Performance Analysis of Classification Algorithms

Presented by

Sai Hemanth Thota

Veera Reddy Vangala

Kapil Dharao

Rahul

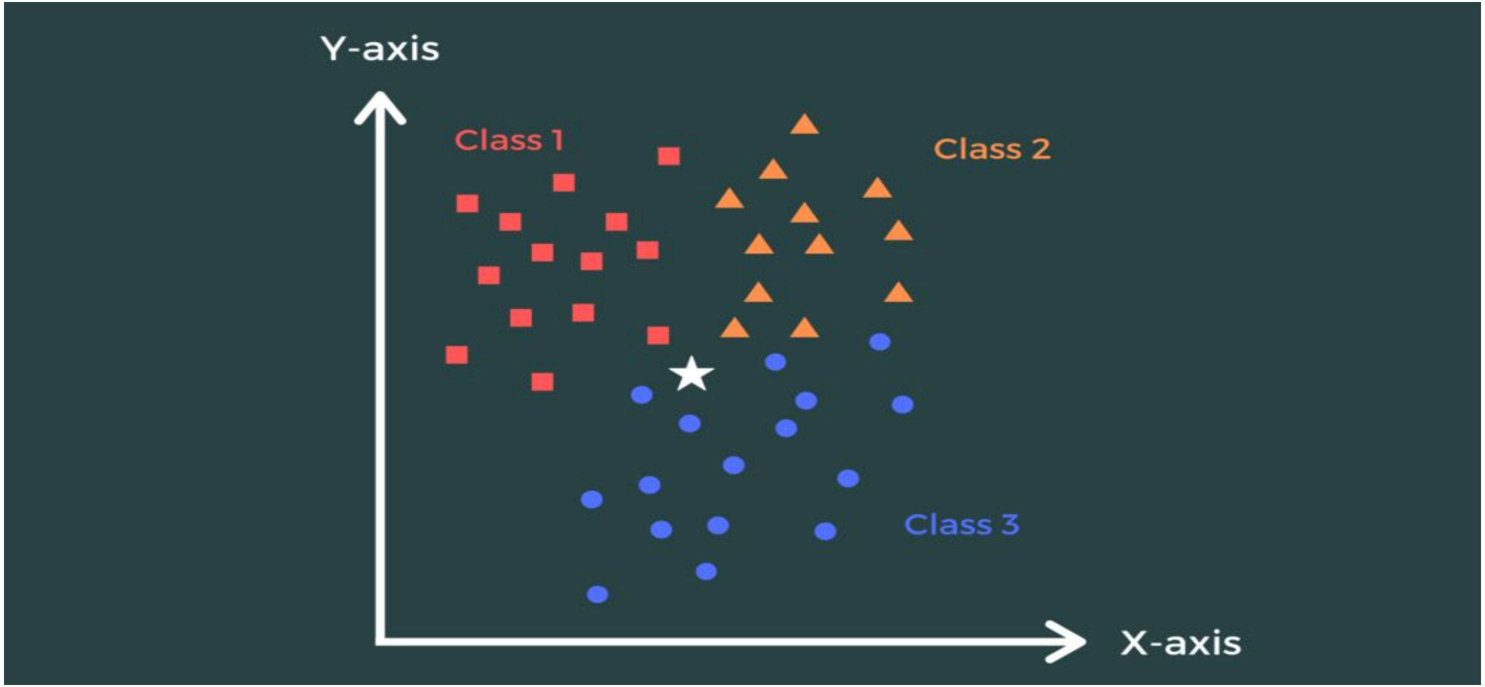
Problem Statement

- To identify for which type of dataset which classification technique is more suitable.
- Wrong selection of classification algorithm will certainly lead to bad classification model and bad results.
- It is also very hard to write each and every time code to analyze the dataset and create prediction model for it.
- So we designed a GUI which handles all these issues.

Classification Algorithms Used

- KNN
- Naive Bayes
- Decision Tree
- Random Forest

KNN Algorithm

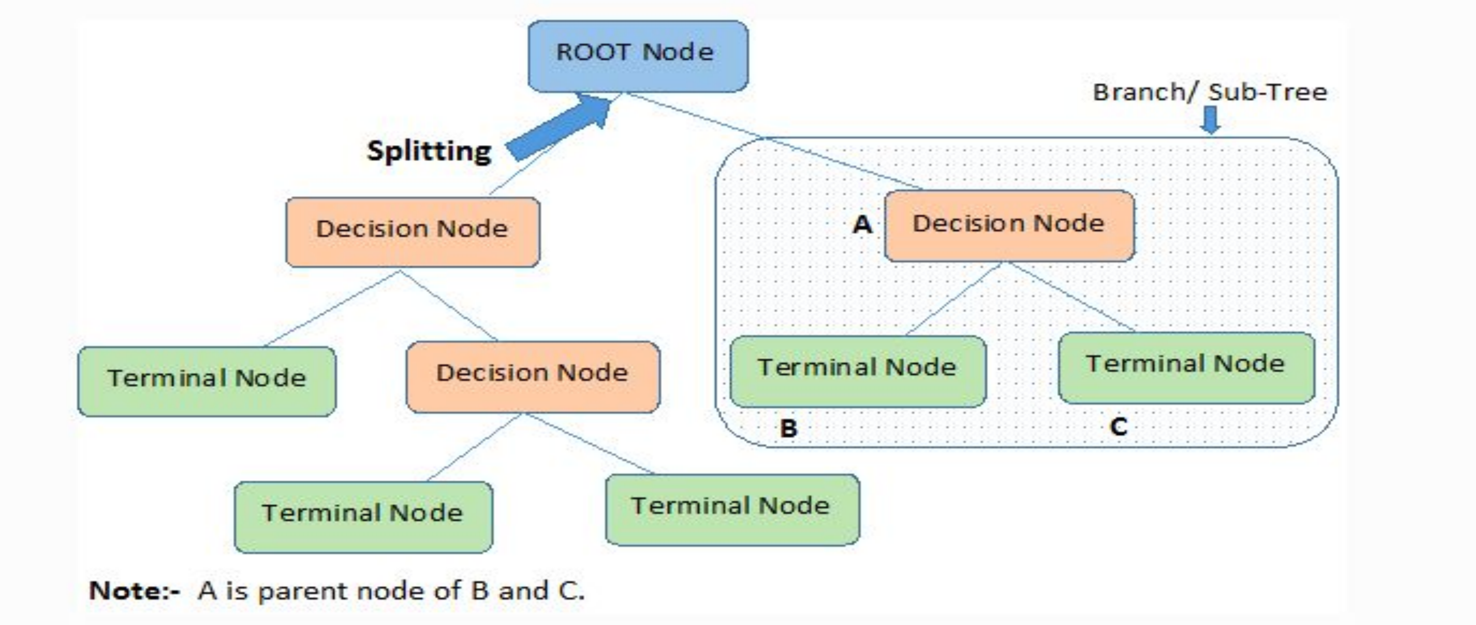


Naive Bayes

$$\overset{\text{Posterior}}{P(A|B)} = \frac{\overset{\text{Likelihood}}{P(B|A)}\,\overset{\text{Prior}}{P(A)}}{\underset{\text{Normalizing constant}}{P(B)}}$$

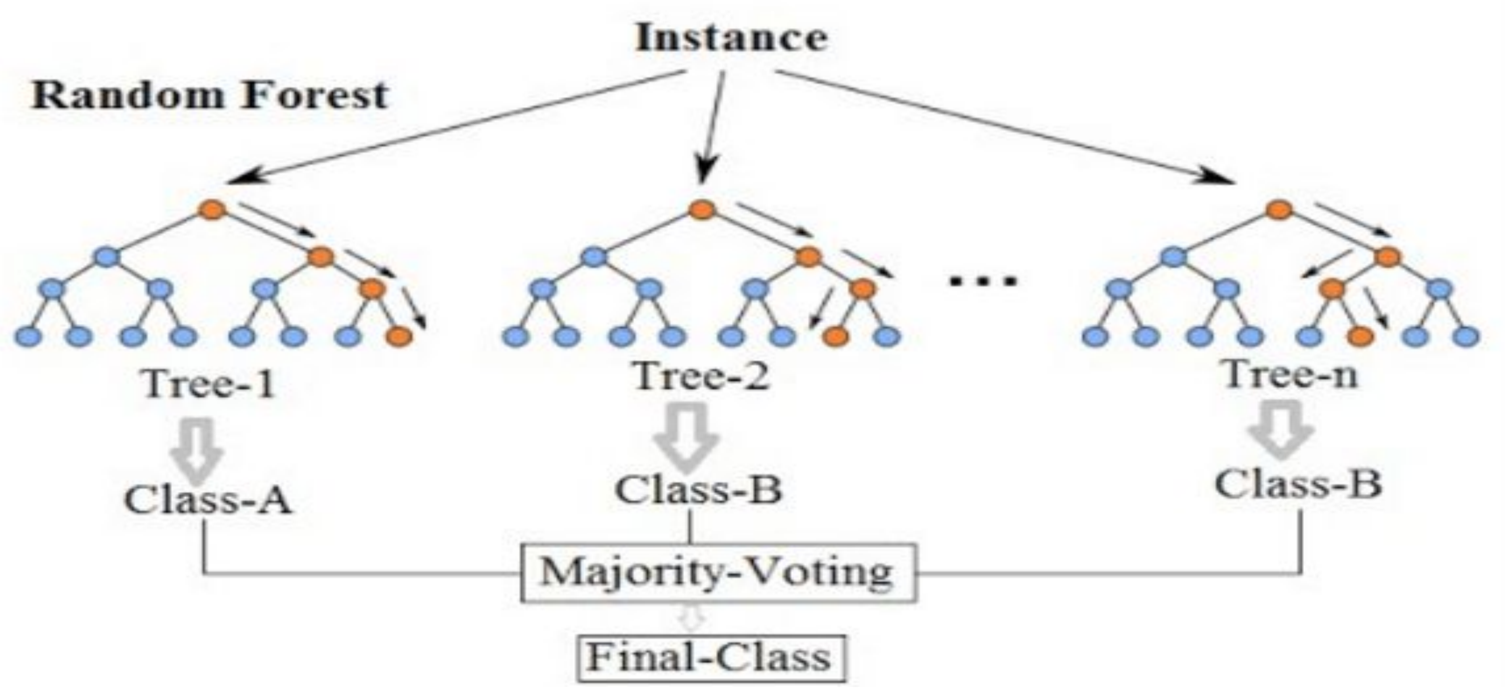
$$P(B) = \sum_{Y} P(B|A)P(A)$$

Decision Tree



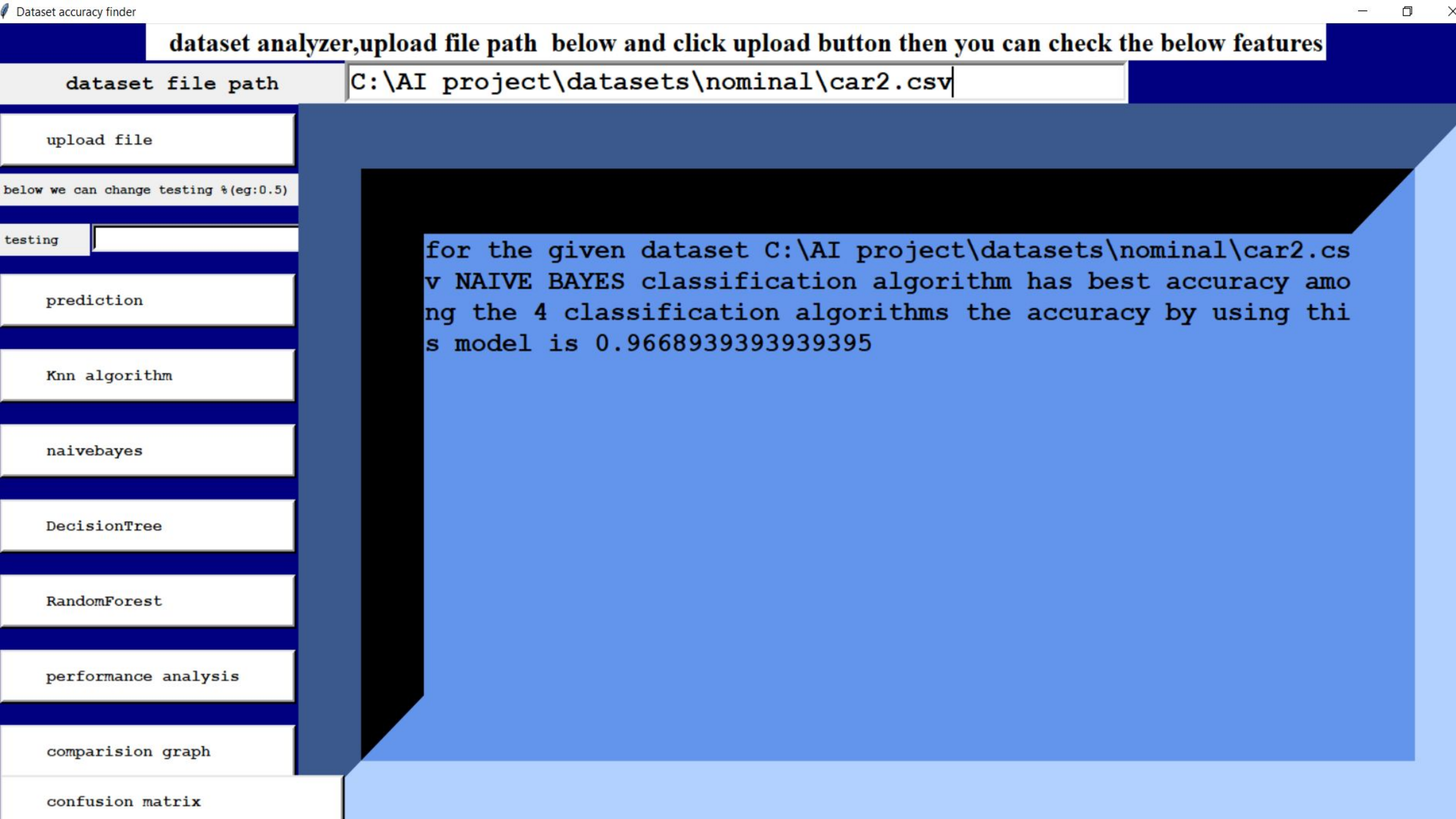
Note:- A is parent node of B and C.

Random Forest



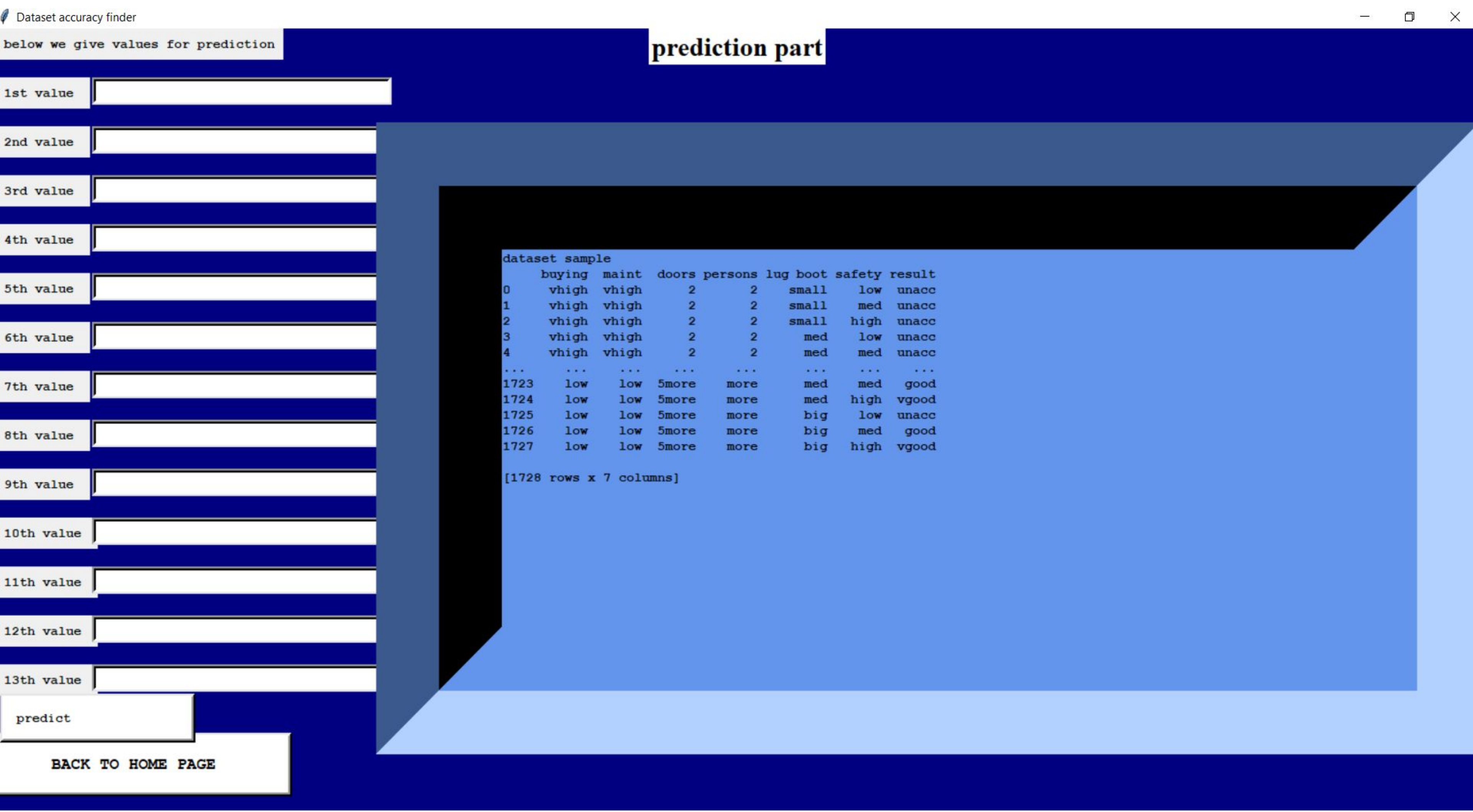
Tools and Technologies

- Jupyter
- Python Libraries
  - Pandas
  - Sklearn
  - Tkinter
  - Matplotlib
  - Numpy



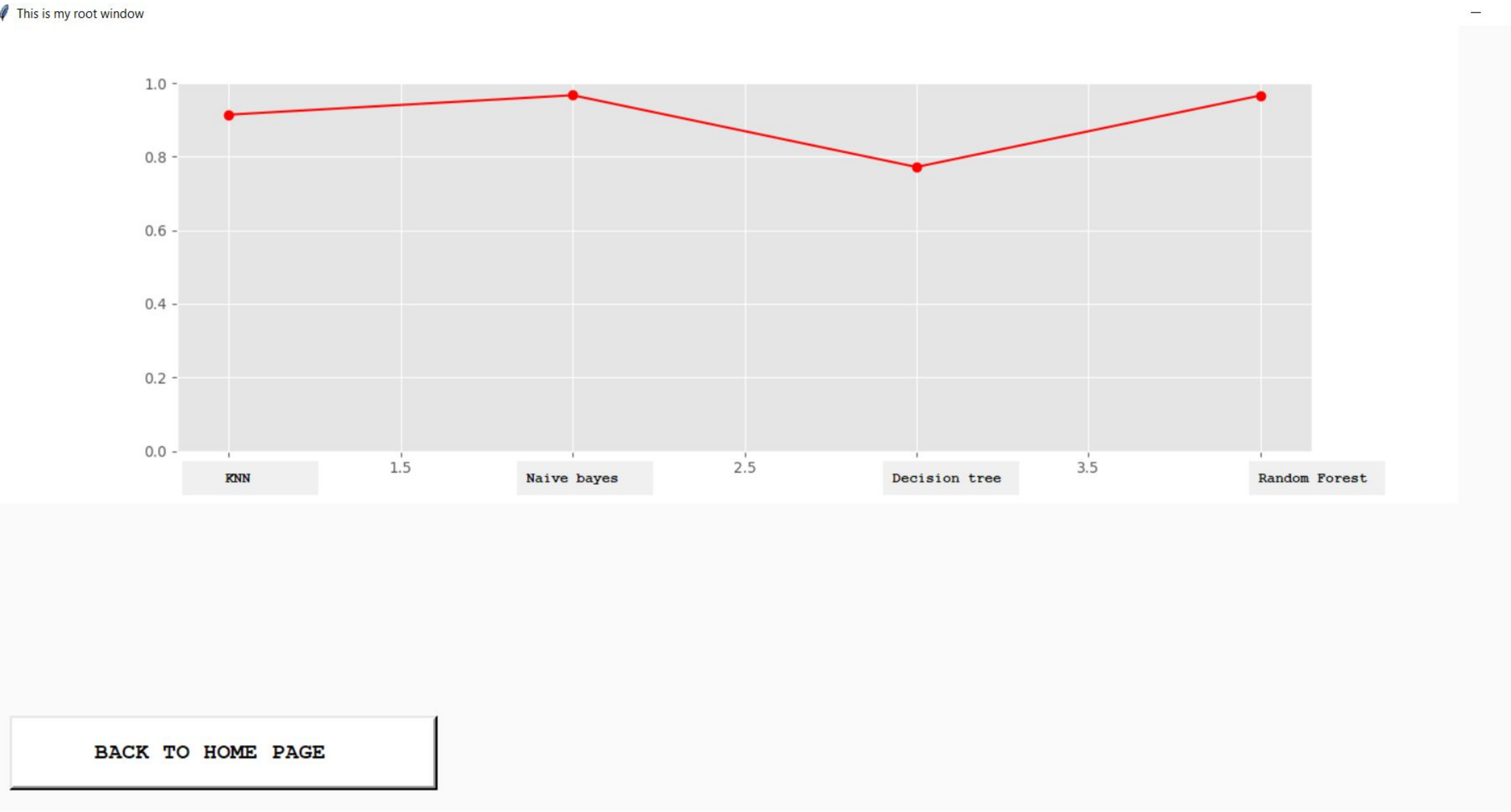
dataset file path: C:\AI project\datasets\nominal\car2.csv

```
for the given dataset C:\AI project\datasets\nominal\car2.cs
v NAIVE BAYES classification algorithm has best accuracy amo
ng the 4 classification algorithms the accuracy by using thi
s model is 0.9668939393939395
```

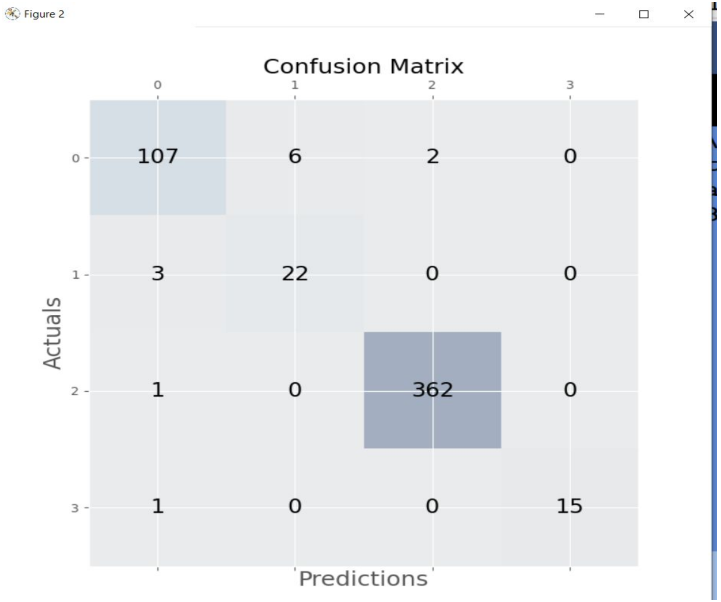


```
dataset sample
     buying  maint  doors persons lug boot safety result
0     vhigh  vhigh      2       2    small    low  unacc
1     vhigh  vhigh      2       2    small    med  unacc
2     vhigh  vhigh      2       2    small   high  unacc
3     vhigh  vhigh      2       2      med    low  unacc
4     vhigh  vhigh      2       2      med    med  unacc
...     ...    ...    ...     ...      ...    ...    ...
1723    low    low  5more    more      med    med   good
1724    low    low  5more    more      med   high  vgood
1725    low    low  5more    more      big    low  unacc
1726    low    low  5more    more      big    med   good
1727    low    low  5more    more      big   high  vgood

[1728 rows x 7 columns]
```


This is my root window
1.0
0.8
0.6
0.4
0.2
0.0
KNN
1.5
Naive bayes
2.5
Decision tree
3.5
Random Forest
BACK TO HOME PAGE

Confusion Matrix



	0	1	2	3
0	107	6	2	0
1	3	22	0	0
2	1	0	362	0
3	1	0	0	15

Dataset Types

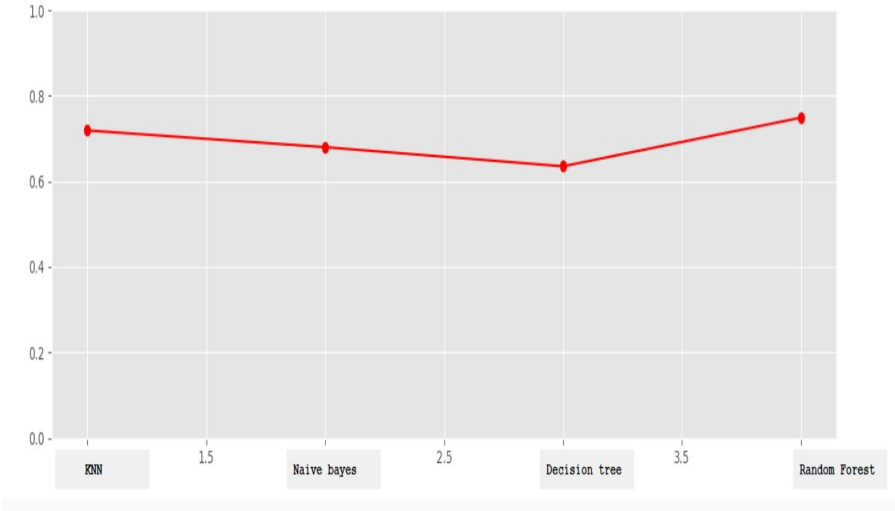
- Based on content
  - Mixed (with text and numbers)
  - Numeric (with only numbers)
  - Nominal (with only text)
- Based on size
  - Small (<500)
  - Medium (<10000)
  - Large (>10000)

Diabetes Prediction- Numeric - Medium-786 Records

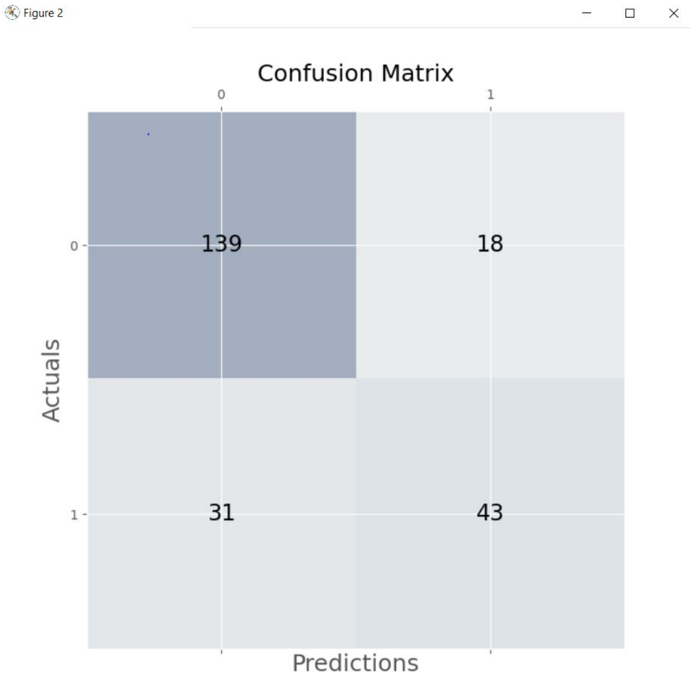
NO	NAME OF ATTRIBUTES	TYPE
1	Number	Numeric
2	Glucose	Numeric
3	Blood Pressure(mm HG)	Numeric
4	Skin Thickness	Numeric
5	Insulin	Numeric
6	Body Mass Index(BMI)	Numeric
7	Diabetes Pedigree function	Numeric
8	Age(years)	Numeric
9	Outcome	Numeric(0 or 1)

Accuracies

● KNN	: 0.718
● Naive Bayes	: 0.679
● Decision Tree	: 0.635
● Random Forest	: 0.748



Confusion matrix For best algorithm



	0	1
0	139	18
1	31	43

Numeric - Other Datasets

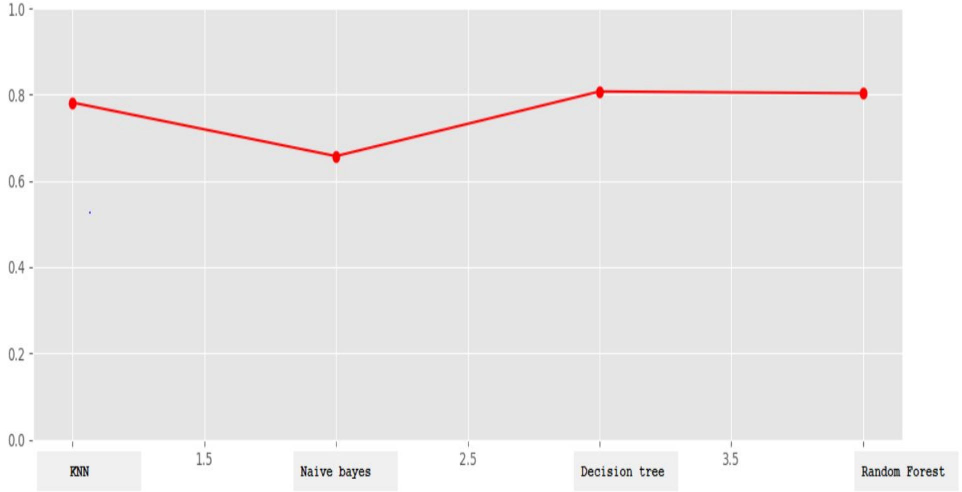
- Prediction of sepsis - 110k Records - Large
  - Accuracies:
    - KNN : 0.9179
    - Naive Bayes : 0.9261
    - Decision Trees : 0.9266
    - Random Forest: 0.9263
- Heart - 300 Records - Small
  - Accuracies:
    - KNN : 0.806
    - Naive Bayes : 0.778
    - Decision Trees : 0.821
    - Random Forest: 0.816

German Credit Data- Mixed- small- 600 Records

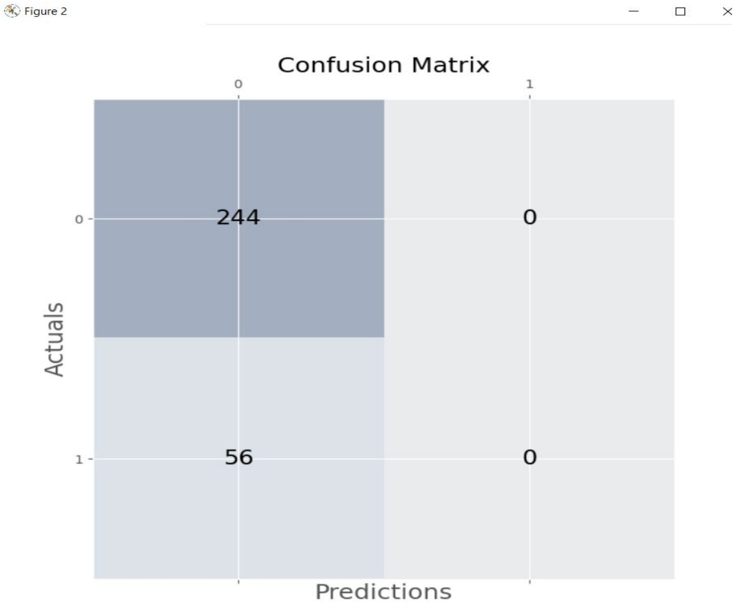
NO	NAME OF ATTRIBUTES	TYPE
1	Age	Numeric
2	Sex	Text
3	No of Jobs	Numeric
4	Housing	Numeric
5	Savings	Text
6	Checkings	Text
7	Credit Amount	Numeric
8	Duration	Numeric
9	Purpose	Text(comfort or essential)

Accuracies

- KNN : 0.781
- Naive Bayes : 0.657
- Decision Tree : 0.807
- Random Forest : 0.802



Confusion matrix For best algorithm



	0	1
0	244	0
1	56	0

Mixed - Other Datasets

- Power System - 11k Records - Large
    - Accuracies:
        - KNN : 0.928
        - Naive Bayes : 0.977
        - Decision Trees : 0.977
        - Random Forest : 0.999
- Abalone - 4k Records - Medium
    - Accuracies:
        - KNN : 0.523
        - Naive Bayes : 0.479
        - Decision Trees : 0.365
        - Random Forest : 0.540

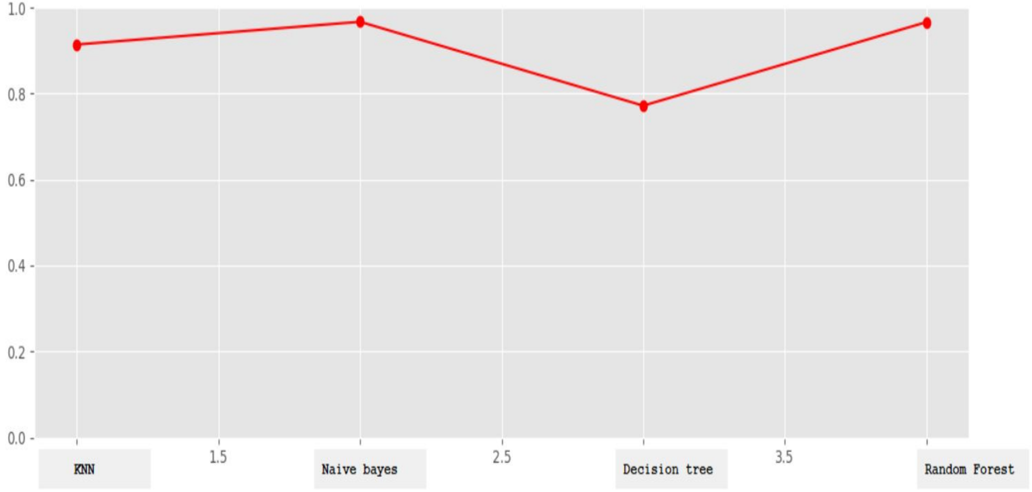
Nominal - Car Evaluation - Medium -1800 Records

NO	NAME OF ATTRIBUTES	TYPE
1	Buying	Nominal
2	Maintenance	Nominal
3	Doors	Nominal
4	Persons	Nominal
5	Luggage Boot	Nominal
6	Safety	Nominal

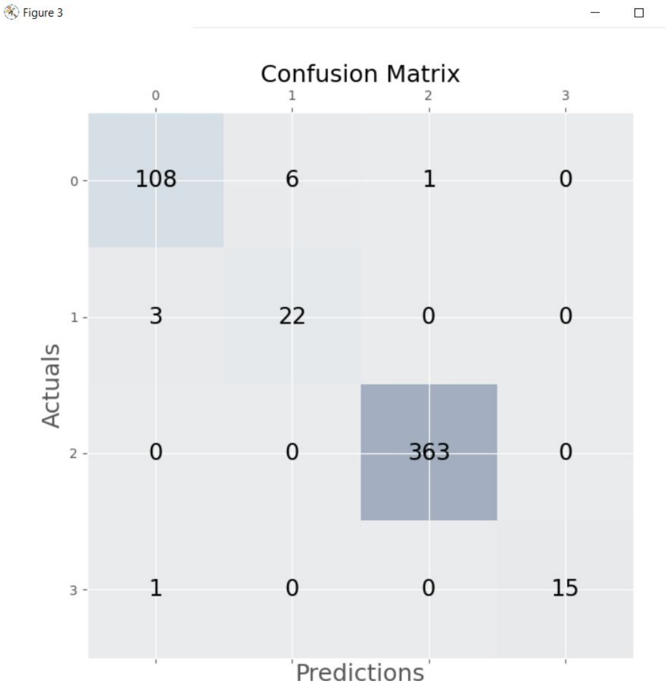
buying	v-high, high, med, low
maintenance	v-high, high, med, low
doors	2, 3, 4,5
persons	2,4,5
luggage boot	small, med, med
safety	low, med, high

Accuracies

KNN	: 0.913
Naive Bayes	: 0.966
Decision Tree	: 0.771
Random Forest	: 0.966



Confusion matrix For best algorithm



	0	1	2	3
0	108	6	1	0
1	3	22	0	0
2	0	0	363	0
3	1	0	0	15

Nominal - Other Datasets

- Perform- 500 Records - small
  - Accuracies:
    - KNN : 0.634
    - Naive Bayes : 0.669
    - Decision Trees : 0.675
    - Random Forest: 0.722
- Animals - 100k Records - Large
  - Accuracies:
    - KNN : 0.971
    - Naive Bayes : 0.969
    - Decision Trees : 0.785
    - Random Forest : 0.972

Conclusion

	Small	Medium	Large
Numeric	Decision Tree/ Random Forest	Random Forest	Decision Tree/ Random Forest
Mixed	Decision Tree/ Random Forest	Random Forest	Naive Bayes/ Decision Tree
Nominal	Random Forest	Random Forest/ Decision Tree	Random Forest

Future Work

- Implement many more datasets
- Improve GUI
- Implement other classification algorithms

Thank You