Artificial Intelligence - Project 05: Machine Learning

TEAM:
Renu Rani (111482474)
Anurag Arora (111425080)
Shayan Ray (111424665)
Gourab Bhattacharyya (170048888)

Q1: Clickstream Mining with Decision Trees

```
For each value of threshold p-value
-----------------------------------

p = 0.01
----------

python q1_classifier.py  -p 0.01 -f1 train.csv -f2 test.csv -o output.csv -t tree.pkl
Test 25000 sample
Done predict test set
Node in the decision tree is: 266
Prediction is :  25000
Output files generated

Autograder Results:
--------------------
python autograder_basic.py
Data Loading: done
Tree prediction accuracy:  0.74832
Output file prediction accuracy:  0.72928



p = 0.05
---------
python q1_classifier.py -p 0.05 -f1 train.csv -f2 test.csv -o output.csv -t tree.pkl
Done predict test set
Node in the decision tree is: 420
Prediction is :  25000
Output files generated

Autograder Results:
```

```
--------------------
python autograder_basic.py
Data Loading: done
Tree prediction accuracy:  0.74832
Output file prediction accuracy:  0.73104
```

```
p = 1.0
----------
python q1_classifier.py -p 1 -f1 train.csv -f2 test.csv -o output.csv -t tree.pkl
Test 25000 sample
Done predict test set
Node in the decision tree is: 17120
Prediction is :  25000
Output files generated
```

```
Autograder Results:
--------------------
python autograder_basic.py
Data Loading: done
Tree prediction accuracy:  0.74832
Output file prediction accuracy:  0.66588
```

Observations:

Accuracy is almost ~75% for p =0.01 and 0.05 which is fair. For p = 1.0 it is almost ~65%.
As obvious as the p-value increases the number of nodes to be expanded will increase, the dataset used for training and testing
will increase. This would have more and more uncertain data introduced into the algorithm and test out its generalizability.
Hence the accuracy goes down a bit around p=1.0

Also note, smaller decision trees(lower p-values) would reduce the chances of overfitting the training data and hence the accuracy will be more.
ID3 only works on discrete values and does not work on continuous data.

Improvements:

Initially when the accuracies were lower than this, the improvements made were as follows:
1. Entropy computation was casted to floating point explicitly.
2. Accuracy calculations with True Positives, False Positives etc were rectified and improved.

3. During prediction, sometimes division by zero was encountered, that issue upon resolution improved accuracy.

Ans 2:

All the improvements mentioned above worked well and improved accuracy.
Now ID3 does not guarentee an optimal solution always. It is a greedy approach and can get stuck in local optima.
One improvement that can be done is use backtracking during search for an optimal decision tree.
This would produce smaller trees than without it. Hence the chances of overfitting will be reduced and the model would become
more generalizable, resulting in better prediction accuracy.

Q2: Spam Filter

We used Multinomial Naïve Bayes for Spam/Ham filtering. We tried different values for smoothing factor (alpha) including Laplace and Lidstone.

alpha >= 1 is called Laplace smoothing, while alpha < 1 is called Lidstone smoothing.

Value of alpha	Accuracy %
10	81.4
8	83
5	86.5
2	89.5
1	90.1
0.8	90.3
0.5	90.7
0.1	91.5
0.05	91.9

We used alpha as 0.05 and get accuracy 91.9%.