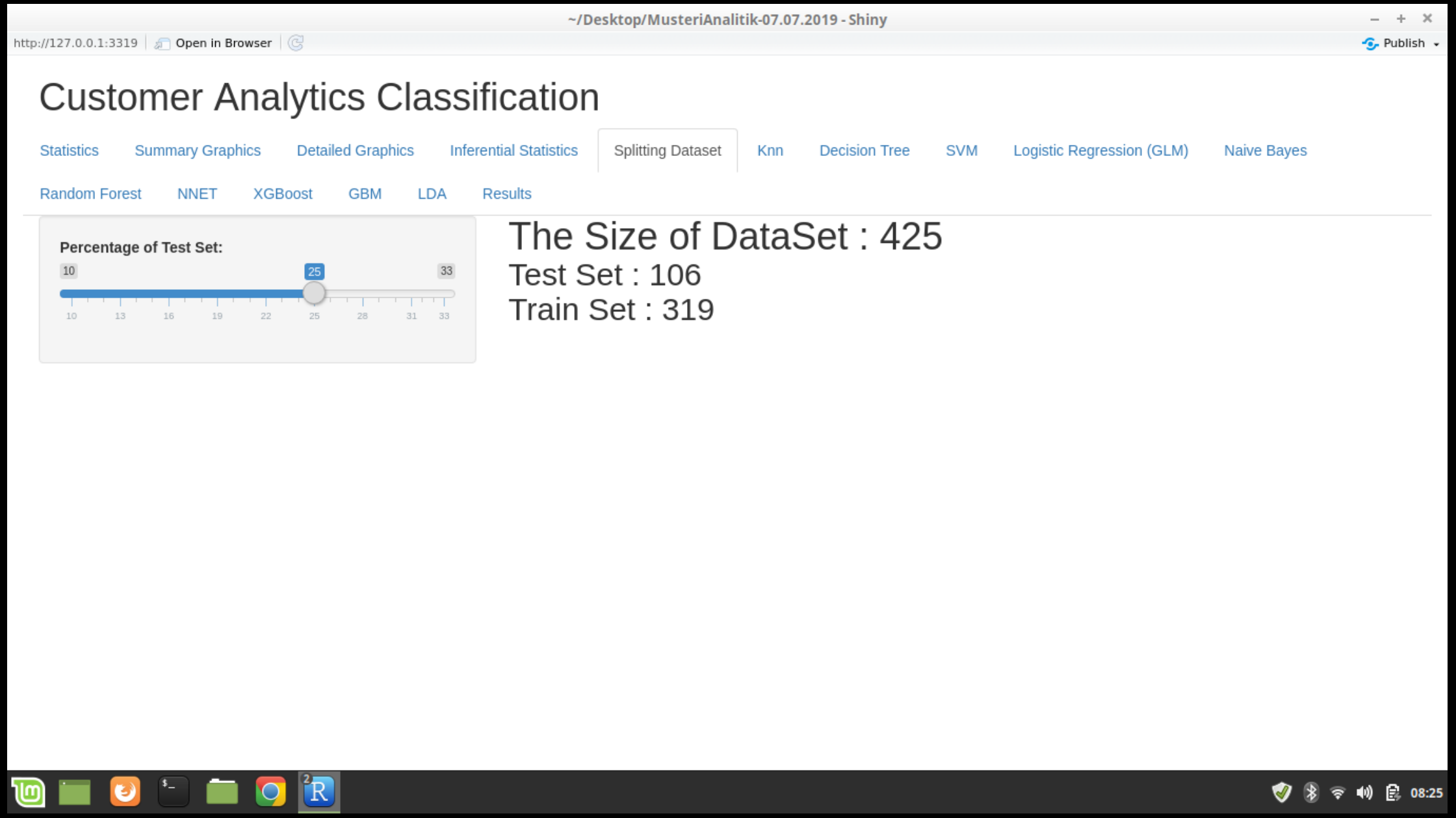
Out[10]:

	Checking	Savings	Months Customer	Months Employed	Gender	Marital Status	Age	Housing	Years	Job	Credit Risk
Checking	1	0.0200792	-0.0362263	-0.00919098	0.0207947	-0.0175604	-0.0023693	0.0396903	-0.0355347	0.0067802	-0.351463
Savings	0.0200792	1	-0.0504407	0.048978	-0.04985	0.0251077	-0.0282859	-0.0422778	-0.0193616	0.0186301	-0.141489
Months Customer	-0.0362263	-0.0504407	1	0.0544558	0.125811	-0.133166	-0.00378591	0.220757	0.0831089	0.16697	-0.0838113
Months Employed	-0.00919098	0.048978	0.0544558	1	0.246575	-0.259574	0.306799	0.0486035	0.28823	0.0113563	0.081447
Gender	0.0207947	-0.04985	0.125811	0.246575	1	-0.489119	0.156988	-0.0449949	-0.0214099	-0.0414339	0.066945
Marital Status	-0.0175604	0.0251077	-0.133166	-0.259574	-0.489119	1	-0.221221	-0.0948081	-0.149197	0.0243386	-0.131088
Age	-0.0023693	-0.0282859	-0.00378591	0.306799	0.156988	-0.221221	1	0.094231	0.240027	-0.134193	0.0479164
Housing	0.0396903	-0.0422778	0.220757	0.0486035	-0.0449949	-0.0948081	0.094231	1	0.335998	0.049325	-0.0314407
Years	-0.0355347	-0.0193616	0.0831089	0.28823	-0.0214099	-0.149197	0.240027	0.335998	1	-0.0348175	0.0933453
Job	0.0067802	0.0186301	0.16697	0.0113563	-0.0414339	0.0243386	-0.134193	0.049325	-0.0348175	1	0.0625659

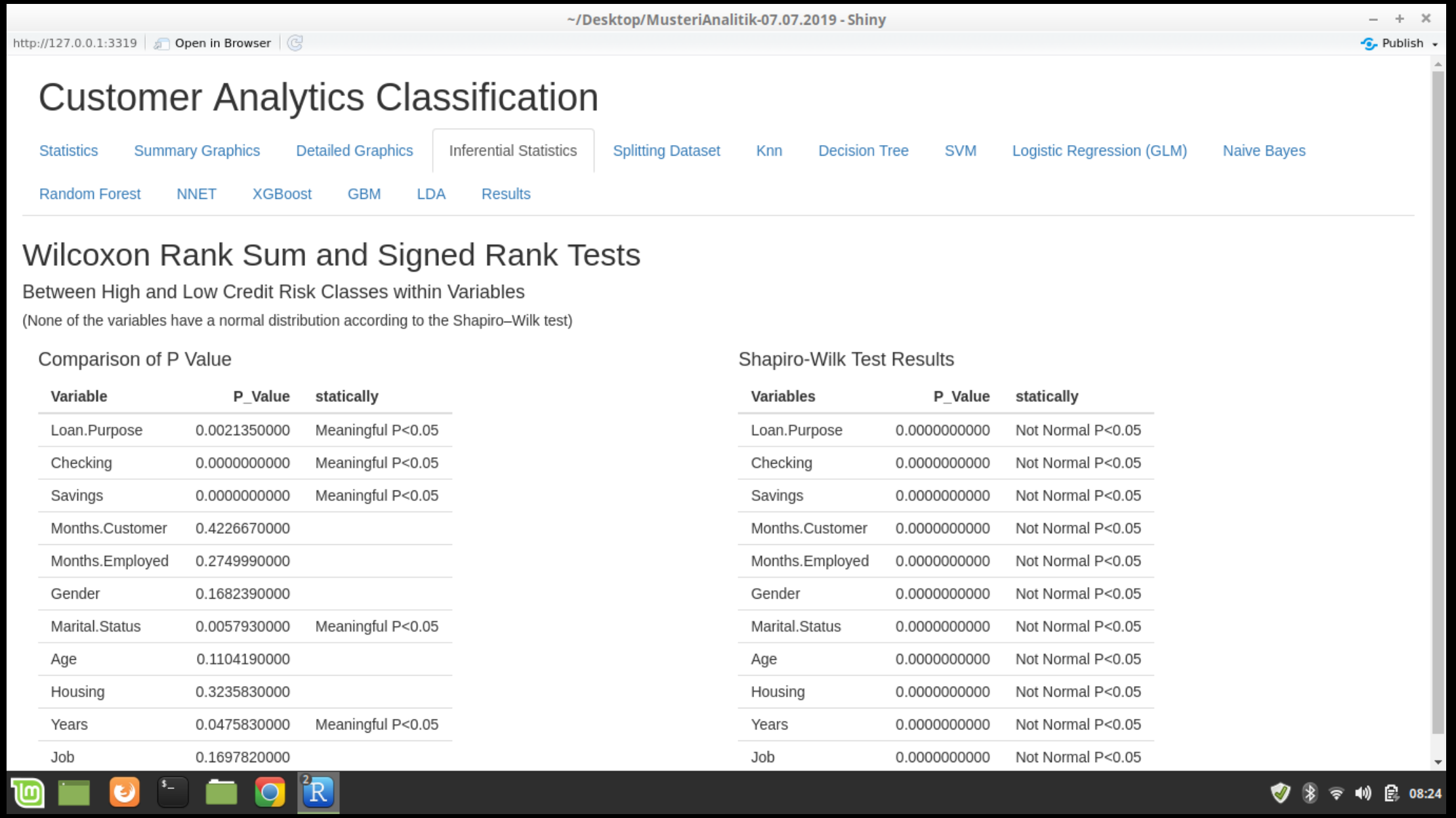
```
aler().fit(A).transform(A)

ite-packages/sklearn/preprocessing/data.py:645: DataConv
 converted to float64 by StandardScaler.
X, y)
ite-packages/ipykernel_launcher.py:1: DataConversionWarn
to float64 by StandardScaler.
hing an IPython kernel.
```

```
Out[9]: array([[ 1.14004792, -0.3333931 , -0.29878856, ..., -0.62151966,
                  0.14734777, -1.78085412],
                [-0.84704859, -0.3333931 , -0.16213587, ..., -0.62151966,
                 -1.6944994 ,  0.6685407 ],
                [-0.18468309, -0.3333931 , -0.39619882, ..., -0.62151966,
                  1.06827136, -0.55615671],
                ...,
                [-0.18468309, -0.3333931 , -0.50446338, ..., -0.62151966,
                 -0.77357581,  0.6685407 ],
                [-0.18468309, -0.3333931 , -0.30630306, ..., -0.62151966,
                 -0.77357581,  0.6685407 ],
                [-0.18468309, -0.3333931 , -0.25064005, ..., -0.62151966,
                  0.14734777, -0.55615671]])
```

Modelling (Data Spilit)



Modelling (Inferential)



Modelling (Decision Tree)

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish

Customer Analytics Classification

Statistics | Summary Graphics | Detailed Graphics | Inferential Statistics | Splitting Dataset | Knn | Decision Tree | SVM | Logistic Regression (GLM) | Naive Bayes

Random Forest | NNET | XGBoost | GBM | LDA | Results

Conditional Inference Trees Algorithm

Accuracy Rate :

0.92

Prediction	Reference	Freq
0	0	19
1	0	3
0	1	5
1	1	79

Predict Status

Gender (0-1)

Loan Purpose (1-

Checking

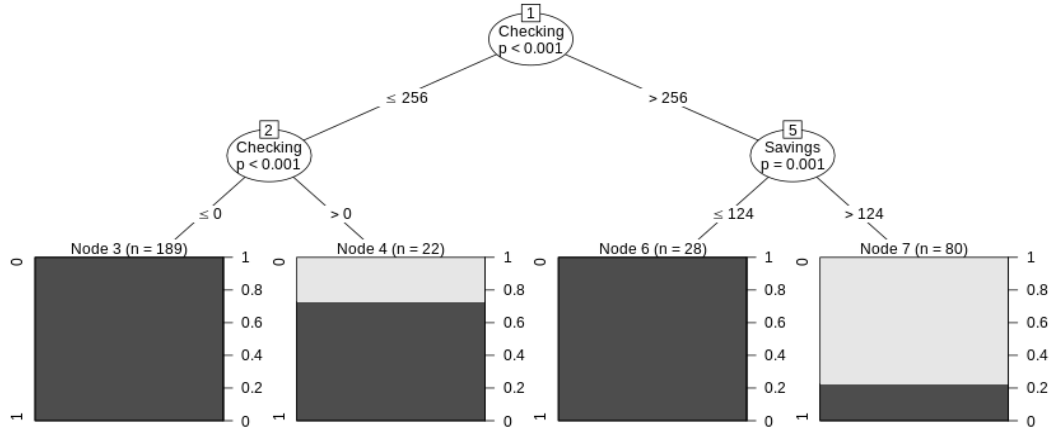
Savings

Months Customer

Months Employed

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish



Confusion Matrix and Statistics

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 19  5
         1  3 79

               Accuracy : 0.925
                 95% CI : (0.857, 0.967)
```

08:25

Modelling (Decision Tree-Pyhton)

In [55]:
```
credittree.fit(X_train,y_train)
```

Out[55]:
```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=4,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
```

In [56]:
```
predTree = credittree.predict(X_test)
```

In [57]:
```
print (predTree [0:12])
print (y_test [0:12])
```

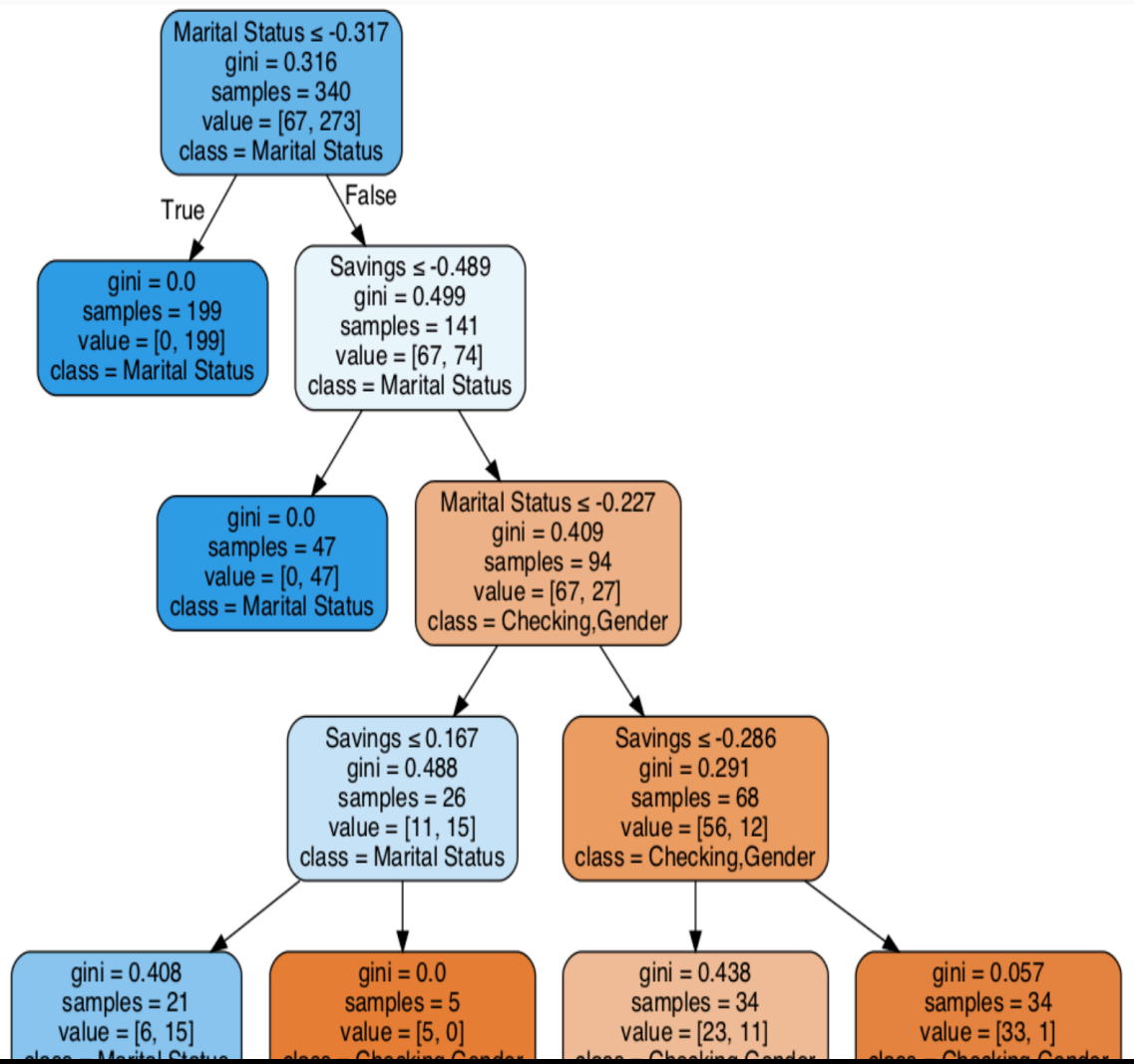
```
[0. 0. 0. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
[1. 1. 0. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
```

In [58]:
```
from sklearn import metrics
import matplotlib.pyplot as plt
print("DecisionTrees's Accuracy: ", metrics.accuracy_score(y_test, predTree))
```

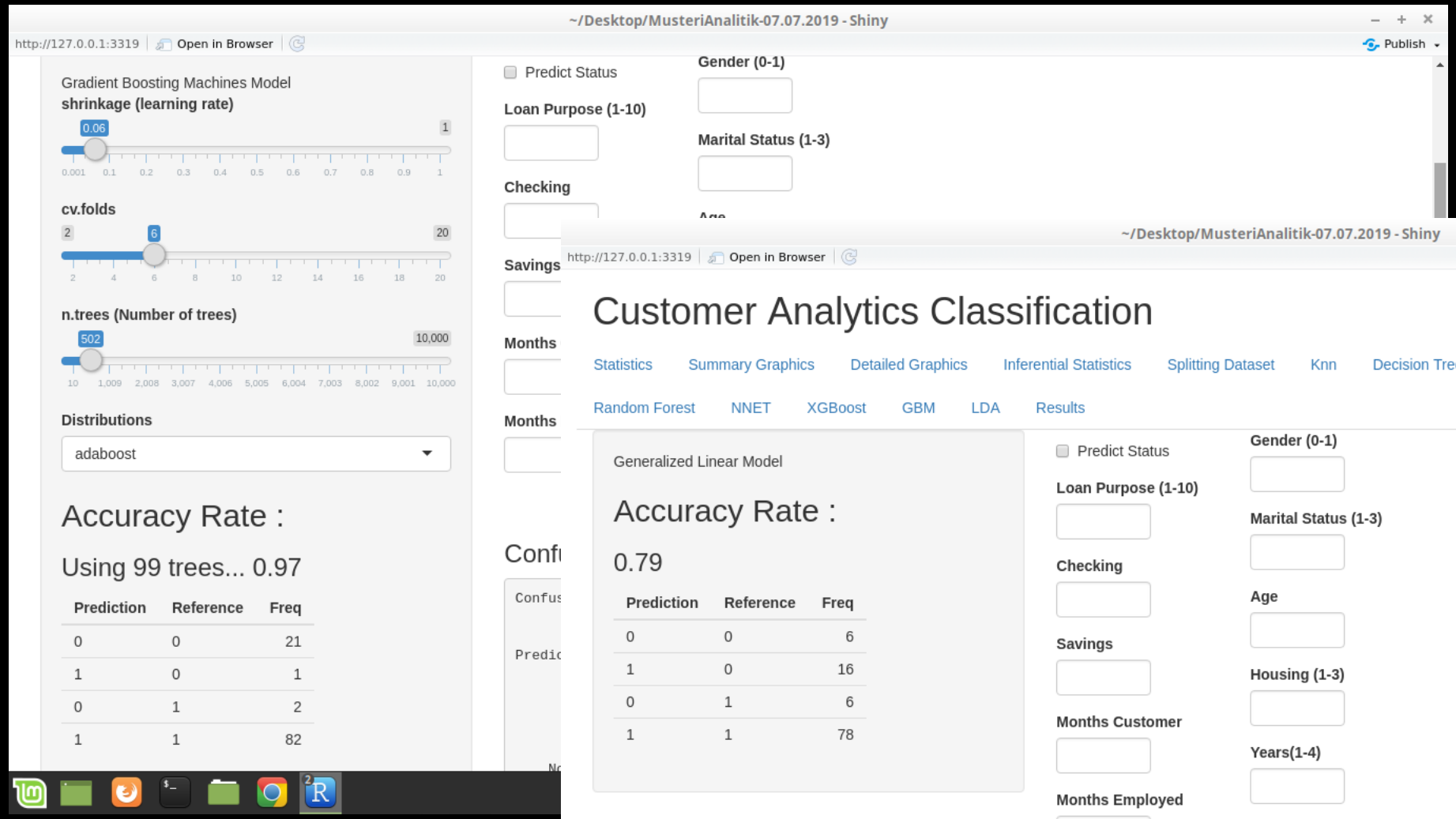
```
DecisionTrees's Accuracy:  0.9058823529411765
```

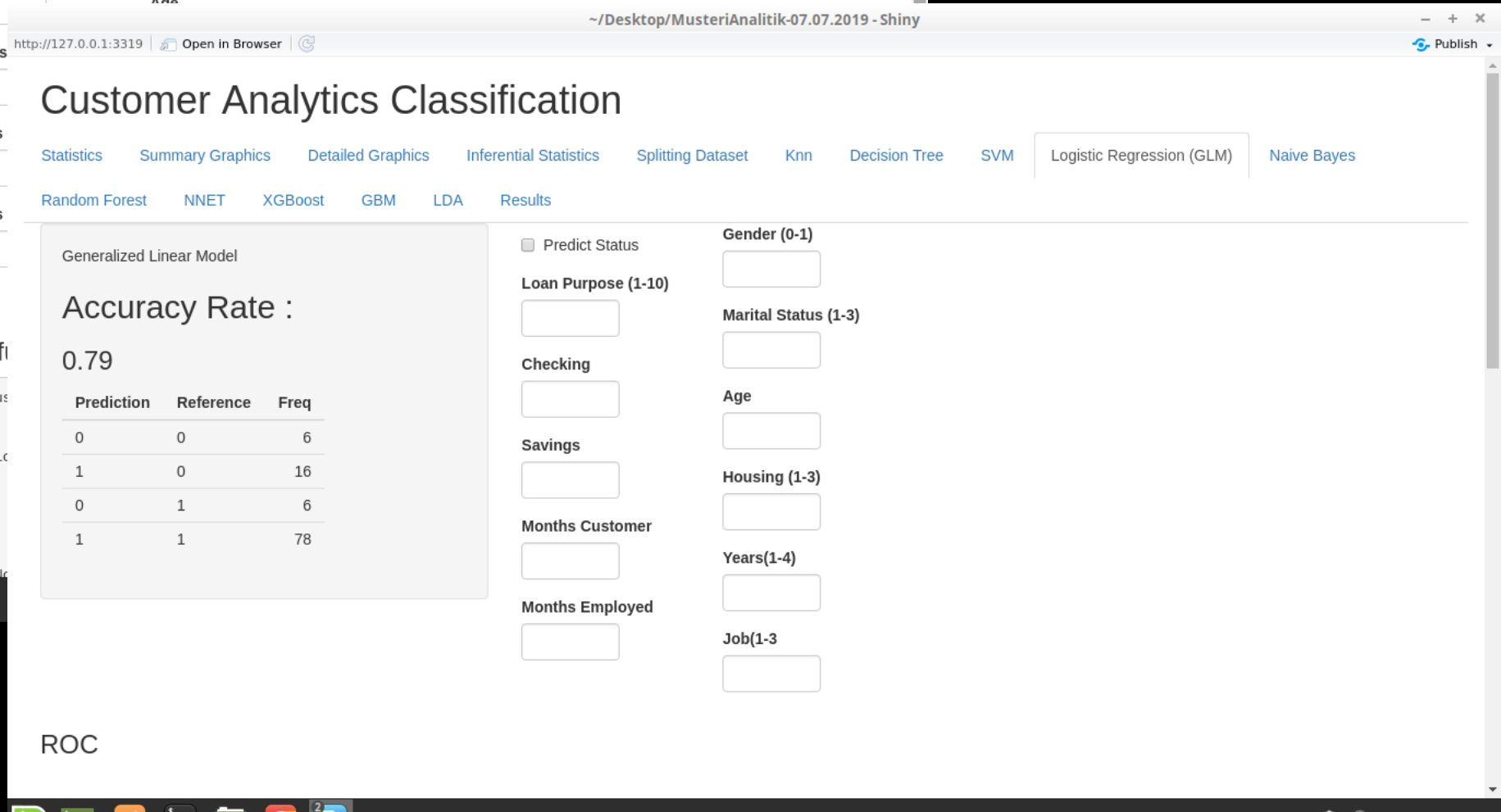
In [59]:
```
from sklearn.tree import export_graphviz
from sklearn.externals.six import StringIO
from IPython.display import Image
import pydotplus

col = ["Checking,""Gender","Marital Status","Savings","Months Customer","Months Employed","Marital Status","Age","Housi

dot_data = StringIO()
export_graphviz(credittree, out_file=dot_data,
                filled=True, rounded=True,
                special_characters=True,feature_names = col,class_names = col)
graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
graph.write_png('decison.png')
Image(graph.create_png())
```

Out[59]:

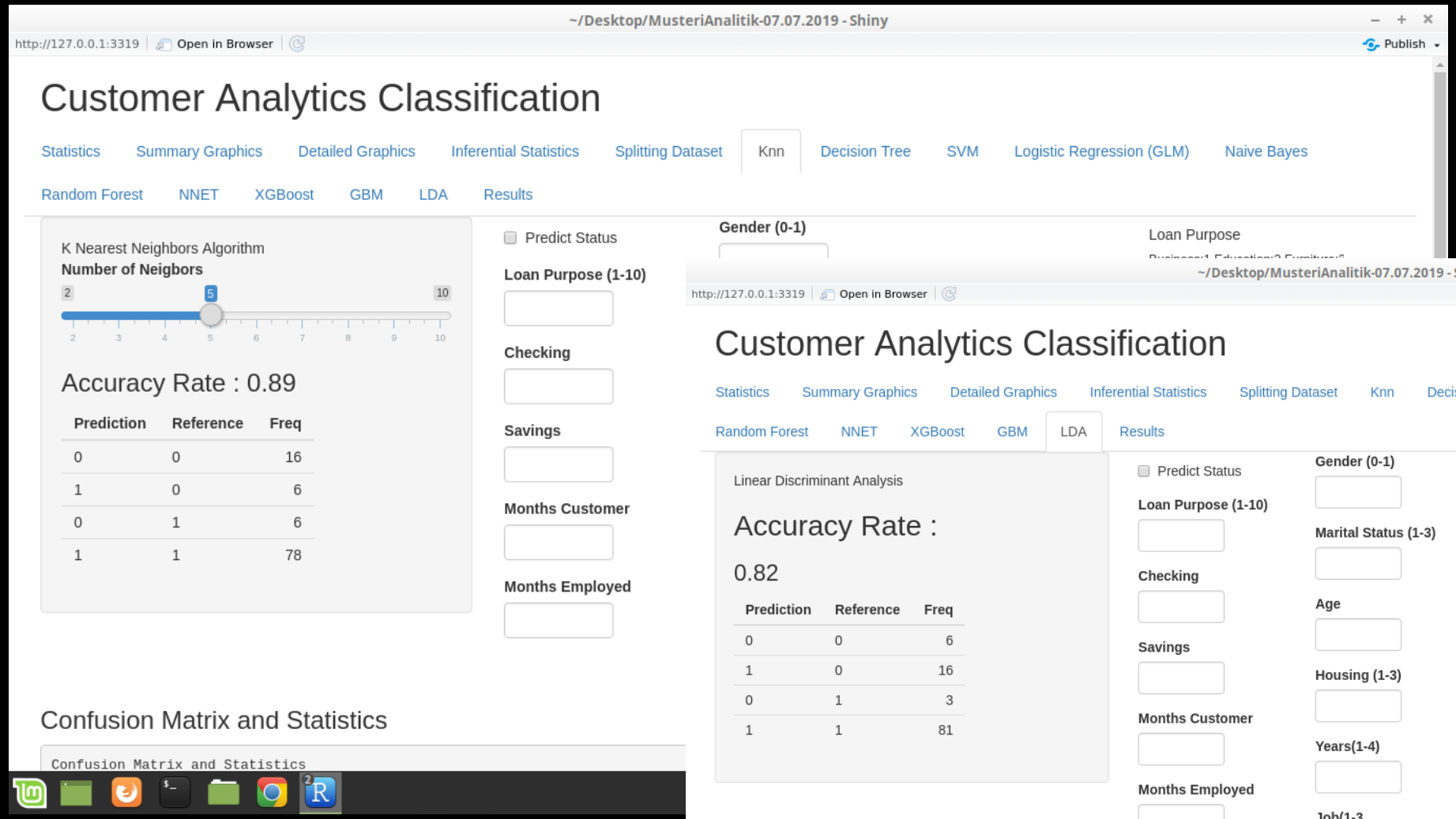


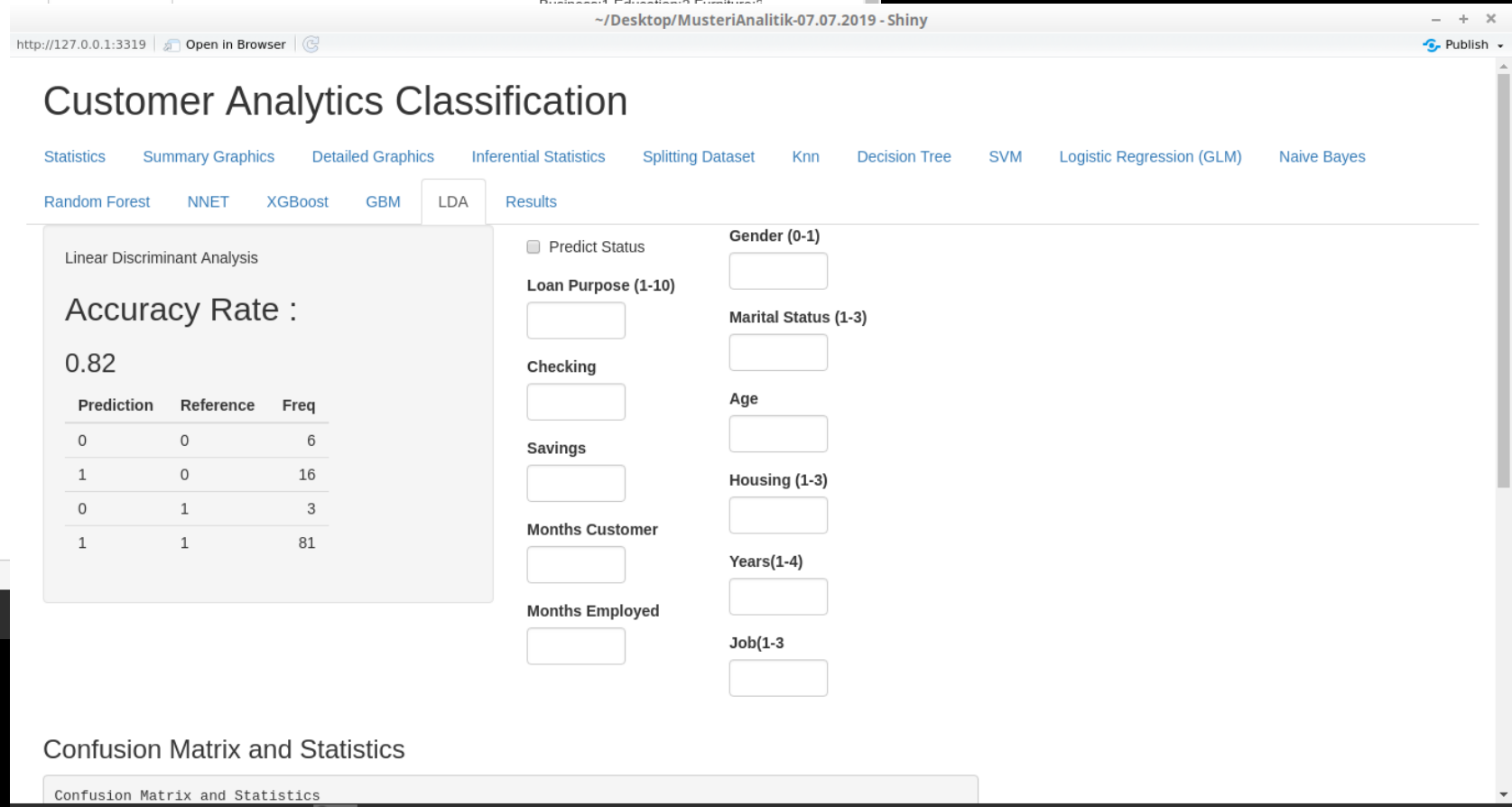
Modelling (GBM-GLM)





Modelling (KNN-LDA)





Modelling (NaiveBayes - NNET)

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish

Customer Analytics Classification

Statistics | Summary Graphics | Detailed Graphics | Inferential Statistics | Splitting Dataset | Knn | Decision Tree | SVM | Logistic Regression (GLM) | Naive Bayes | Random Forest | NNET | XGBoost | GBM | LDA | Results

Naive Bayes

Accuracy Rate :

0.81

Prediction	Reference	Freq
0	0	7
1	0	15
0	1	5
1	1	79

Predict Status

Gender (0-1)

Loan Purpose (1-10

Checking

Savings

Months Customer

Months Employed

ROC

~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish

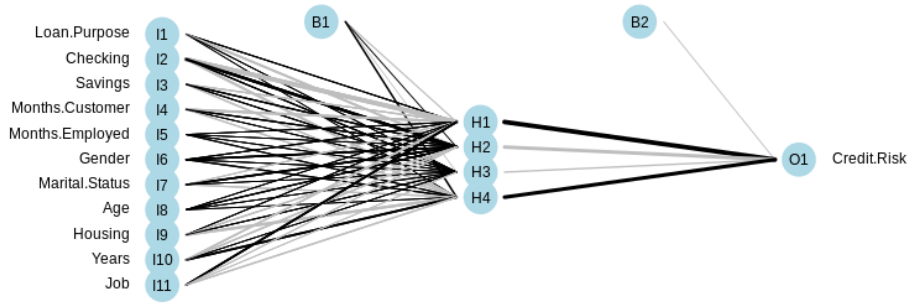
Accuracy Rate :

0.83

Prediction	Reference	Freq
0	0	8
1	0	14
0	1	4
1	1	80

Months Employed

Job(1-3

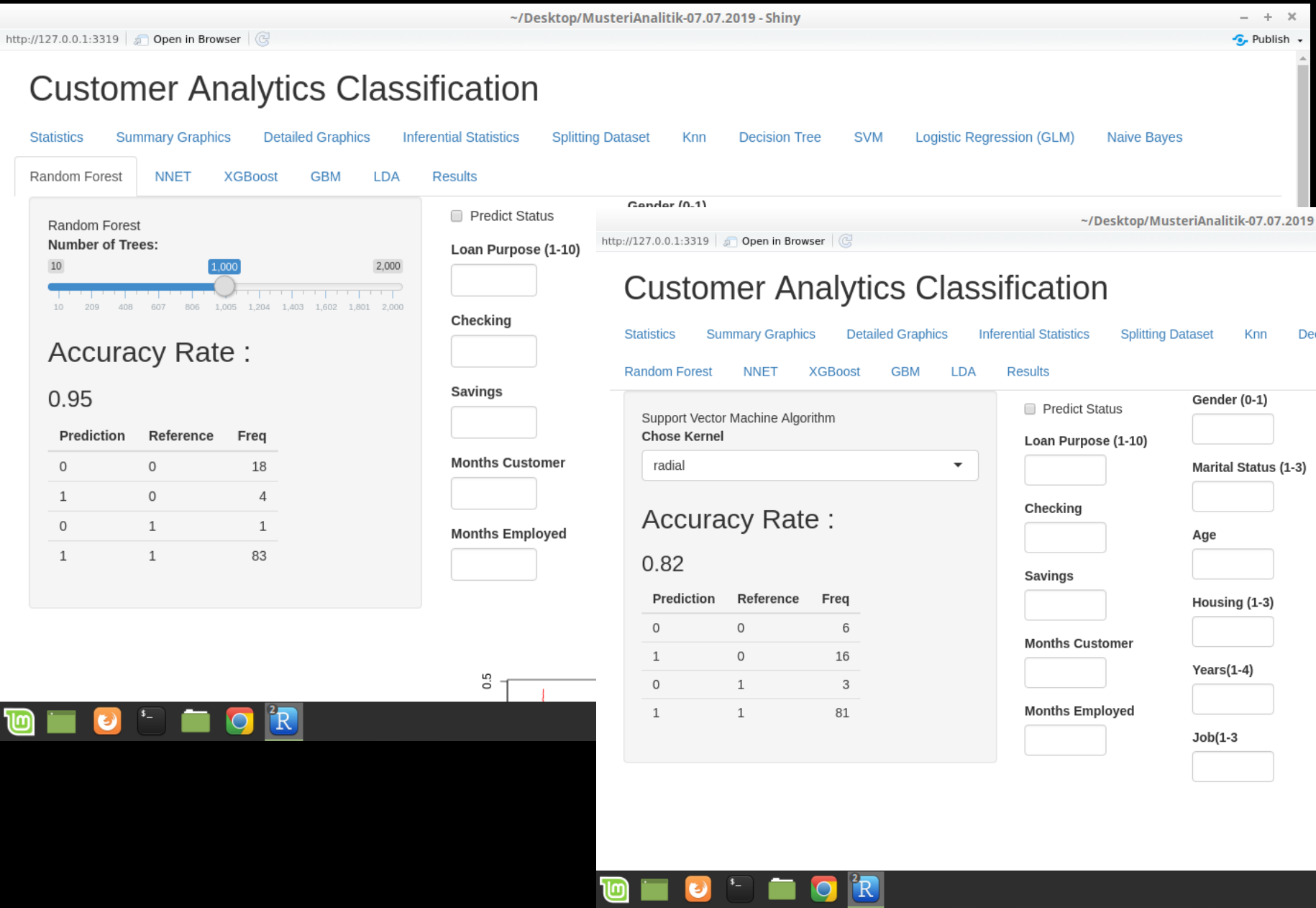


Confusion Matrix and Statistics

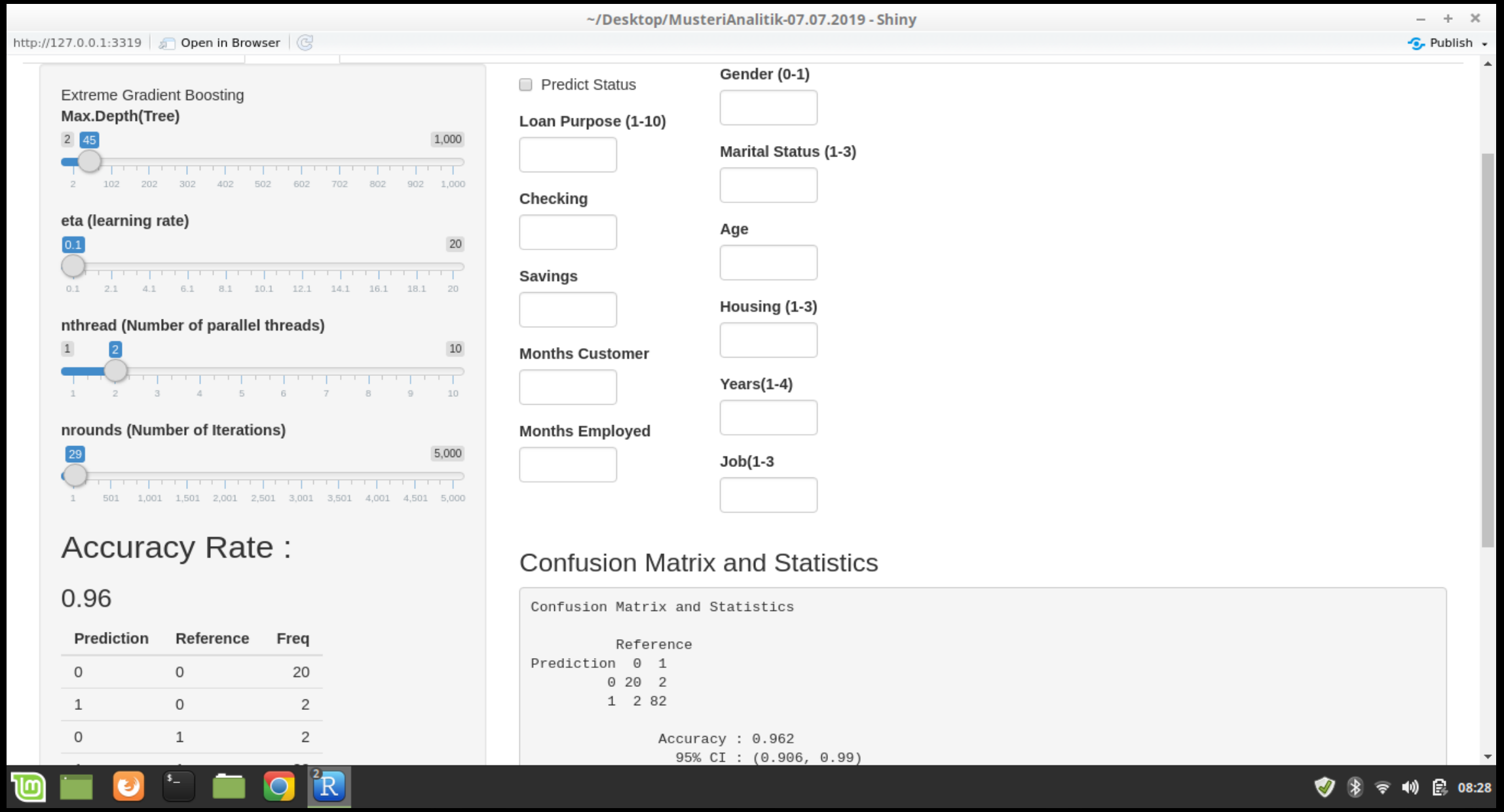
```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0  8  4
```

08:27

Modelling (Random Forest-SVM)



Modelling (XGBOOST)



Result

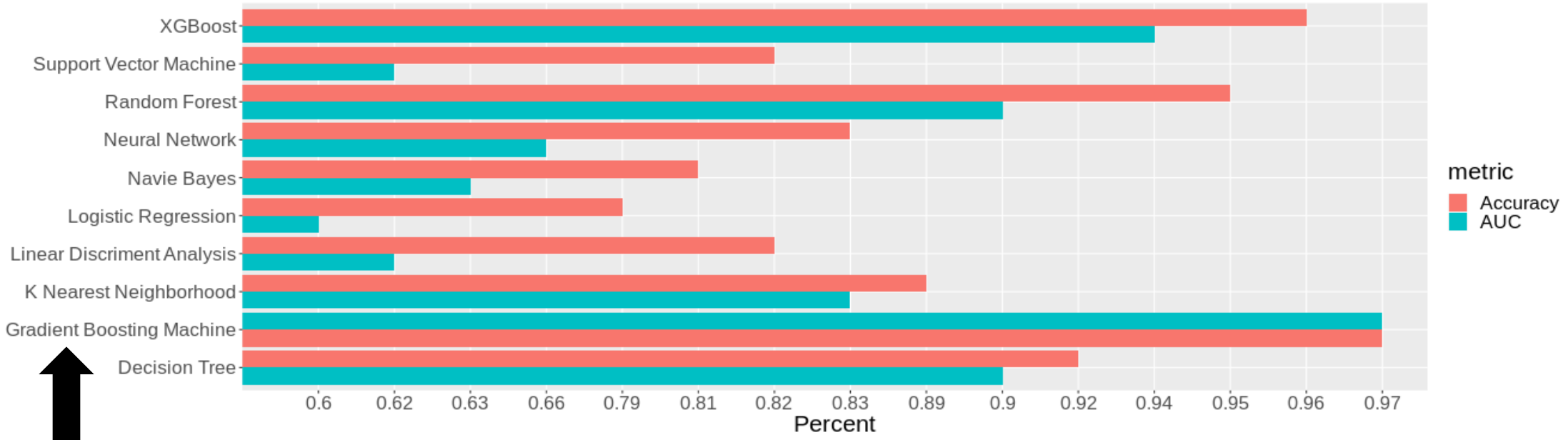
~/Desktop/MusteriAnalitik-07.07.2019 - Shiny

http://127.0.0.1:3319 | Open in Browser | Publish

Customer Analytics Classification

Statistics | Summary Graphics | Detailed Graphics | Inferential Statistics | Splitting Dataset | Knn | Decision Tree | SVM | Logistic Regression (GLM) | Naive Bayes | Random Forest | NNET | XGBoost | GBM | LDA | Results

Performance Metrics



Most Accuracy Model GBM

08:29

Thank You