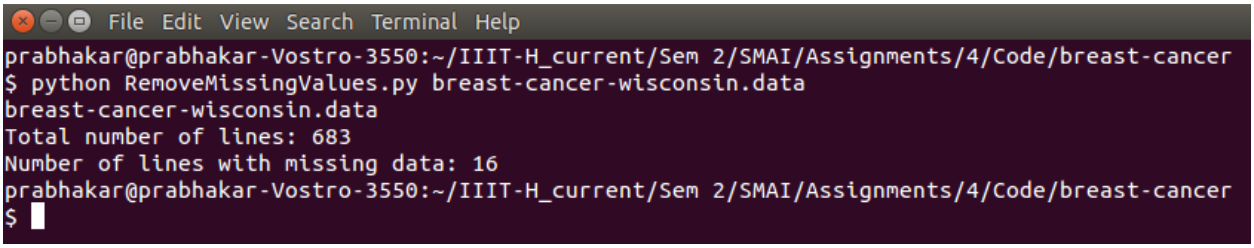
B. Prabhakar

201505618

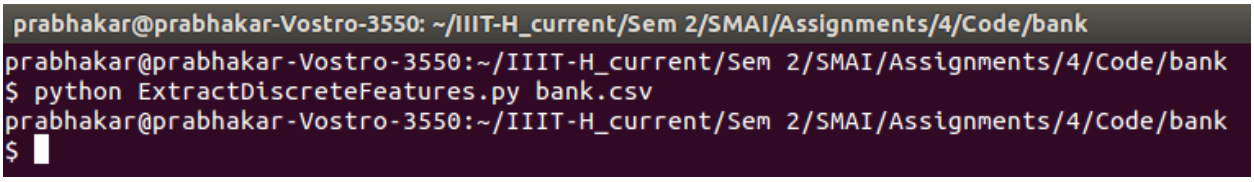
Assignment IV

Experiment 1

$ python RemoveMissingValues.py breast-cancer-wisconsin.data

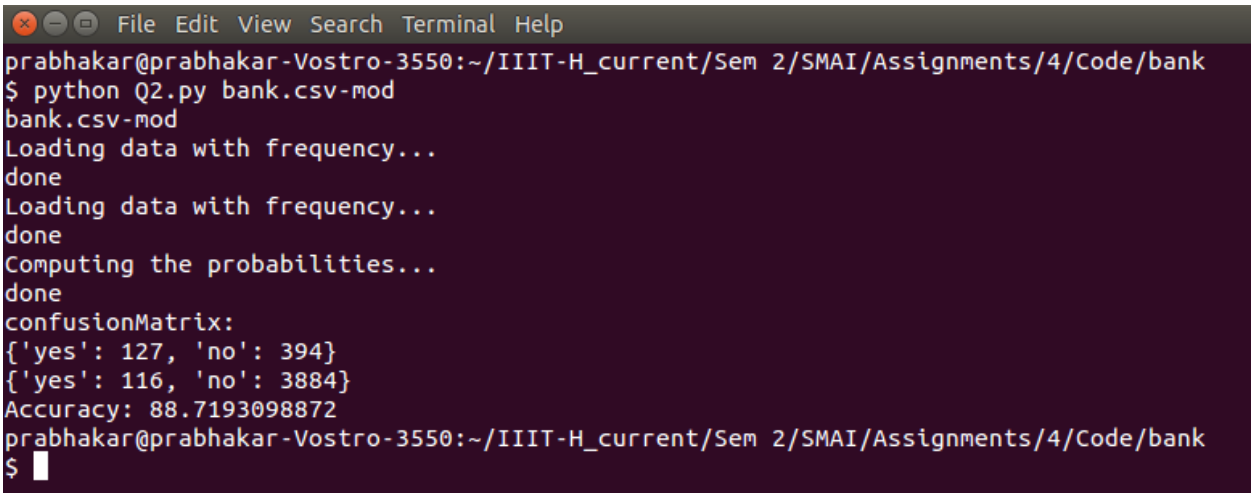


$ python ExtractDiscreteFeatures.py bank.csv



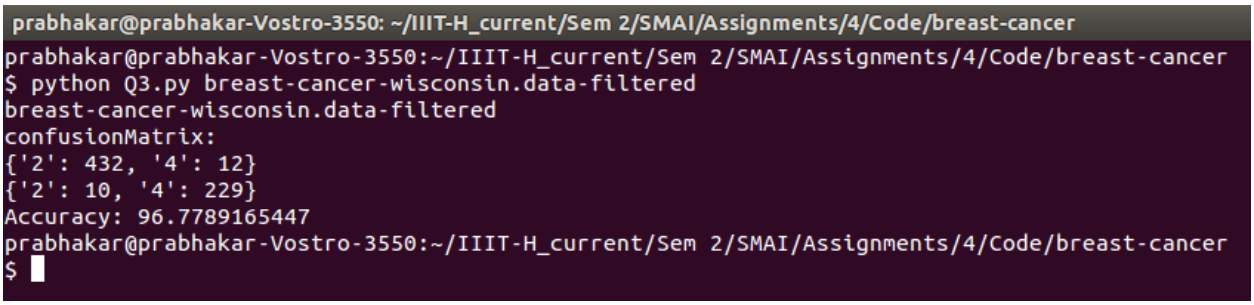
Experiment 2, 10

$ python Q2.py bank.csv-mod



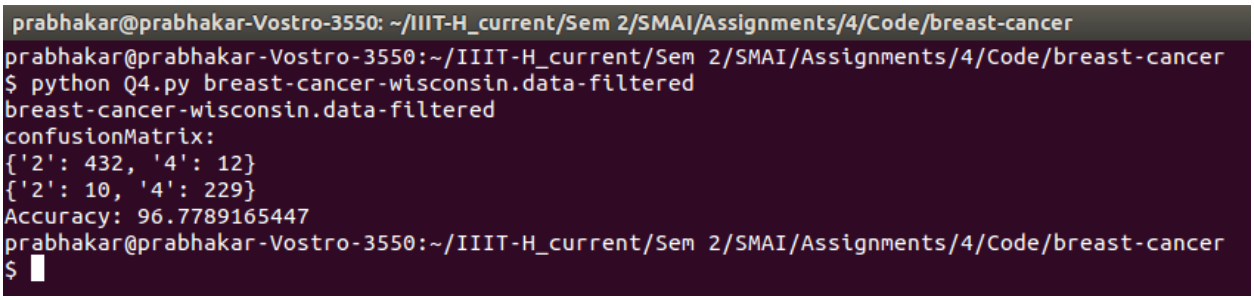
Experiment 3, 10

$ python Q3.py breast-cancer-wisconsin.data-filtered



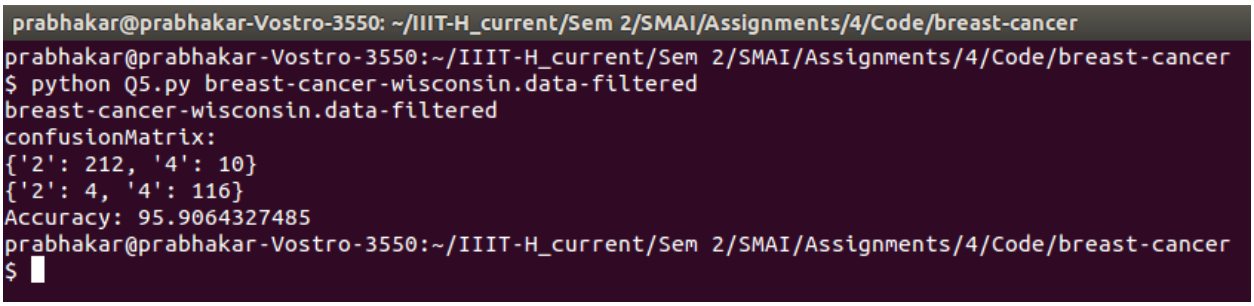
Experiment 4

$ python Q4.py breast-cancer-wisconsin.data-filtered



Experiment 5,6,7

$ python Q5.py breast-cancer-wisconsin.data-filtered



Experiment 8

$ python Q8.py breast-cancer-wisconsin.data-filtered

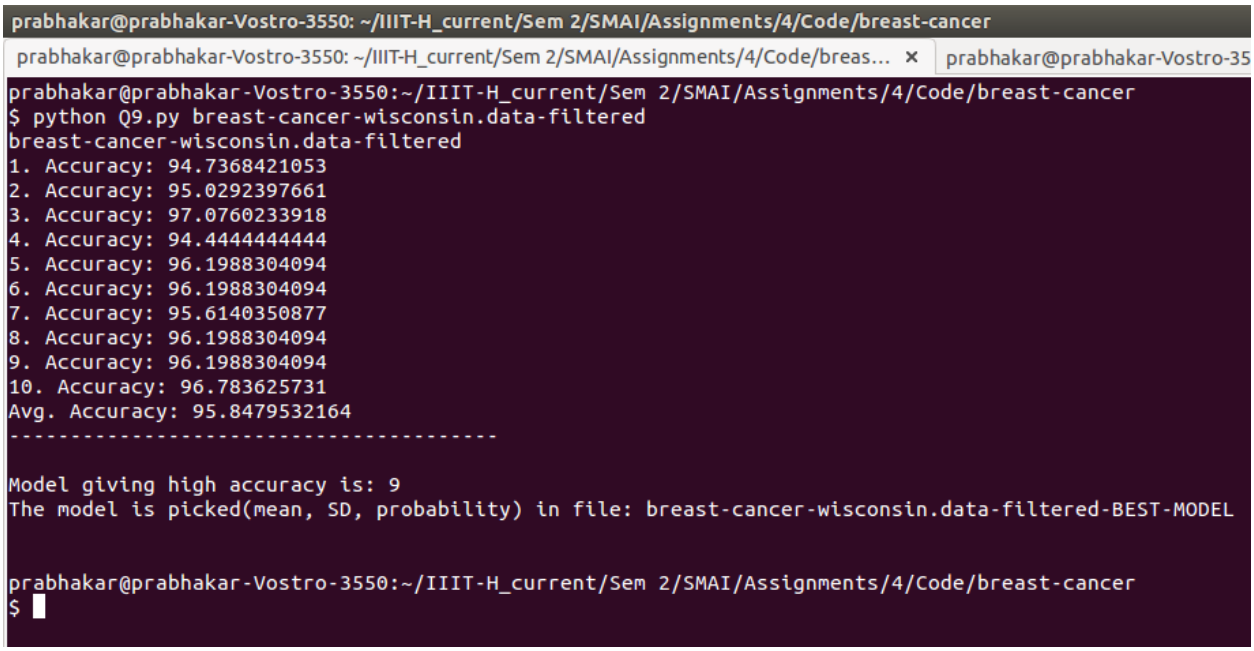
```
prabhakar@prabhakar-Vostro-3550: ~/IIIT-H_current/Sem 2/SMAI/Assignments/4 Naive Bayes (NB) CLASS
prabhakar@prabhakar-Vostro-3550:~/IIIT-H_current/Sem 2/SMAI/Assignments/4 Naive
Bayes (NB) CLASSIFIER/Code/breast-cancer
$ python Q8.py breast-cancer-wisconsin.data-filtered
breast-cancer-wisconsin.data-filtered
1. Accuracy: 96.1988304094
2. Accuracy: 96.783625731
3. Accuracy: 95.3216374269
4. Accuracy: 97.6608187135
5. Accuracy: 95.3216374269
6. Accuracy: 94.4444444444
7. Accuracy: 96.4912280702
8. Accuracy: 95.6140350877
9. Accuracy: 94.4444444444
10. Accuracy: 95.3216374269
Avg. Accuracy: 95.7602339181
Standard Dev.: 0.971974861906
prabhakar@prabhakar-Vostro-3550:~/IIIT-H_current/Sem 2/SMAI/Assignments/4 Naive
Bayes (NB) CLASSIFIER/Code/breast-cancer
$
```

$ python Q8.py bank.csv-mod

```
prabhakar@prabhakar-Vostro-3550:~/IIIT
$ python Q8.py bank.csv-mod
bank.csv-mod
1. Accuracy: 88.1637168142
2. Accuracy: 87.389380531
3. Accuracy: 87.8318584071
4. Accuracy: 86.9469026549
5. Accuracy: 88.6061946903
6. Accuracy: 88.0530973451
7. Accuracy: 88.9380530973
8. Accuracy: 88.6061946903
9. Accuracy: 88.6061946903
10. Accuracy: 89.3805309735
Avg. Accuracy: 88.2522123894
Standard Dev.: 0.692233975538
```

Experiment 9

$ python Q9.py breast-cancer-wisconsin.data-filtered



Questions

Question 1:

Dataset: breast-cancer-wisconsin.data-filtered

Features: [5, 4, 4, 5, 7, 10, 3, 2, 1]

Prob: {'2': 3.9406537264581124e-17, '4': 1.0}

Prediction: 4

class: 2

Features: [6, 8, 8, 1, 3, 4, 3, 7, 1]

Prob: {'2': 2.0487223717408941e-21, '4': 1.0}

Prediction: 4

class: 2

Features: [3, 1, 1, 3, 8, 1, 5, 8, 1]

Prob: {'2': 0.00032719585990578917, '4': 0.9996728041400943}

Prediction: 4

class: 2

Dataset: bank.csv-mod

Features: ['student' 'single' 'secondary' 'no' 'no' 'no' 'cellular' 'apr' 'unknown']

Prob: {'yes': 1.9397681910904754e-05, 'no': 2.0540225518229928e-05}

Prediction: no

class: yes

Features: ['retired' 'divorced' 'secondary' 'no' 'no' 'no' 'telephone' 'jul'
 'unknown']

Prob: {'yes': 2.9071078295234765e-06, 'no': 6.8231395083710569e-06}

Prediction: no

class: yes

Features: ['management' 'single' 'tertiary' 'no' 'yes' 'no' 'cellular' 'aug'
 'unknown']

Prob: {'yes': 0.00010369572533677281, 'no': 0.00048144202303075287}

Prediction: no

class: yes

The features that are associated the above misclassified samples, are acting like an outliers for the class 2. Hence the Bayesian Classifier has misclassified them.

Question 2:

In case if *a prior* probability for feature given a class turns out to be zero, the posterior probability for all the novel patterns in which the feature appears will also turn out to be zero. This may lead to misclassification of all such novel patterns, which is not a desirable thing.

In order to avoid such thing, we add 1 to the numerator of the probability of all the features given a class. However, this addition will lead to violation of the probability law, “Sum of all the probabilities quantize to 1”. For making this quantization happen, we add the “Count of the classes” to the denominator. This is called *Laplacian smoothing.*

Question 3:

If dataset contains mixed values, while finding the *a prior probabilities* consider the frequency of each discrete features, while considering the value of the continuous features.

Question 4:

If the feature for which the value is missing is a continuous value, then consider the missing value as zero.

If the feature for which the value is missing is a discrete value, then consider the missing value as “-”.