CS205 Project2

NAME Zijian Li
SID 862545559
URL of my code: git@github.com:DKZF91/CS205-project2.git

File Structure

There are two python files and three data files.
Project2_part1.py: This file contains the code to complete the part 1 task, including the binary classification KNN algorithm, leaving-one-out algorithm, forward selection algorithm, and backward elimination algorithm. There are also some user interaction codes (including asking the user to select the data file to be processed and the algorithm to be selected).
Project2_part2.py: This file includes the code to complete part 2, including cleaning and organizing the data, and referring to the code in part 1 to complete the classification task.
CS205_large_Data__49.txt: This file is the large dataset I should choose.
CS205_small_Data__38.txt: This file is the small dataset I should choose.
Algerian_forest_fires_dataset_UPDATE.csv: This dataset is the one I chose from the link given in the assignment instructions. I chose to use pandas to import the data instead of directly referencing the data with import, see the code *Project2_part2.py* for details. This dataset is a binary classification dataset. It includes 14 features and correctly classified labels. It includes 244 instances.

How to run

Part 1

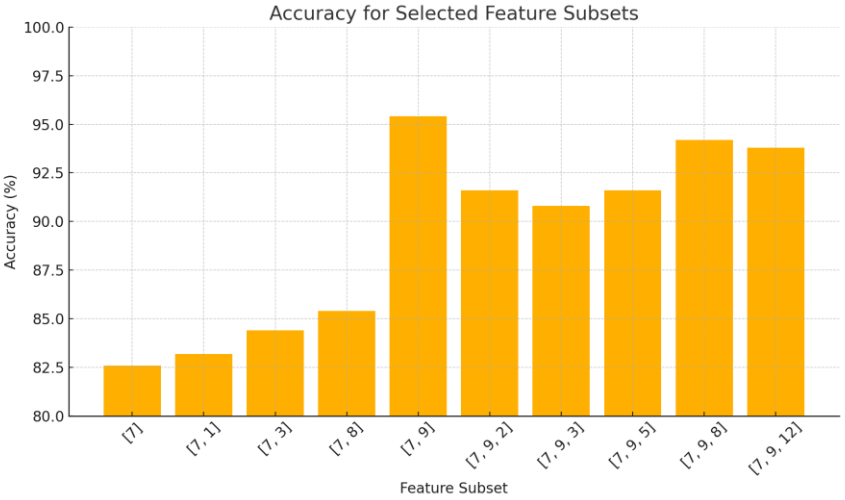
1. Please make sure that the two datasets and the python file are in the same directory. Then run the code directly.
2. Enter the file name as required (no need to enter the directory) and select the model (1 for Forward Selection, 2 for Backward Elimination)
3. Wait for the terminal to output the results (it will take a long time to run Backward Elimination using *CS205_large_Data__49.txt,* please be patient)

Part 2

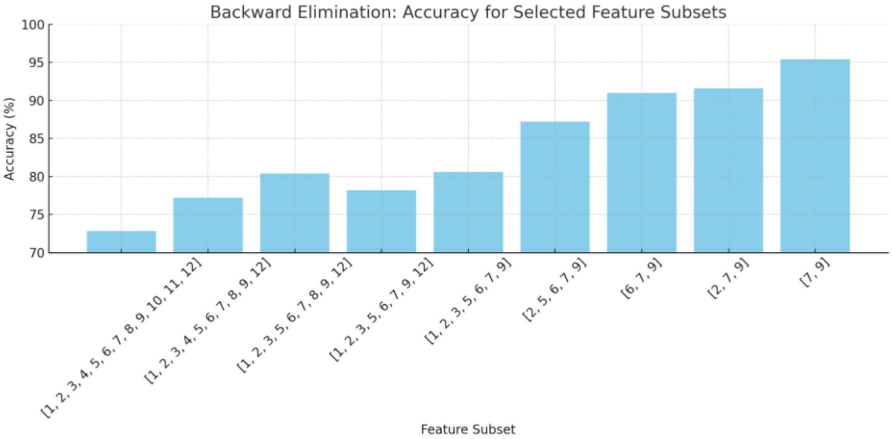
1. Make sure the two Python files and the dataset are in the same directory. Run the code directly. Wait for the terminal to output the results.

Results

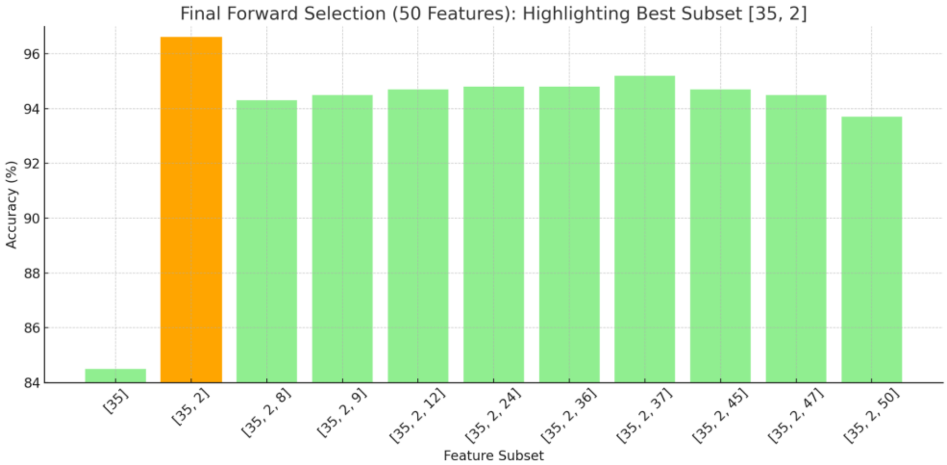
Part 1



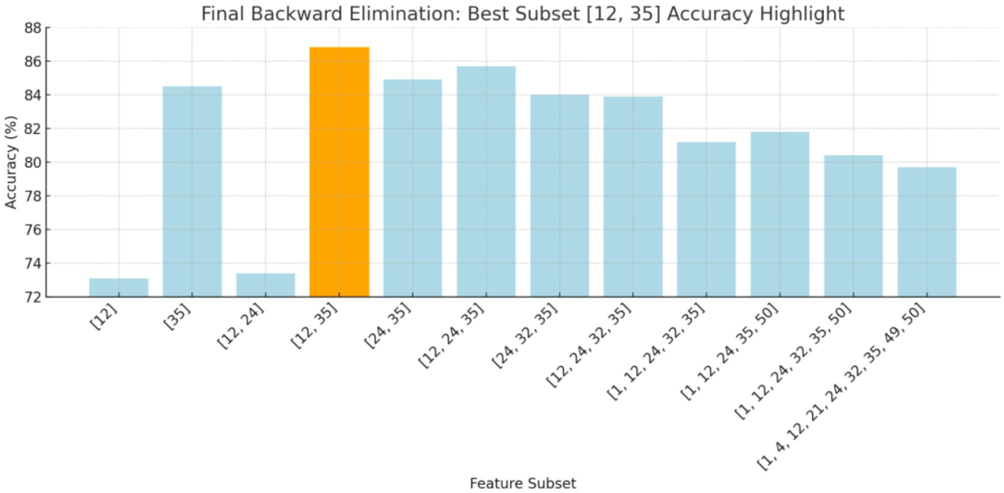
As shown in the figure, this is the accuracy for selected feature subsets obtained by Forward Selection for the CS205_small_Data__38.txt dataset. In order to save space, some data has been omitted and only important data is retained. From the figure, we can see that the set with the highest accuracy is [7,9], with an accuracy of 95.4%.



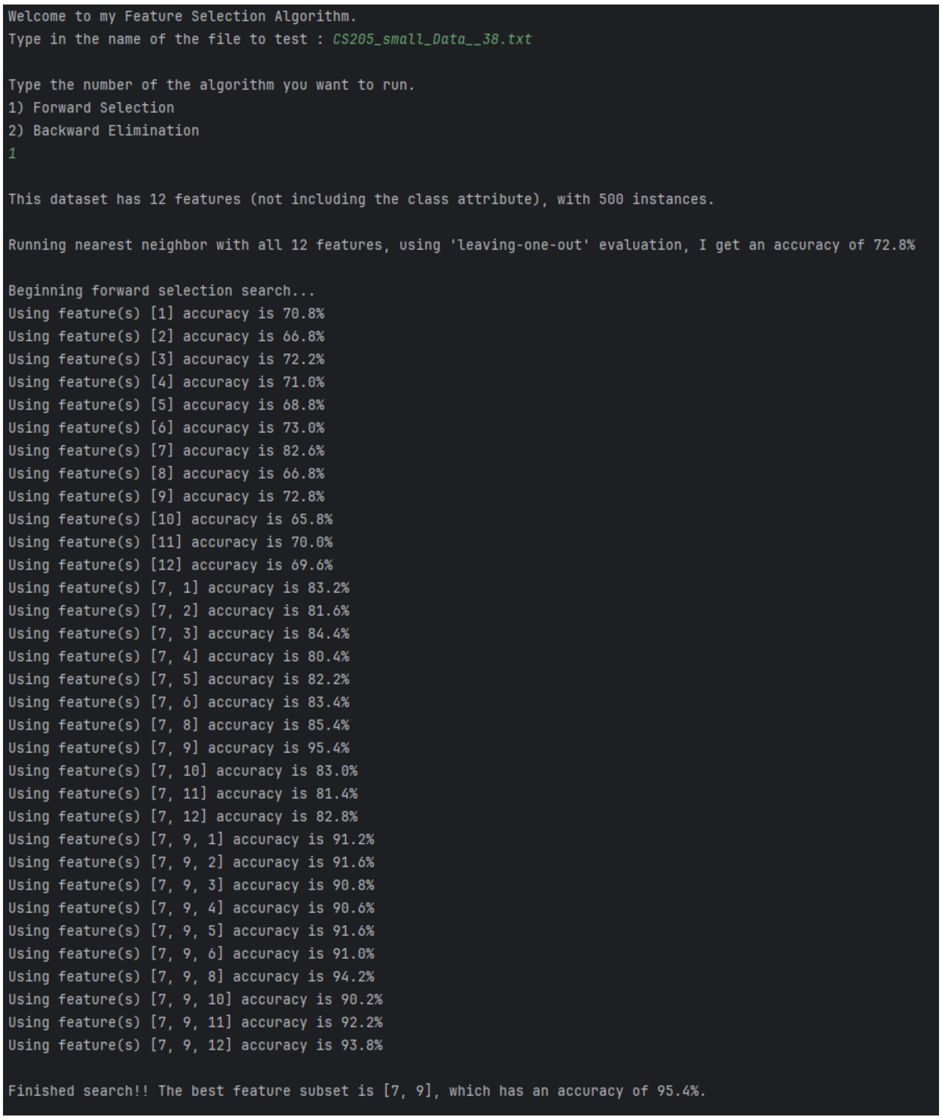
As shown in the figure, this is the accuracy for selected feature subsets obtained by Backward Elimination for the CS205_small_Data__38.txt dataset. In order to save space, some data has been omitted and only important data is retained. From the figure, we can see that the set with the highest accuracy is [7,9], with an accuracy of 95.4%.



As shown in the figure, this is the accuracy for selected feature subsets obtained by Forward Selection for the CS205_large_Data__49.txt dataset. In order to save space, some data has been omitted and only important data is retained. From the figure, we can see that the set with the highest accuracy is [35,2], with an accuracy of 96.6%.



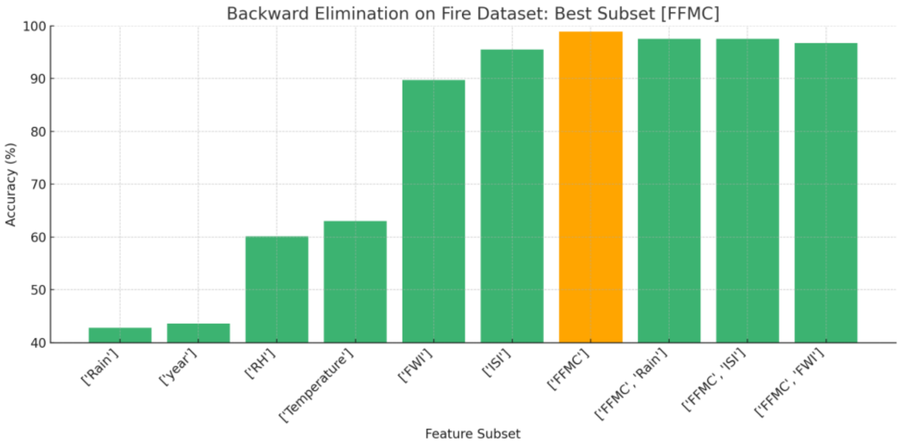
As shown in the figure, this is the accuracy for selected feature subsets obtained by Backward Elimination for the CS205_large_Data__49.txt dataset. In order to save space, some data has been omitted and only important data is retained. From the figure, we can see that the set with the highest accuracy is [12,35], with an accuracy of 86.8%.



```
Welcome to my Feature Selection Algorithm.
Type in the name of the file to test : CS205_small_Data__38.txt

Type the number of the algorithm you want to run.
1) Forward Selection
2) Backward Elimination
1

This dataset has 12 features (not including the class attribute), with 500 instances.

Running nearest neighbor with all 12 features, using 'leaving-one-out' evaluation, I get an accuracy of 72.8%

Beginning forward selection search...
Using feature(s) [1] accuracy is 70.8%
Using feature(s) [2] accuracy is 66.8%
Using feature(s) [3] accuracy is 72.2%
Using feature(s) [4] accuracy is 71.0%
Using feature(s) [5] accuracy is 68.8%
Using feature(s) [6] accuracy is 73.0%
Using feature(s) [7] accuracy is 82.6%
Using feature(s) [8] accuracy is 66.8%
Using feature(s) [9] accuracy is 72.8%
Using feature(s) [10] accuracy is 65.8%
Using feature(s) [11] accuracy is 70.0%
Using feature(s) [12] accuracy is 69.6%
Using feature(s) [7, 1] accuracy is 83.2%
Using feature(s) [7, 2] accuracy is 81.6%
Using feature(s) [7, 3] accuracy is 84.4%
Using feature(s) [7, 4] accuracy is 80.4%
Using feature(s) [7, 5] accuracy is 82.2%
Using feature(s) [7, 6] accuracy is 83.4%
Using feature(s) [7, 8] accuracy is 85.4%
Using feature(s) [7, 9] accuracy is 95.4%
Using feature(s) [7, 10] accuracy is 83.0%
Using feature(s) [7, 11] accuracy is 81.4%
Using feature(s) [7, 12] accuracy is 82.8%
Using feature(s) [7, 9, 1] accuracy is 91.2%
Using feature(s) [7, 9, 2] accuracy is 91.6%
Using feature(s) [7, 9, 3] accuracy is 90.8%
Using feature(s) [7, 9, 4] accuracy is 90.6%
Using feature(s) [7, 9, 5] accuracy is 91.6%
Using feature(s) [7, 9, 6] accuracy is 91.0%
Using feature(s) [7, 9, 8] accuracy is 94.2%
Using feature(s) [7, 9, 10] accuracy is 90.2%
Using feature(s) [7, 9, 11] accuracy is 92.2%
Using feature(s) [7, 9, 12] accuracy is 93.8%

Finished search!! The best feature subset is [7, 9], which has an accuracy of 95.4%.
```

Here is an image of the complete results for reference.

Part 2



As shown in the figure, this is the accuracy for selected feature subsets obtained by Backward Elimination for the

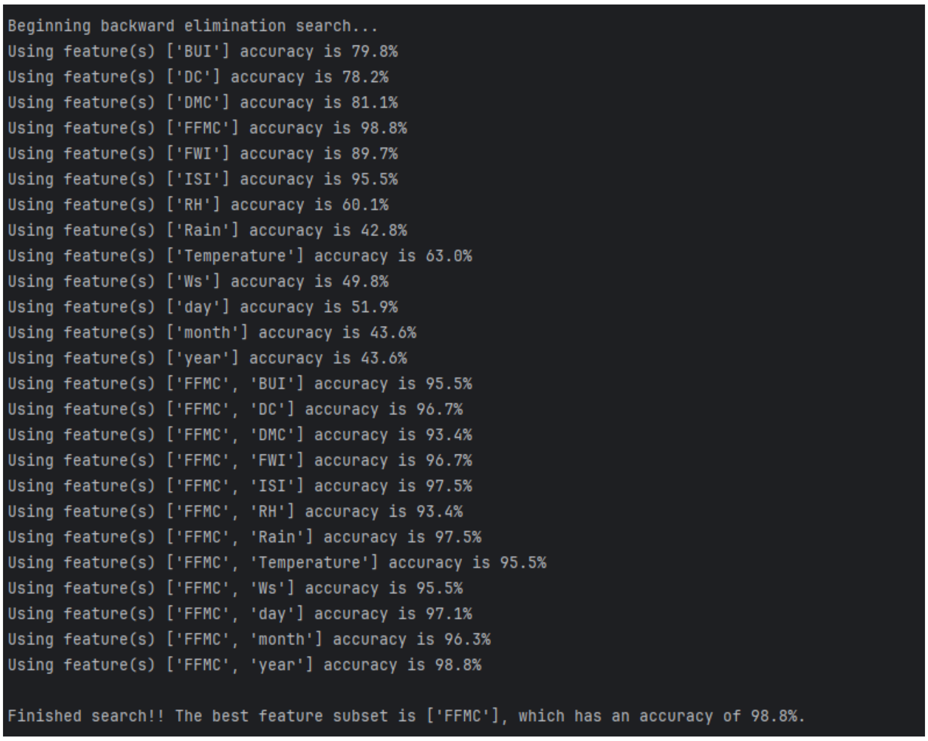
Algerian_forest_fires_dataset_UPDATE.csv dataset. In order to save space, some data has been omitted and only important data is retained. From the figure, we can see that the set with the highest accuracy is [‘FFMC’], with an accuracy of 98.8%.

First of all, we need to understand the meaning of each feature in the data set.

1. Date : (DD/MM/YYYY) Day, month ('june' to 'september'), year (2012)
2. Temp : temperature noon (temperature max) in Celsius degrees: 22 to 42
3. RH : Relative Humidity in %: 21 to 90
4. Ws :Wind speed in km/h: 6 to 29
5. Rain: total day in mm: 0 to 16.8
6. Fine Fuel Moisture Code (FFMC) index from the FWI system: 28.6 to 92.5
7. Duff Moisture Code (DMC) index from the FWI system: 1.1 to 65.9
8. Drought Code (DC) index from the FWI system: 7 to 220.4
9. Initial Spread Index (ISI) index from the FWI system: 0 to 18.5
10. Buildup Index (BUI) index from the FWI system: 1.1 to 68
11. Fire Weather Index (FWI) Index: 0 to 31.1

The above variable information comes from the dataset website (see reference). Note that there is a feature called region in the original data (divided into two categories, Bejaia and Sidi-Bel Abbes), and in the original table, the author separated the data of the two places into two tables. I merged the two tables before cleaning the data, so I deleted the region feature (because it is a Categorical type and does not affect the final classification result).

The algorithm first adds the feature Fine Fuel Moisture Code (FFMC). This makes sense because FFMC is used to measure the dryness of flammable fine fuels (such as dead plant matter on the surface, such as fallen leaves, hay, needles, etc.) [d]. Usually when the FFMC value is high, it means that the fine fuel is very easy to burn. This is also the most direct cause of natural forest fires. According to the paper [a], FFMC is calculated based on meteorological factors (temperature, relative humidity, wind speed and rainfall). Therefore, with the FFMC indicator, we do not need natural factors such as temperature and rainfall. The date factor is not considered because it does not affect the prediction. Therefore, we only need one feature FFMC to get the highest accuracy result: 98.8%.



```
Beginning backward elimination search...
Using feature(s) ['BUI'] accuracy is 79.8%
Using feature(s) ['DC'] accuracy is 78.2%
Using feature(s) ['DMC'] accuracy is 81.1%
Using feature(s) ['FFMC'] accuracy is 98.8%
Using feature(s) ['FWI'] accuracy is 89.7%
Using feature(s) ['ISI'] accuracy is 95.5%
Using feature(s) ['RH'] accuracy is 60.1%
Using feature(s) ['Rain'] accuracy is 42.8%
Using feature(s) ['Temperature'] accuracy is 63.0%
Using feature(s) ['Ws'] accuracy is 49.8%
Using feature(s) ['day'] accuracy is 51.9%
Using feature(s) ['month'] accuracy is 43.6%
Using feature(s) ['year'] accuracy is 43.6%
Using feature(s) ['FFMC', 'BUI'] accuracy is 95.5%
Using feature(s) ['FFMC', 'DC'] accuracy is 96.7%
Using feature(s) ['FFMC', 'DMC'] accuracy is 93.4%
Using feature(s) ['FFMC', 'FWI'] accuracy is 96.7%
Using feature(s) ['FFMC', 'ISI'] accuracy is 97.5%
Using feature(s) ['FFMC', 'RH'] accuracy is 93.4%
Using feature(s) ['FFMC', 'Rain'] accuracy is 97.5%
Using feature(s) ['FFMC', 'Temperature'] accuracy is 95.5%
Using feature(s) ['FFMC', 'Ws'] accuracy is 95.5%
Using feature(s) ['FFMC', 'day'] accuracy is 97.1%
Using feature(s) ['FFMC', 'month'] accuracy is 96.3%
Using feature(s) ['FFMC', 'year'] accuracy is 98.8%

Finished search!! The best feature subset is ['FFMC'], which has an accuracy of 98.8%.
```

Here is an image of the complete results for reference.

References

[a] Abid, F., & Izeboudjen, N. (2020). Predicting forest fire in algeria using data mining techniques: Case study of the decision tree algorithm. In *Advanced Intelligent Systems for Sustainable Development (AI2SD'2019) Volume 4-Advanced Intelligent Systems for Applied Computing Sciences* (pp. 363-370). Springer International Publishing.

[b] https://archive.ics.uci.edu/dataset/547/algerian+forest+fires+dataset

[c] https://www.geeksforgeeks.org/k-nearest-neighbours/

[d]https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-fwi-system#:~:text=The%20Fine%20Fuel%20Moisture%20Code,It%20ranges%20from%200%2D101.