CS205 Project 2
Feature Selection with Nearest Neighbor

Hugo Wan, 862180666, twan012

Due Date: June 7th, 2025

Cover Page

Outline of this report:

My Github ID: **HugoWan0504** & Link to play test my codes:

`https://github.com/HugoWan0504/CS205_Project_2_Feature_Selection_with_Nearest_Neighbor`

1 Introduction

This is a report that summarizes my findings in feature selection with nearest neighbor using two different algorithms. My choice of coding language: C++:

1. Forward Selection (FS)
2. Backward Elimination (BE)

2 Coding Page

Link to play test my codes:

`https://github.com/HugoWan0504/CS205_Project_2_Feature_Selection_with_Nearest_Neighbor`

Files and their functionalities:

1. **main.cpp**: Prompts the user to select either Forward Selection or Backward Elimination. Asks the user for a dataset file. Loads data using load_data function. Calls either forward_selection or backward_elimination from header_functions.h. Then, writes output to results.csv.
2. **helper_function.h**:
   (a) `load_data`: Loads labels and feature values from dataset.
   (b) `nearest_neighbor_classification`: Implements Leave-One-Out Nearest Neighbor accuracy computation.
   (c) `forward_selection`: Adds features incrementally to improve accuracy. Stops if accuracy drops for LMT rounds. Logs successful and final attempts in results.csv.
   (d) `backward_elimination`: Starts with all features and removes one-by-one to improve accuracy. Also logs progress and final result to results.csv.
3. **result_graph.py**: Accuracy values are displayed inside each bar in bold text. Feature sets are shown below each bar, formatted with 5 features per line and enclosed in braces. The second-to-last bar, representing the final best subset before accuracy decreases, is highlighted in orange, while others remain blue. The legend is positioned top-left outside the plotting area for clarity. The script automatically adjusts bar width and layout based on the number of feature subsets.

3 Part 1 - Assigned Datasets

These are datasets that are assigned to me:

1. CS205_small_Data__26.txt
2. CS205_large_Data__36.txt

Report - Small Dataset

For the **small dataset**, the best **forward selection** subset was **11, 6, 10**, which achieved a high accuracy of **95.4%**. Feature 11 stood out initially with a large jump in accuracy, likely due to its strong predictive power. Features 6 and 10 further improved classification results, while additional features caused performance to drop slightly, suggesting diminishing returns and potential noise.

In contrast, **backward elimination** selected a broader set **2, 3, 4, 6, 7, 9, 10, 11, 12** with a lower final accuracy of **80.2%**. This indicates that some features not included in the forward selection were likely redundant or even harmful. Overall, the forward search algorithm was more effective on this dataset. Both searches for the small dataset finished in **under one minute**.

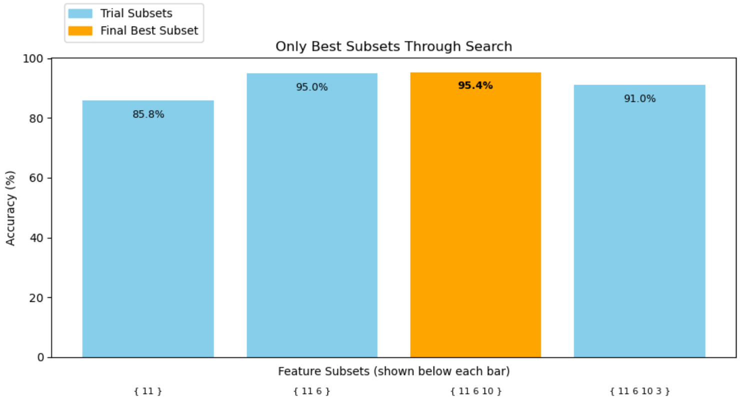


Figure 1: **Small Dataset - Forward Selection**: The best subset was {11, 6, 10}, with an accuracy of **95.4%**.

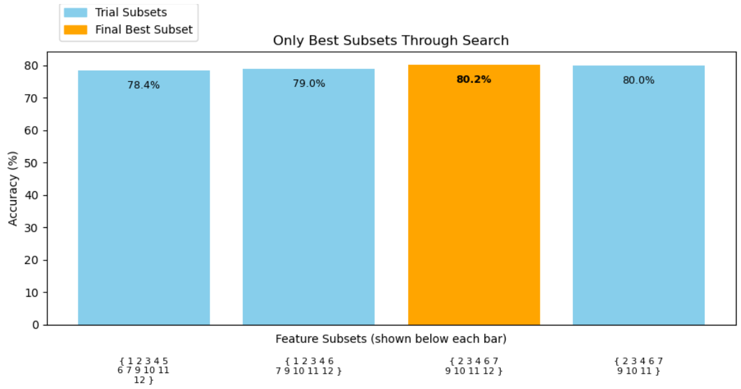


Figure 2: **Small Dataset - Backward Elimination**: The best subset was {2, 3, 4, 6, 7, 9, 10, 11, 12}, achieving an accuracy of **80.2%**.

Report - Large Dataset

For the **large dataset**, **forward selection** chose just two features, **9, 48**, and got a very high accuracy of **97.1%**. Even feature 9 by itself gave a strong result, and adding 48 made it even better. This search finished in **under one minute**.

Backward elimination, on the other hand, kept a much larger subset of **41 features** and reached **73.5%** accuracy. It took about **5 minutes** to run this full backward elimination, which was the longest of all the runs.

In comparison, forward selection for the large dataset finished in under a minute and gave better results with fewer features. This shows that **forward selection worked well for both datasets in terms of both speed and accuracy.**

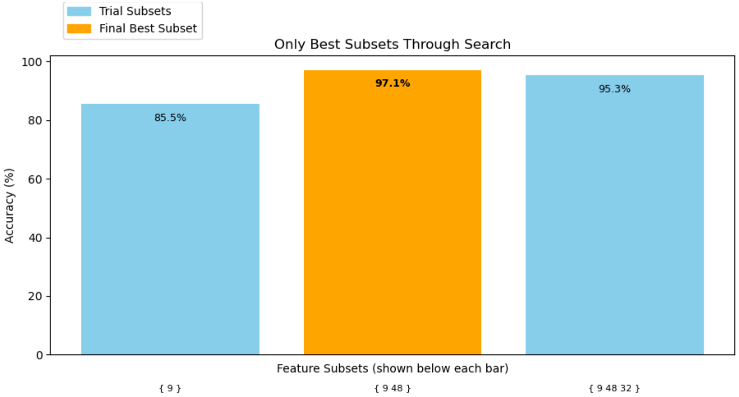


Figure 3: Large Dataset - Forward Selection: The best subset was {9, 48}, achieving an accuracy of **97.1%**.

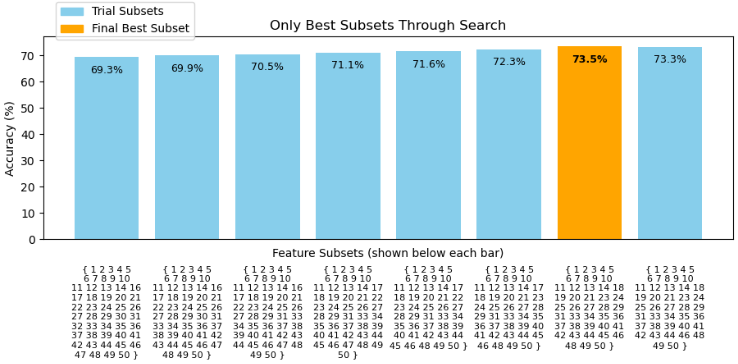


Figure 4: Large Dataset - Backward Elimination: The best subset was all but without the features {15, 16, 17, 22, 30, 32, 47}, with an accuracy of **73.5%**.

FS Traceback of Small Dataset

```
Welcome to Hugo Wan's Feature Selection Algorithm!
Type in the name of the file to test: CS205_small_Data__26.txt
Type the number of the algorithm you want to run.
1) Forward Selection
2) Backward Elimination
Your choice: 1
This dataset has 500 records and 12 features.
Beginning search.
     Current feature(s) { 1 } with accuracy 69.6%
     Current feature(s) { 2 } with accuracy 68.0%
     Current feature(s) { 3 } with accuracy 66.8%
     Current feature(s) { 4 } with accuracy 71.4%
     Current feature(s) { 5 } with accuracy 68.6%
     Current feature(s) { 6 } with accuracy 72.0%
     Current feature(s) { 7 } with accuracy 70.2%
     Current feature(s) { 8 } with accuracy 70.2%
     Current feature(s) { 9 } with accuracy 71.4%
     Current feature(s) { 10 } with accuracy 71.0%
     Current feature(s) { 11 } with accuracy 85.8%
     Current feature(s) { 12 } with accuracy 71.2%
Current best overall is { 11 } with accuracy 85.8%
     Current feature(s) { 11 1 } with accuracy 80.6%
     Current feature(s) { 11 2 } with accuracy 84.8%
     Current feature(s) { 11 3 } with accuracy 82.4%
     Current feature(s) { 11 4 } with accuracy 86.2%
     Current feature(s) { 11 5 } with accuracy 83.2%
     Current feature(s) { 11 6 } with accuracy 95.0%
     Current feature(s) { 11 7 } with accuracy 85.6%
     Current feature(s) { 11 8 } with accuracy 82.8%
     Current feature(s) { 11 9 } with accuracy 84.2%
     Current feature(s) { 11 10 } with accuracy 86.4%
     Current feature(s) { 11 12 } with accuracy 82.6%
Current best overall is { 11 6 } with accuracy 95.0%
     Current feature(s) { 11 6 1 } with accuracy 90.6%
     Current feature(s) { 11 6 2 } with accuracy 90.6%
     Current feature(s) { 11 6 3 } with accuracy 89.6%
     Current feature(s) { 11 6 4 } with accuracy 89.4%
     Current feature(s) { 11 6 5 } with accuracy 92.6%
     Current feature(s) { 11 6 7 } with accuracy 91.0%
     Current feature(s) { 11 6 8 } with accuracy 88.4%
     Current feature(s) { 11 6 9 } with accuracy 92.4%
     Current feature(s) { 11 6 10 } with accuracy 95.4%
     Current feature(s) { 11 6 12 } with accuracy 90.8%
Current best overall is { 11 6 10 } with accuracy 95.4%
```

```
     Current feature(s) { 11 6 10 1 } with accuracy 90.0%
     Current feature(s) { 11 6 10 2 } with accuracy 90.6%
     Current feature(s) { 11 6 10 3 } with accuracy 91.0%
     Current feature(s) { 11 6 10 4 } with accuracy 88.6%
     Current feature(s) { 11 6 10 5 } with accuracy 90.8%
     Current feature(s) { 11 6 10 7 } with accuracy 88.6%
     Current feature(s) { 11 6 10 8 } with accuracy 89.4%
     Current feature(s) { 11 6 10 9 } with accuracy 85.8%
     Current feature(s) { 11 6 10 12 } with accuracy 89.0%
The accuracy is decreasing!
Current round feature(s): { 11 6 10 3 } with accuracy 91.0%,
lower than best 95.4%
Best feature subset is { 11 6 10 } with accuracy 95.4%
```

BE Traceback of Small Dataset

```
Welcome to Hugo Wan's Feature Selection Algorithm!
Type in the name of the file to test: CS205_small_Data__26.txt
Type the number of the algorithm you want to run.
1) Forward Selection
2) Backward Elimination
Your choice: 2
This dataset has 500 records and 12 features.
Initial accuracy with all features: 76.6%
Beginning search.
     Current feature(s) { 2 3 4 5 6 7 8 9 10 11 12 } with accuracy 75.2%
     Current feature(s) { 1 3 4 5 6 7 8 9 10 11 12 } with accuracy 75.6%
     Current feature(s) { 1 2 4 5 6 7 8 9 10 11 12 } with accuracy 77.4%
     Current feature(s) { 1 2 3 5 6 7 8 9 10 11 12 } with accuracy 78.0%
     Current feature(s) { 1 2 3 4 6 7 8 9 10 11 12 } with accuracy 77.8%
     Current feature(s) { 1 2 3 4 5 7 8 9 10 11 12 } with accuracy 74.6%
     Current feature(s) { 1 2 3 4 5 6 8 9 10 11 12 } with accuracy 75.4%
     Current feature(s) { 1 2 3 4 5 6 7 9 10 11 12 } with accuracy 78.4%
     Current feature(s) { 1 2 3 4 5 6 7 8 10 11 12 } with accuracy 74.0%
     Current feature(s) { 1 2 3 4 5 6 7 8 9 11 12 } with accuracy 74.2%
     Current feature(s) { 1 2 3 4 5 6 7 8 9 10 12 } with accuracy 71.0%
     Current feature(s) { 1 2 3 4 5 6 7 8 9 10 11 } with accuracy 77.0%
Current best overall is { 1 2 3 4 5 6 7 9 10 11 12 } with accuracy 78.4%
     Current feature(s) { 2 3 4 5 6 7 9 10 11 12 } with accuracy 78.0%
     Current feature(s) { 1 3 4 5 6 7 9 10 11 12 } with accuracy 77.2%
     Current feature(s) { 1 2 4 5 6 7 9 10 11 12 } with accuracy 76.2%
     Current feature(s) { 1 2 3 5 6 7 9 10 11 12 } with accuracy 77.8%
     Current feature(s) { 1 2 3 4 6 7 9 10 11 12 } with accuracy 79.0%
     Current feature(s) { 1 2 3 4 5 7 9 10 11 12 } with accuracy 74.8%
     Current feature(s) { 1 2 3 4 5 6 9 10 11 12 } with accuracy 77.0%
     Current feature(s) { 1 2 3 4 5 6 7 10 11 12 } with accuracy 75.4%
     Current feature(s) { 1 2 3 4 5 6 7 9 11 12 } with accuracy 77.8%
     Current feature(s) { 1 2 3 4 5 6 7 9 10 12 } with accuracy 73.4%
     Current feature(s) { 1 2 3 4 5 6 7 9 10 11 } with accuracy 78.8%
Current best overall is { 1 2 3 4 6 7 9 10 11 12 } with accuracy 79.0%
     Current feature(s) { 2 3 4 6 7 9 10 11 12 } with accuracy 80.2%
     Current feature(s) { 1 3 4 6 7 9 10 11 12 } with accuracy 77.2%
     Current feature(s) { 1 2 4 6 7 9 10 11 12 } with accuracy 79.2%
     Current feature(s) { 1 2 3 6 7 9 10 11 12 } with accuracy 77.4%
     Current feature(s) { 1 2 3 4 7 9 10 11 12 } with accuracy 76.4%
     Current feature(s) { 1 2 3 4 6 9 10 11 12 } with accuracy 77.0%
     Current feature(s) { 1 2 3 4 6 7 10 11 12 } with accuracy 77.6%
     Current feature(s) { 1 2 3 4 6 7 9 11 12 } with accuracy 76.6%
     Current feature(s) { 1 2 3 4 6 7 9 10 12 } with accuracy 72.4%
     Current feature(s) { 1 2 3 4 6 7 9 10 11 } with accuracy 79.4%
Current best overall is { 2 3 4 6 7 9 10 11 12 } with accuracy 80.2%
     Current feature(s) { 3 4 6 7 9 10 11 12 } with accuracy 76.4%
     Current feature(s) { 2 4 6 7 9 10 11 12 } with accuracy 78.0%
     Current feature(s) { 2 3 6 7 9 10 11 12 } with accuracy 78.4%
     Current feature(s) { 2 3 4 7 9 10 11 12 } with accuracy 74.4%
     Current feature(s) { 2 3 4 6 9 10 11 12 } with accuracy 79.6%
     Current feature(s) { 2 3 4 6 7 10 11 12 } with accuracy 78.4%
     Current feature(s) { 2 3 4 6 7 9 11 12 } with accuracy 76.6%
     Current feature(s) { 2 3 4 6 7 9 10 12 } with accuracy 70.8%
     Current feature(s) { 2 3 4 6 7 9 10 11 } with accuracy 80.0%
The accuracy is decreasing!
Current round feature(s): { 2 3 4 6 7 9 10 11 } with accuracy 80.0%,
lower than best 80.2%
Best feature subset is { 2 3 4 6 7 9 10 11 12 } with accuracy 80.2%
```

FS Traceback of Large Dataset

```
Welcome to Hugo Wan's Feature Selection Algorithm!
Type in the name of the file to test: CS205_large_Data__36.txt
Type the number of the algorithm you want to run.
1) Forward Selection
2) Backward Elimination
Your choice: 1
This dataset has 1000 records and 50 features.
Beginning search.
     Current feature(s) { 1 } with accuracy 65.7%
     Current feature(s) { 2 } with accuracy 67.4%
     Current feature(s) { 3 } with accuracy 68.8%
     Current feature(s) { 4 } with accuracy 68.1%
     Current feature(s) { 5 } with accuracy 68.4%
     Current feature(s) { 6 } with accuracy 67.0%
     Current feature(s) { 7 } with accuracy 65.7%
     Current feature(s) { 8 } with accuracy 68.1%
     Current feature(s) { 9 } with accuracy 85.5%
     Current feature(s) { 10 } with accuracy 67.3%
     Current feature(s) { 11 } with accuracy 65.5%
     Current feature(s) { 12 } with accuracy 67.4%
     Current feature(s) { 13 } with accuracy 65.9%
     Current feature(s) { 14 } with accuracy 66.1%
     Current feature(s) { 15 } with accuracy 68.0%
     Current feature(s) { 16 } with accuracy 65.5%
     Current feature(s) { 17 } with accuracy 65.5%
     Current feature(s) { 18 } with accuracy 66.3%
     Current feature(s) { 19 } with accuracy 68.1%
     Current feature(s) { 20 } with accuracy 70.0%
     Current feature(s) { 21 } with accuracy 69.1%
     Current feature(s) { 22 } with accuracy 67.1%
     Current feature(s) { 23 } with accuracy 65.9%
     Current feature(s) { 24 } with accuracy 67.7%
     Current feature(s) { 25 } with accuracy 69.0%
     Current feature(s) { 26 } with accuracy 67.1%
     Current feature(s) { 27 } with accuracy 69.7%
     Current feature(s) { 28 } with accuracy 70.7%
     Current feature(s) { 29 } with accuracy 66.4%
     Current feature(s) { 30 } with accuracy 66.3%
     Current feature(s) { 31 } with accuracy 68.0%
     Current feature(s) { 32 } with accuracy 67.6%
     Current feature(s) { 33 } with accuracy 65.6%
     Current feature(s) { 34 } with accuracy 64.7%
     Current feature(s) { 35 } with accuracy 67.0%
     Current feature(s) { 36 } with accuracy 66.1%
     Current feature(s) { 37 } with accuracy 65.8%
     Current feature(s) { 38 } with accuracy 66.0%
     Current feature(s) { 39 } with accuracy 66.4%
     Current feature(s) { 40 } with accuracy 69.4%
     Current feature(s) { 41 } with accuracy 65.8%
     Current feature(s) { 42 } with accuracy 66.3%
     Current feature(s) { 43 } with accuracy 68.3%
     Current feature(s) { 44 } with accuracy 69.3%
     Current feature(s) { 45 } with accuracy 70.8%
     Current feature(s) { 46 } with accuracy 68.9%
     Current feature(s) { 47 } with accuracy 67.6%
     Current feature(s) { 48 } with accuracy 72.1%
     Current feature(s) { 49 } with accuracy 68.5%
     Current feature(s) { 50 } with accuracy 67.7%
Current best overall is { 9 } with accuracy 85.5%
     Current feature(s) { 9 1 } with accuracy 81.9%
     Current feature(s) { 9 2 } with accuracy 84.4%
     Current feature(s) { 9 3 } with accuracy 83.9%
     Current feature(s) { 9 4 } with accuracy 83.7%
     Current feature(s) { 9 5 } with accuracy 82.5%
     Current feature(s) { 9 6 } with accuracy 83.5%
     Current feature(s) { 9 7 } with accuracy 83.6%
     Current feature(s) { 9 8 } with accuracy 82.4%
     Current feature(s) { 9 10 } with accuracy 83.6%
     Current feature(s) { 9 11 } with accuracy 82.8%
     Current feature(s) { 9 12 } with accuracy 84.0%
     Current feature(s) { 9 13 } with accuracy 82.7%
     Current feature(s) { 9 14 } with accuracy 82.7%
     Current feature(s) { 9 15 } with accuracy 84.2%
     Current feature(s) { 9 16 } with accuracy 82.6%
     Current feature(s) { 9 17 } with accuracy 84.2%
     Current feature(s) { 9 18 } with accuracy 83.8%
     Current feature(s) { 9 19 } with accuracy 81.2%
     Current feature(s) { 9 20 } with accuracy 85.5%
     Current feature(s) { 9 21 } with accuracy 82.7%
     Current feature(s) { 9 22 } with accuracy 85.2%
     Current feature(s) { 9 23 } with accuracy 84.6%
     Current feature(s) { 9 24 } with accuracy 84.8%
     Current feature(s) { 9 25 } with accuracy 83.4%
     Current feature(s) { 9 26 } with accuracy 83.8%
     Current feature(s) { 9 27 } with accuracy 84.6%
     Current feature(s) { 9 28 } with accuracy 81.1%
     Current feature(s) { 9 29 } with accuracy 82.4%
     Current feature(s) { 9 30 } with accuracy 86.3%
     Current feature(s) { 9 31 } with accuracy 84.4%
     Current feature(s) { 9 32 } with accuracy 85.0%
     Current feature(s) { 9 33 } with accuracy 84.1%
     Current feature(s) { 9 34 } with accuracy 82.6%
     Current feature(s) { 9 35 } with accuracy 84.4%
     Current feature(s) { 9 36 } with accuracy 84.5%
     Current feature(s) { 9 37 } with accuracy 84.1%
     Current feature(s) { 9 38 } with accuracy 84.8%
     Current feature(s) { 9 39 } with accuracy 86.3%
     Current feature(s) { 9 40 } with accuracy 83.9%
     Current feature(s) { 9 41 } with accuracy 84.6%
     Current feature(s) { 9 42 } with accuracy 83.2%
     Current feature(s) { 9 43 } with accuracy 84.7%
     Current feature(s) { 9 44 } with accuracy 85.3%
     Current feature(s) { 9 45 } with accuracy 83.9%
     Current feature(s) { 9 46 } with accuracy 84.6%
     Current feature(s) { 9 47 } with accuracy 84.0%
     Current feature(s) { 9 48 } with accuracy 97.1%
     Current feature(s) { 9 49 } with accuracy 84.0%
```

```
     Current feature(s) { 9 50 } with accuracy 83.2%
Current best overall is { 9 48 } with accuracy 97.1%
     Current feature(s) { 9 48 1 } with accuracy 92.4%
     Current feature(s) { 9 48 2 } with accuracy 92.7%
     Current feature(s) { 9 48 3 } with accuracy 93.2%
     Current feature(s) { 9 48 4 } with accuracy 91.7%
     Current feature(s) { 9 48 5 } with accuracy 91.2%
     Current feature(s) { 9 48 6 } with accuracy 93.3%
     Current feature(s) { 9 48 7 } with accuracy 93.6%
     Current feature(s) { 9 48 8 } with accuracy 92.9%
     Current feature(s) { 9 48 10 } with accuracy 93.3%
     Current feature(s) { 9 48 11 } with accuracy 91.6%
     Current feature(s) { 9 48 12 } with accuracy 91.4%
     Current feature(s) { 9 48 13 } with accuracy 92.5%
     Current feature(s) { 9 48 14 } with accuracy 93.2%
     Current feature(s) { 9 48 15 } with accuracy 93.2%
     Current feature(s) { 9 48 16 } with accuracy 92.1%
     Current feature(s) { 9 48 17 } with accuracy 92.4%
     Current feature(s) { 9 48 18 } with accuracy 93.3%
     Current feature(s) { 9 48 19 } with accuracy 92.4%
     Current feature(s) { 9 48 20 } with accuracy 91.1%
     Current feature(s) { 9 48 21 } with accuracy 92.4%
     Current feature(s) { 9 48 22 } with accuracy 91.5%
     Current feature(s) { 9 48 23 } with accuracy 93.4%
     Current feature(s) { 9 48 24 } with accuracy 92.3%
     Current feature(s) { 9 48 25 } with accuracy 92.6%
     Current feature(s) { 9 48 26 } with accuracy 92.5%
     Current feature(s) { 9 48 27 } with accuracy 91.4%
     Current feature(s) { 9 48 28 } with accuracy 92.7%
     Current feature(s) { 9 48 29 } with accuracy 91.7%
     Current feature(s) { 9 48 30 } with accuracy 92.8%
     Current feature(s) { 9 48 31 } with accuracy 93.3%
     Current feature(s) { 9 48 32 } with accuracy 95.3%
     Current feature(s) { 9 48 33 } with accuracy 94.2%
     Current feature(s) { 9 48 34 } with accuracy 92.8%
     Current feature(s) { 9 48 35 } with accuracy 93.7%
     Current feature(s) { 9 48 36 } with accuracy 93.2%
     Current feature(s) { 9 48 37 } with accuracy 92.2%
     Current feature(s) { 9 48 38 } with accuracy 91.3%
     Current feature(s) { 9 48 39 } with accuracy 93.0%
     Current feature(s) { 9 48 40 } with accuracy 93.1%
     Current feature(s) { 9 48 41 } with accuracy 92.2%
     Current feature(s) { 9 48 42 } with accuracy 91.0%
     Current feature(s) { 9 48 43 } with accuracy 92.9%
     Current feature(s) { 9 48 44 } with accuracy 93.6%
     Current feature(s) { 9 48 45 } with accuracy 92.0%
     Current feature(s) { 9 48 46 } with accuracy 91.7%
     Current feature(s) { 9 48 47 } with accuracy 92.9%
     Current feature(s) { 9 48 49 } with accuracy 91.9%
     Current feature(s) { 9 48 50 } with accuracy 93.4%
The accuracy is decreasing!
Current round feature(s): { 9 48 32 } with accuracy 95.3%,
lower than best 97.1%
Best feature subset is { 9 48 } with accuracy 97.1%
```

BE Traceback of Large Dataset

```
Welcome to Hugo Wan's Feature Selection Algorithm!
Type in the name of the file to test: CS205_large_Data__36.txt
Type the number of the algorithm you want to run.
1) Forward Selection
2) Backward Elimination
Your choice: 2
This dataset has 1000 records and 50 features.
Initial accuracy with all features: 68.1%
Beginning search.
     Current feature(s) { ... // all but feature 1 } with accuracy 66.6%
     Current feature(s) { ... // all but feature 2 } with accuracy 67.0%
     Current feature(s) { ... // all but feature 3 } with accuracy 67.1%.

     ... // listed all till feature 50, shortened for report readability

Current best overall is { ... // all but feature 15 } with accuracy 69.3%

     ... // similar patterns as above until the first accuracy decrease

The accuracy is decreasing!
Current round feature(s): { 1 2 3 4 5 6 7 8 9 10 11 12 13 14 18 19 20 21
  23 24 25 26 27 28 29 31 33 34 35 36 37 38 39 40 41 42 43 44 46 48 49 50 }
  with accuracy 73.3%, lower than best 73.5%
Best feature subset is { 1 2 3 4 5 6 7 8 9 10 11 12 13 14 18 19 20 21
  23 24 25 26 27 28 29 31 33 34 35 36 37 38 39 40 41 42 43 44 45 46 48 49 50 }
  with accuracy 73.5%
```

4 Part 2 - Real World Datasets

The real world dataset I chose for Part 2 is:

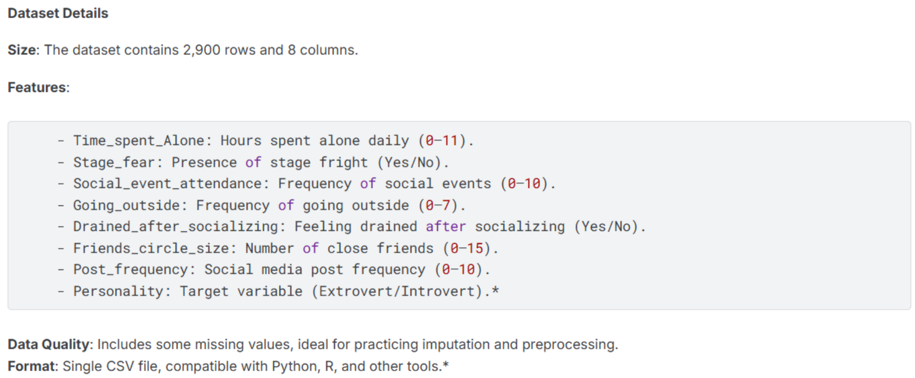
- Extrovert vs. Introvert Behavior Data
- Link: https://www.kaggle.com/datasets/rakeshkapilavai/extrovert-vs-introvert-behavior-data?resource=download

The Reformation of Real-World Dataset

The original real-world dataset contains 2,900 rows and 8 columns (features). To make it compatible with the format used in the assigned small and large datasets, I made a few adjustments. First, I moved the last column — `Personality`, which labels each row as either Extrovert or Introvert — to the first position and encoded the values as **1.0 for Extrovert** and **2.0 for Introvert**. This effectively turned the label into the "class" column.

Next, to match the numeric formatting used in the provided datasets, I converted the Yes/No values in the `Stage_fear` and `Drained_after_socializing` features to 1.0 and 2.0, respectively. This standardized the data and brought the total number of features down to 7, just like the format used in Part 1.

Although the dataset initially had 2,900 rows, some rows contained missing values in one or more columns. To ensure clean input and better accuracy during analysis, I filtered out those incomplete rows. After this step, I was left with 2,477 fully populated rows, which I saved in a cleaned version named `cleaned_data.txt`.



Dataset Details

Size: The dataset contains 2,900 rows and 8 columns.

Features:

```
- Time_spent_Alone: Hours spent alone daily (0-11).
- Stage_fear: Presence of stage fright (Yes/No).
- Social_event_attendance: Frequency of social events (0-10).
- Going_outside: Frequency of going outside (0-7).
- Drained_after_socializing: Feeling drained after socializing (Yes/No).
- Friends_circle_size: Number of close friends (0-15).
- Post_frequency: Social media post frequency (0-10).
- Personality: Target variable (Extrovert/Introvert).*
```

Data Quality: Includes some missing values, ideal for practicing imputation and preprocessing.
Format: Single CSV file, compatible with Python, R, and other tools.*

Figure 5: **Real-World Dataset Details**: 2,900 rows and 8 columns (features).

Important Note

Although this real-world dataset only contains 7 features, making it relatively small, I still included a backward elimination analysis for completeness—even though it wasn't strictly required for Part 2. I felt it was helpful to show how the algorithm behaves on a large real-world dataset with limited features.

I also chose to keep the full feature trace in the backward plot for this dataset. This is different from the plot for my assigned large dataset, where I trimmed the feature trace for readability, since it had too many features. Including the full trace here helps emphasize the simplicity and clarity of this dataset's structure.

Report - Real World Dataset

For the **real-world dataset**, the best forward selection subset was **2**, which alone reached the highest accuracy of 92.5%. This shows that Feature 2 is highly predictive of personality type. Adding a second feature didn't improve the result, suggesting the added feature might be redundant.

In contrast, **backward elimination** found the best subset to be **2, 4, 5, 7**, also reaching **92.5%**. This subset preserved the key signal from Feature 2 while trimming away less helpful features. Accuracy slightly dipped after removing one more feature in the last step, confirming that the selected subset maintained important traits.

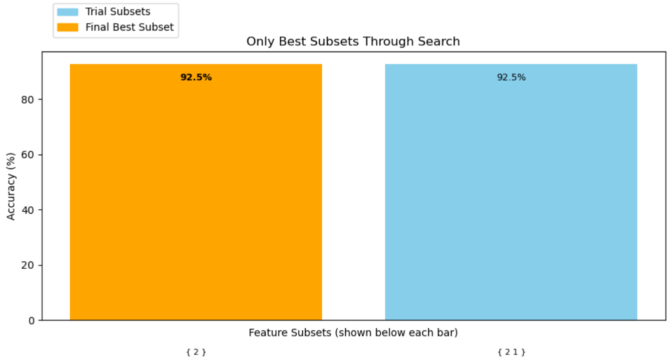


Figure 6: **Real-World Dataset - Forward Selection**: the best subset was {2}, which achieved the highest accuracy of **92.5%**.

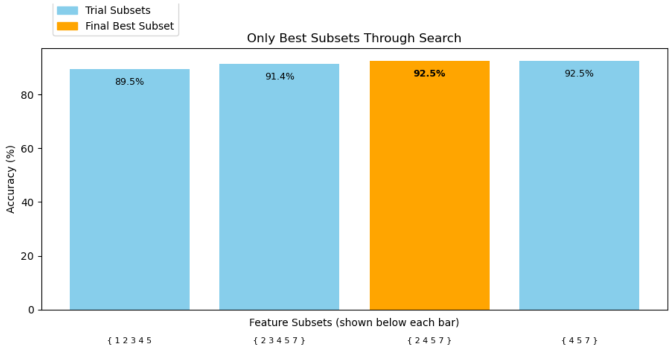


Figure 7: **Real-World Dataset - Backward Elimination**: the best subset is {2, 4, 5, 7}, achieving the same accuracy of **92.5%**.

Both search methods performed equally well, reinforcing that a small subset of features can be powerful. As noted earlier, this dataset only has 7 usable features and required reformatting. The full search and graphing processes for **both directions finished in under one minute**.

FS Traceback of Real-World Dataset

```
Welcome to Hugo Wan's Feature Selection Algorithm!
Type in the name of the file to test: cleaned_data.txt
Type the number of the algorithm you want to run.
1) Forward Selection
2) Backward Elimination
Your choice: 1
This dataset has 2477 records and 7 features.
Beginning search.
     Current feature(s) { 1 } with accuracy 87.0%
     Current feature(s) { 2 } with accuracy 92.5%
     Current feature(s) { 3 } with accuracy 91.9%
     Current feature(s) { 4 } with accuracy 86.7%
     Current feature(s) { 5 } with accuracy 92.5%
     Current feature(s) { 6 } with accuracy 89.2%
     Current feature(s) { 7 } with accuracy 86.4%
Current best overall is { 2 } with accuracy 92.5%
     Current feature(s) { 2 1 } with accuracy 92.5%
     Current feature(s) { 2 3 } with accuracy 92.5%
     Current feature(s) { 2 4 } with accuracy 92.5%
     Current feature(s) { 2 5 } with accuracy 92.5%
     Current feature(s) { 2 6 } with accuracy 92.5%
     Current feature(s) { 2 7 } with accuracy 92.5%
The accuracy is decreasing!
Current round feature(s): { 2 1 } with accuracy 92.5%,
lower than best 92.5%
Best feature subset is { 2 } with accuracy 92.5%
```

BE Traceback of Real-World Dataset

```
Welcome to Hugo Wan's Feature Selection Algorithm!
Type in the name of the file to test: cleaned_data.txt
Type the number of the algorithm you want to run.
1) Forward Selection
2) Backward Elimination
Your choice: 2
This dataset has 2477 records and 7 features.
Initial accuracy with all features: 88.5%
Beginning search.
     Current feature(s) { 2 3 4 5 6 7 } with accuracy 89.3%
     Current feature(s) { 1 3 4 5 6 7 } with accuracy 88.5%
     Current feature(s) { 1 2 4 5 6 7 } with accuracy 89.2%
     Current feature(s) { 1 2 3 5 6 7 } with accuracy 88.5%
     Current feature(s) { 1 2 3 4 6 7 } with accuracy 88.5%
     Current feature(s) { 1 2 3 4 5 7 } with accuracy 89.5%
     Current feature(s) { 1 2 3 4 5 6 } with accuracy 88.3%
Current best overall is { 1 2 3 4 5 7 } with accuracy 89.5%
     Current feature(s) { 2 3 4 5 7 } with accuracy 91.4%
     Current feature(s) { 1 3 4 5 7 } with accuracy 89.5%
     Current feature(s) { 1 2 4 5 7 } with accuracy 91.2%
     Current feature(s) { 1 2 3 5 7 } with accuracy 90.3%
     Current feature(s) { 1 2 3 4 7 } with accuracy 89.5%
     Current feature(s) { 1 2 3 4 5 } with accuracy 90.8%
Current best overall is { 2 3 4 5 7 } with accuracy 91.4%
     Current feature(s) { 3 4 5 7 } with accuracy 91.4%
     Current feature(s) { 2 4 5 7 } with accuracy 92.5%
     Current feature(s) { 2 3 5 7 } with accuracy 92.5%
     Current feature(s) { 2 3 4 7 } with accuracy 91.4%
     Current feature(s) { 2 3 4 5 } with accuracy 92.5%
Current best overall is { 2 4 5 7 } with accuracy 92.5%
     Current feature(s) { 4 5 7 } with accuracy 92.5%
     Current feature(s) { 2 5 7 } with accuracy 92.5%
     Current feature(s) { 2 4 7 } with accuracy 92.5%
     Current feature(s) { 2 4 5 } with accuracy 92.5%
The accuracy is decreasing!
Current round feature(s): { 4 5 7 } with accuracy 92.5%,
lower than best 92.5%
Best feature subset is { 2 4 5 7 } with accuracy 92.5%
```

5 Citations / References

1. Keogh, Eamonn. “Project2_Spring_2025.” Eamonn Keogh, 2 May 2025.

2. Keogh, Eamonn. “Project_2_sample_report.” Eamonn Keogh, 2 May 2025.

3. Keogh, Eamonn. “Project_2_Briefing.” Eamonn Keogh, 2 May 2025.

4. Keogh, Eamonn. “7__MachineLearning002.” Eamonn Keogh, Mar. 2025.

5. Keogh, Eamonn. “9__MachineLearning004.” Eamonn Keogh, 10 May 2025.

6. scikit, learn. “1.13. Feature Selection.” Scikit, 2025, scikit-learn.org/stable/modules/feature_selection.html.

7. scikit, learn. “1.6. Nearest Neighbors.” Scikit, 2025, scikit-learn.org/stable/modules/neighbors.html.

8. Lupi, Nicolas. “Feature Selection with Optuna.” Towards Data Science, 9 May 2024, towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e.

9. Wikipedia. “Feature Selection.” Wikipedia, Wikimedia Foundation, 24 May 2025, en.wikipedia.org/wiki/Feature_selection.

10. Wikipedia. “K-Nearest Neighbors Algorithm.” Wikipedia, Wikimedia Foundation, 16 Apr. 2025, en.wikipedia.org/wiki/K-nearest_neighbors_algorithm.

11. pandas. “Pandas Documentation#.” Pandas Documentation - Pandas 2.2.3 Documentation, 20 Sept. 2024, pandas.pydata.org/docs/.

12. NumPy. “NumPy Documentation.” NumPy Documentation, 2022, numpy.org/doc/.

13. scikit, learn. “6.3. Preprocessing Data.” Scikit, 2025, scikit-learn.org/stable/modules/preprocessing.html.

14. Matplotlib, et al. “Examples#.” Examples - Matplotlib 3.10.3 Documentation, 2025, matplotlib.org/stable/gallery/index.html.

15. Kapilavayi, Rakesh. “Extrovert vs. Introvert Behavior Data.” Kaggle, 21 May 2025, www.kaggle.com/datasets/rakeshkapilavai/extrovert-vs-introvert-behavior-data?resource=download.

16. Solomon, Brad. “Python Plotting with Matplotlib (Guide).” Real Python, Real Python, 1 Dec. 2023, realpython.com/python-matplotlib-guide/.