Forest Cover Type Analysis

Christen Ye & Mariana Chen & Anastasia Ivanova

April 6, 2023

Agenda

- Dataset
- Classification Tree
- Random Forest
- Naïve Bayes Classifier
- K-Nearest Neighbors Classifier
- Feature Importance
- Next Steps

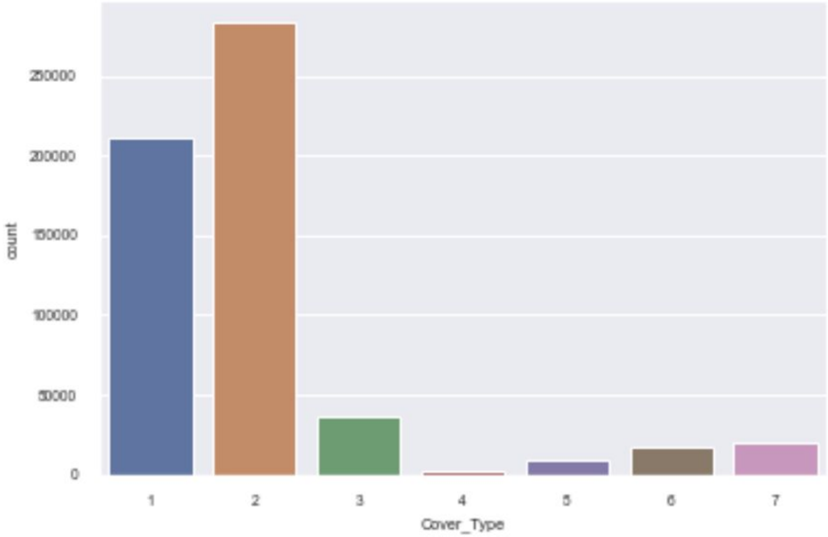
Dataset

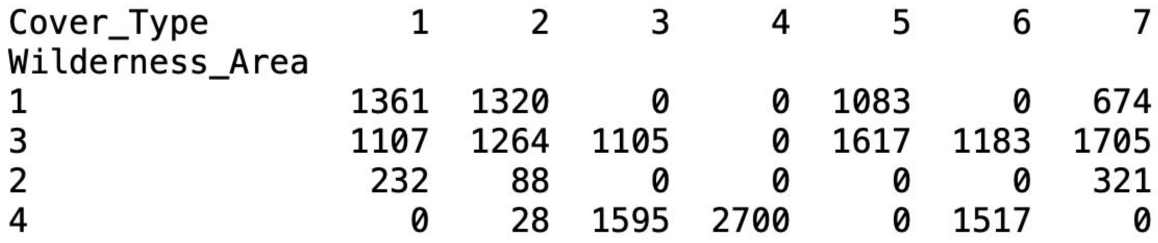
Initial Dataset

- 581,012 records
- 54 features
- Response - cover type

Transformed Dataset

- 2,700 records from each cover type
- Removed some of the soil types and wilderness areas
- Transformed aspect
- = 18,900 records and 20 features





```
Cover_Type          1     2     3     4     5     6     7
Wilderness_Area
1                1361  1320     0     0  1083     0   674
3                1107  1264  1105     0  1617  1183  1705
2                 232    88     0     0     0     0   321
4                   0    28  1595  2700     0  1517     0
```

Dataset

Elevation	Slope	HDTH	VDTH	HDTR	H9	HN	H3	HDTF	TA
2952	30	67	38	2614	238	169	41	2213	0.309017
3134	6	90	0	750	204	234	169	1140	0.777146
3292	19	175	7	4226	230	196	90	3588	0.515038

WA3	WA4	S3	S4	S10	S23	S29	S30	S32	CT
1	0	0	0	0	0	0	0	0	1
0	0	0	0	0	1	0	0	0	1
1	0	0	0	0	1	0	0	0	1

Table 1. Snapshot of dataset

Classification Tree – All features

CV	fit_time	accuracy
1	0.2	78.2%
2	0.2	79.0%
3	0.2	77.8%
4	0.2	78.4%
5	0.2	76.9%
Average	0.2	78.1%

Table 2. 5-Fold Cross-Validation Performance

- The average of 5-fold cross-validation accuracy is 78.1%.
- The overall out-of-sample accuracy is 76.5%.

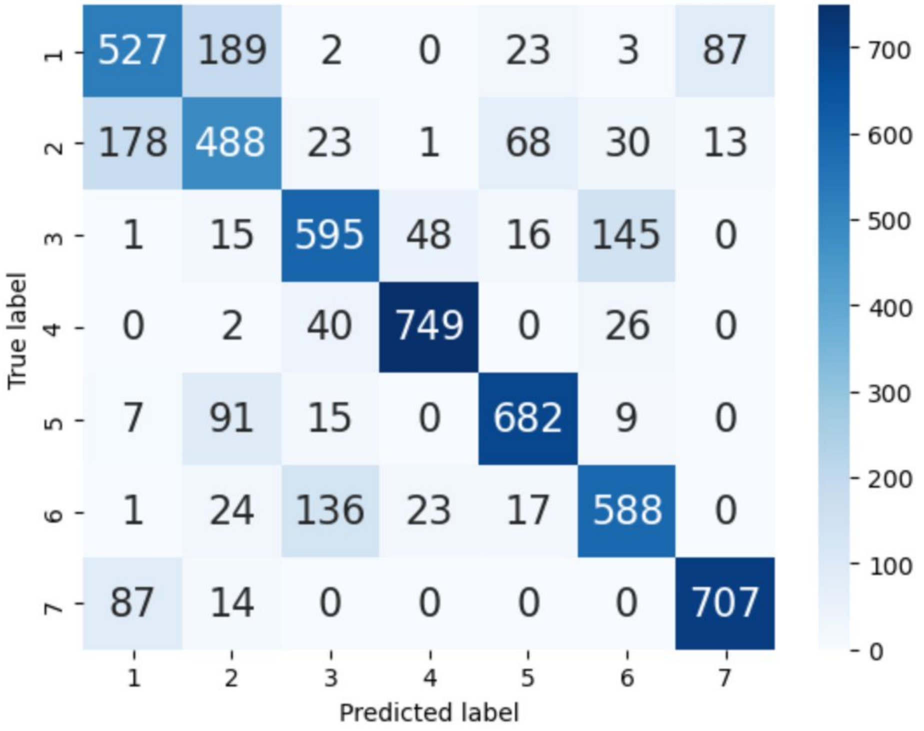
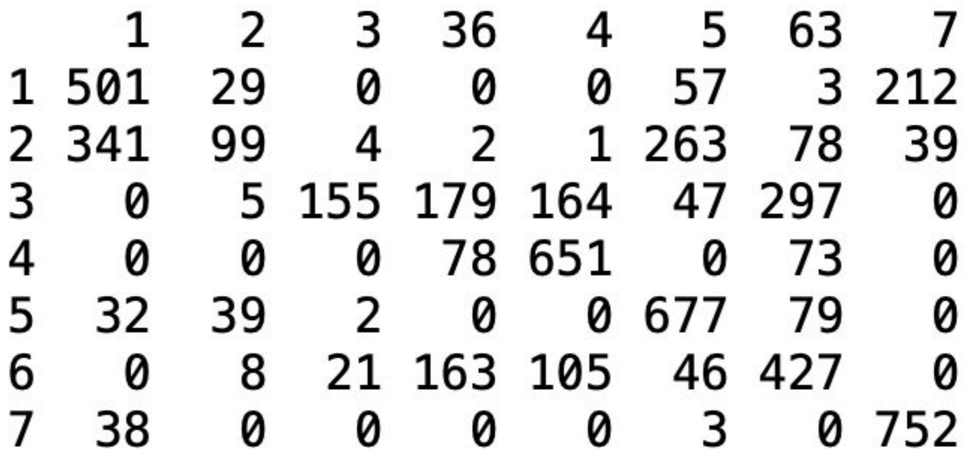


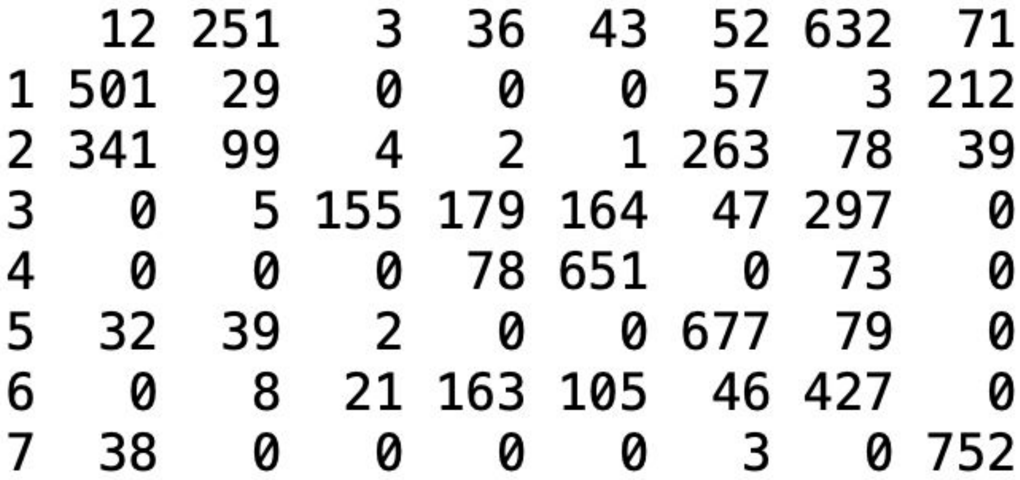
Figure 1. Out-of-sample Confusion Matrix

Classification Tree – All features



```
    1   2   3  36   4   5  63   7
1 501  29   0   0   0  57   3 212
2 341  99   4   2   1 263  78  39
3   0   5 155 179 164  47 297   0
4   0   0   0  78 651   0  73   0
5  32  39   2   0   0 677  79   0
6   0   8  21 163 105  46 427   0
7  38   0   0   0   0   3   0 752
```

Figure 2. Classification Table based on 50% PIs



```
   12 251   3  36  43  52 632  71
1 501  29   0   0   0  57   3 212
2 341  99   4   2   1 263  78  39
3   0   5 155 179 164  47 297   0
4   0   0   0  78 651   0  73   0
5  32  39   2   0   0 677  79   0
6   0   8  21 163 105  46 427   0
7  38   0   0   0   0   3   0 752
```

Figure 3. Classification Table based on 80% PIs

- Based on 50% PIs, the accuracy is 68.8%. Cover type 3 and 6 are harder to discriminate.
- Based on 80% PIs, the accuracy is 88.7%. Cover type 1, 2 and 3,6 are harder to discriminate.

Random Forest – All features

CV	fit_time	accuracy
1	2.9	85.7%
2	3.1	86.1%
3	3.0	85.3%
4	2.9	86.5%
5	2.8	85.4%
Average	2.9	85.8%

Table 3. 5-Fold Cross-Validation Performance

- The average of 5-fold cross-validation accuracy is 85.8%.
- The overall out-of-sample accuracy is 85.3%.

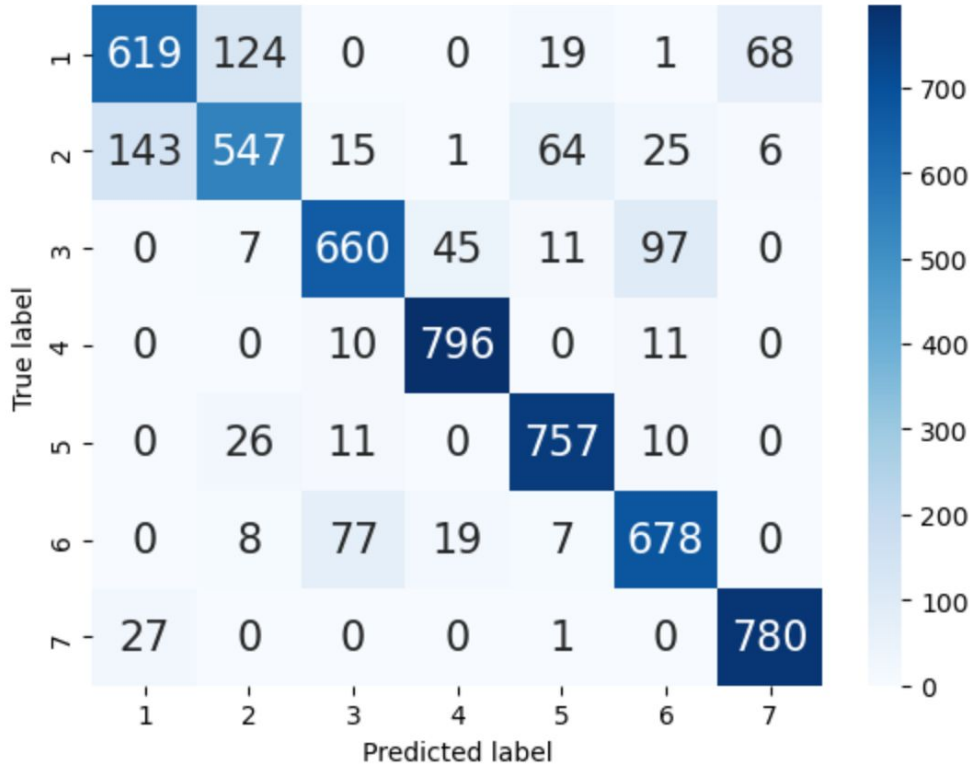
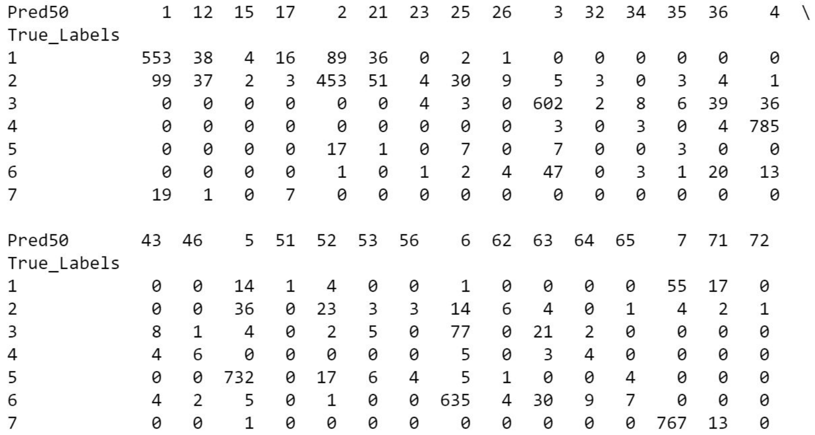


Figure 4. Out-of-sample Confusion Matrix

Random Forest – All features

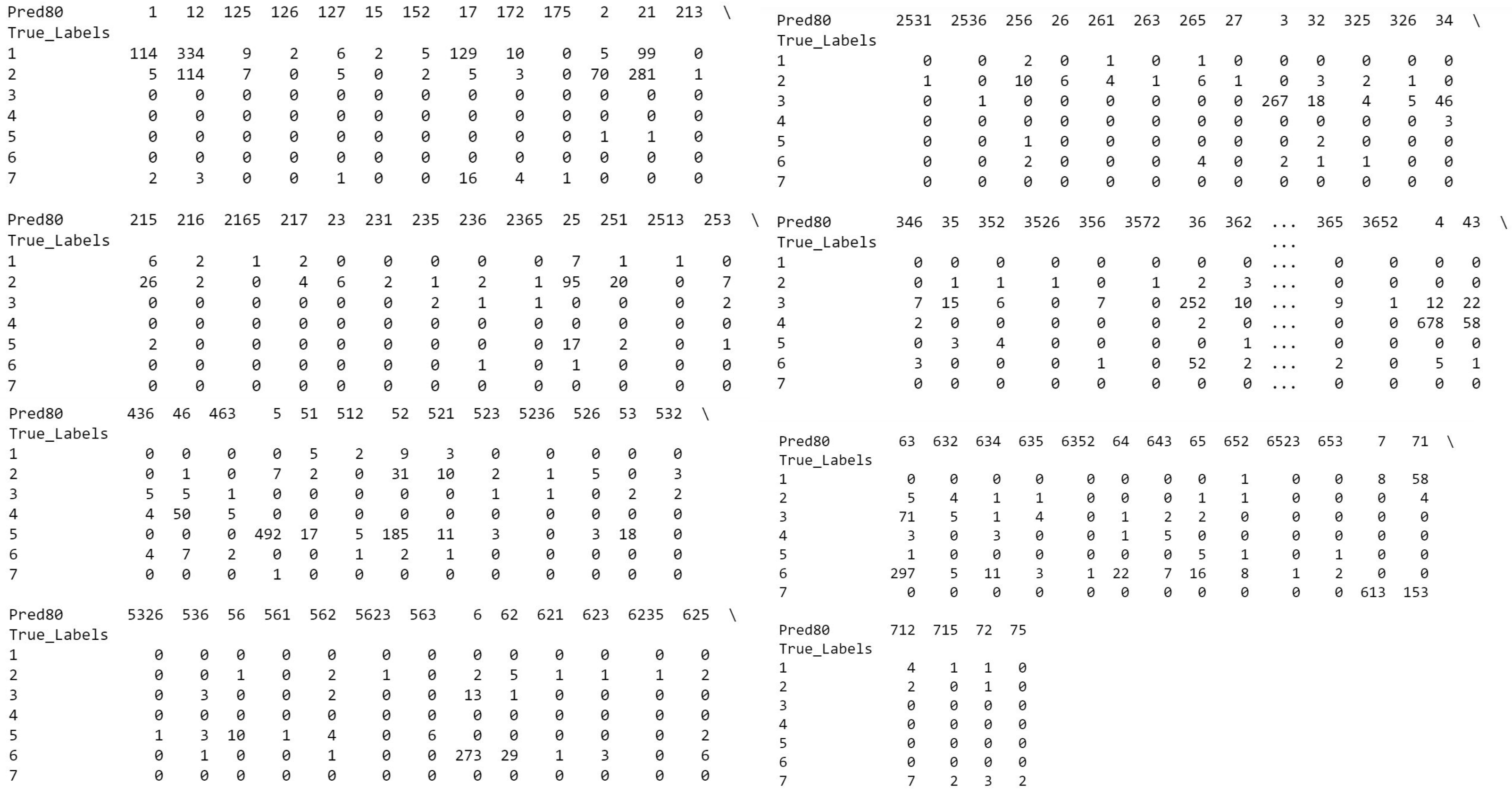


```
Pred50         1   12   15   17    2   21   23   25   26    3   32   34   35   36    4  \
True_Labels
1            553   38    4   16   89   36    0    2    1    0    0    0    0    0    0
2             99   37    2    3  453   51    4   30    9    5    3    0    3    4    1
3              0    0    0    0    0    0    4    3    0  602    2    8    6   39   36
4              0    0    0    0    0    0    0    0    0    3    0    3    0    4  785
5              0    0    0    0   17    1    0    7    0    7    0    0    3    0    0
6              0    0    0    0    1    0    1    2    4   47    0    3    1   20   13
7             19    1    0    7    0    0    0    0    0    0    0    0    0    0    0

Pred50        43   46    5   51   52   53   56    6   62   63   64   65    7   71   72
True_Labels
1              0    0   14    1    4    0    0    1    0    0    0    0   55   17    0
2              0    0   36    0   23    3    3   14    6    4    0    1    4    2    1
3              8    1    4    0    2    5    0   77    0   21    2    0    0    0    0
4              4    6    0    0    0    0    0    5    0    3    4    0    0    0    0
5              0    0  732    0   17    6    4    5    1    0    0    4    0    0    0
6              4    2    5    0    1    0    0  635    4   30    9    7    0    0    0
7              0    0    1    0    0    0    0    0    0    0    0    0  767   13    0
```

Figure 5. Classification Table based on 50% PIs

- Based on 50% PIs, the accuracy is 89.1%. Cover type 1, 2 and 3 are harder to discriminate.
- Based on 80% PIs, the accuracy is 91.7%. Cover type 1, 2 and 3 are harder to discriminate.



```
Pred80         1   12  125  126  127   15  152   17  172  175    2   21  213  \
True_Labels
1            114  334    9    2    6    2    5  129   10    0    5   99    0
2              5  114    7    0    5    0    2    5    3    0   70  281    1
3              0    0    0    0    0    0    0    0    0    0    0    0    0
4              0    0    0    0    0    0    0    0    0    0    0    0    0
5              0    0    0    0    0    0    0    0    0    0    1    1    0
6              0    0    0    0    0    0    0    0    0    0    0    0    0
7              2    3    0    0    1    0    0   16    4    1    0    0    0

Pred80       215  216  2165  217  23  231  235  236  2365  25  251  2513  253  \
True_Labels
1              6    2     1    2   0    0    0    0     0   7    1     1    0
2             26    2     0    4   6    2    1    2     1  95   20     0    7
3              0    0     0    0   0    0    2    1     1   0    0     0    2
4              0    0     0    0   0    0    0    0     0   0    0     0    0
5              2    0     0    0   0    0    0    0     0  17    2     0    1
6              0    0     0    0   0    0    0    1     0   1    0     0    0
7              0    0     0    0   0    0    0    0     0   0    0     0    0

Pred80       436  46  463    5  51  512   52  521  523  5236  526  53  532  \
True_Labels
1              0   0    0    0   5    2    9    3    0     0    0   0    0
2              0   1    0    7   2    0   31   10    2     1    5   0    3
3              5   5    1    0   0    0    0    0    1     1    0   2    2
4              4  50    5    0   0    0    0    0    0     0    0   0    0
5              0   0    0  492  17    5  185   11    3     0    3  18    0
6              4   7    2    0   0    1    2    1    0     0    0   0    0
7              0   0    0    1   0    0    0    0    0     0    0   0    0

Pred80       5326  536  56  561  562  5623  563    6  62  621  623  6235  625  \
True_Labels
1               0    0   0    0    0     0    0    0   0    0    0     0    0
2               0    0   1    0    2     1    0    2   5    1    1     1    2
3               0    3   0    0    2     0    0   13   1    0    0     0    0
4               0    0   0    0    0     0    0    0   0    0    0     0    0
5               1    3  10    1    4     0    6    0   0    0    0     0    2
6               0    1   0    0    1     0    0  273  29    1    3     0    6
7               0    0   0    0    0     0    0    0   0    0    0     0    0

Pred80       2531  2536  256  26  261  263  265  27    3  32  325  326  34  \
True_Labels
1               0     0    2   0    1    0    1   0    0   0    0    0   0
2               1     0   10   6    4    1    6   1    0   3    2    1   0
3               0     1    0   0    0    0    0   0  267  18    4    5  46
4               0     0    0   0    0    0    0   0    0   0    0    0   3
5               0     0    1   0    0    0    0   0    0   2    0    0   0
6               0     0    2   0    0    0    4   0    2   1    1    0   0
7               0     0    0   0    0    0    0   0    0   0    0    0   0

Pred80       346  35  352  3526  356  3572   36  362  ...  365  3652    4  43  \
True_Labels                                          ...
1              0   0    0     0    0     0    0    0  ...    0     0    0   0
2              0   1    1     1    0     1    2    3  ...    0     0    0   0
3              7  15    6     0    7     0  252   10  ...    9     1   12  22
4              2   0    0     0    0     0    2    0  ...    0     0  678  58
5              0   3    4     0    0     0    0    1  ...    0     0    0   0
6              3   0    0     0    1     0   52    2  ...    2     0    5   1
7              0   0    0     0    0     0    0    0  ...    0     0    0   0

Pred80        63  632  634  635  6352  64  643  65  652  6523  653    7   71  \
True_Labels
1              0    0    0    0     0   0    0   0    1     0    0    8   58
2              5    4    1    1     0   0    0   1    1     0    0    0    4
3             71    5    1    4     0   1    2   2    0     0    0    0    0
4              3    0    3    0     0   1    5   0    0     0    0    0    0
5              1    0    0    0     0   0    0   5    1     0    1    0    0
6            297    5   11    3     1  22    7  16    8     1    2    0    0
7              0    0    0    0     0   0    0   0    0     0    0  613  153

Pred80       712  715  72  75
True_Labels
1              4    1   1   0
2              2    0   1   0
3              0    0   0   0
4              0    0   0   0
5              0    0   0   0
6              0    0   0   0
7              7    2   3   2
```

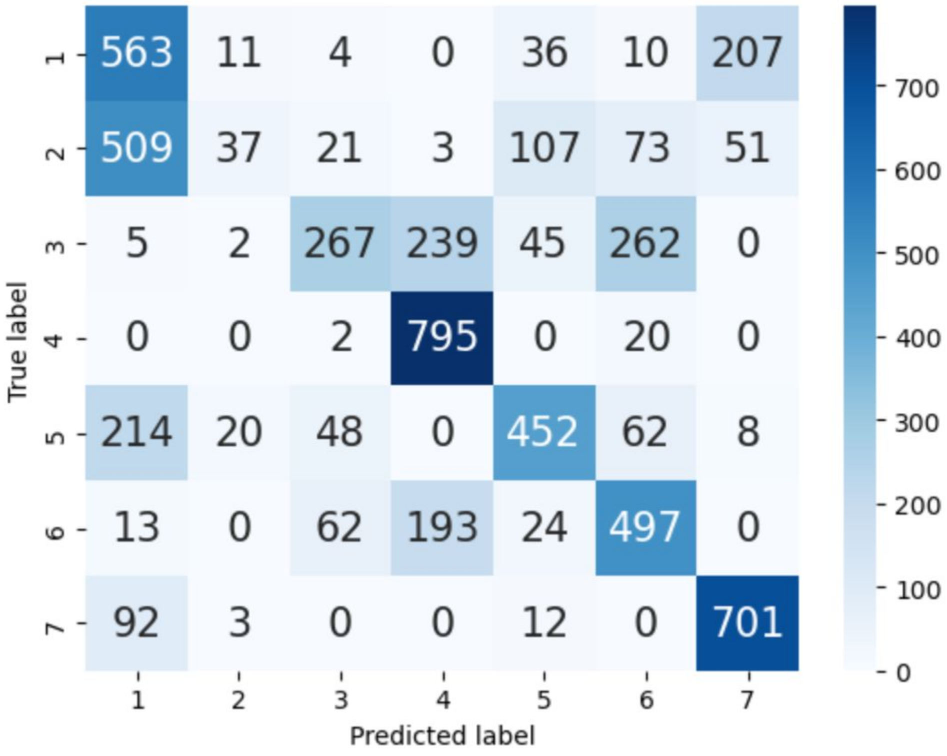
Figure 6. Classification Table based on 80% PIs

Naïve Bayes Classifier – All features

CV	fit_time	accuracy
1	0.01	56.2%
2	0.01	58.7%
3	0.01	57.6%
4	0.01	57.9%
5	0.02	58.1%
Average	0.01	57.7%

Table 4. 5-Fold Cross-Validation Performance

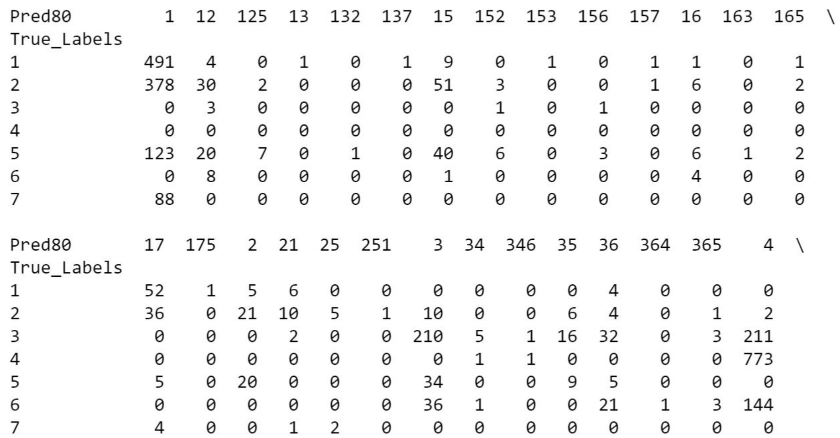
- The average of 5-fold cross-validation accuracy is 57.7%.
- The overall out-of-sample accuracy is 58.4%.



True label \ Predicted label	1	2	3	4	5	6	7
1	563	11	4	0	36	10	207
2	509	37	21	3	107	73	51
3	5	2	267	239	45	262	0
4	0	0	2	795	0	20	0
5	214	20	48	0	452	62	8
6	13	0	62	193	24	497	0
7	92	3	0	0	12	0	701

Figure 7. Out-of-sample Confusion Matrix

Naïve Bayes Classifier – All features

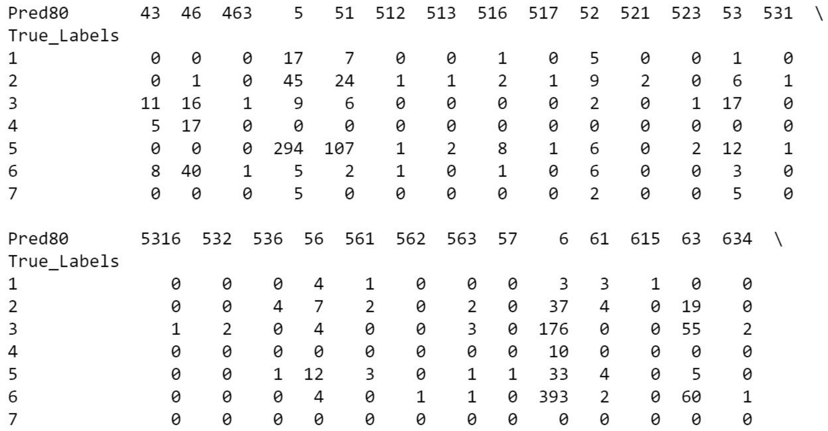


```
Pred80          1  12  125  13  132  137  15  152  153  156  157  16  163  165  \
True_Labels
1             491   4    0   1    0    1   9    0    1    0    1   1    0    1
2             378  30    2   0    0    0  51    3    0    0    1   6    0    2
3               0   3    0   0    0    0   0    1    0    1    0   0    0    0
4               0   0    0   0    0    0   0    0    0    0    0   0    0    0
5             123  20    7   0    1    0  40    6    0    3    0   6    1    2
6               0   8    0   0    0    0   1    0    0    0    0   4    0    0
7              88   0    0   0    0    0   0    0    0    0    0   0    0    0

Pred80        17  175   2  21  25  251    3  34  346  35  36  364  365    4  \
True_Labels
1             52    1   5   6   0    0    0   0    0   0   4    0    0    0
2             36    0  21  10   5    1   10   0    0   6   4    0    1    2
3              0    0   0   2   0    0  210   5    1  16  32    0    3  211
4              0    0   0   0   0    0    0   1    1   0   0    0    0  773
5              5    0  20   0   0    0   34   0    0   9   5    0    0    0
6              0    0   0   0   0    0   36   1    0   0  21    1    3  144
7              4    0   0   1   2    0    0   0    0   0   0    0    0    0
```

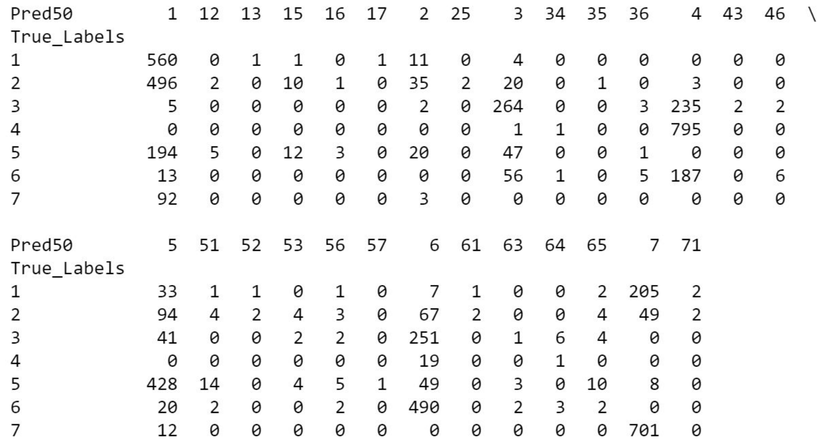
Figure 8. Classification Table based on 50% PIs

- Based on 50% PIs, the accuracy is 68.8%. Cover type 2 is harder to discriminate.
- Based on 80% PIs, the accuracy is 91.7%. Cover types 2 is harder to discriminate.



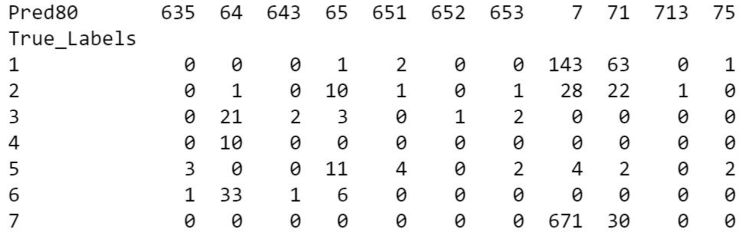
```
Pred80        43  46  463    5   51  512  513  516  517  52  521  523  53  531  \
True_Labels
1              0   0    0   17    7    0    0    1    0   5    0    0   1    0
2              0   1    0   45   24    1    1    2    1   9    2    0   6    1
3             11  16    1    9    6    0    0    0    0   2    0    1  17    0
4              5  17    0    0    0    0    0    0    0   0    0    0   0    0
5              0   0    0  294  107    1    2    8    1   6    0    2  12    1
6              8  40    1    5    2    1    0    1    0   6    0    0   3    0
7              0   0    0    5    0    0    0    0    0   2    0    0   5    0

Pred80        5316  532  536  56  561  562  563  57    6  61  615  63  634  \
True_Labels
1                0    0    0   4    1    0    0   0    3   3    1   0    0
2                0    0    4   7    2    0    2   0   37   4    0  19    0
3                1    2    0   4    0    0    3   0  176   0    0  55    2
4                0    0    0   0    0    0    0   0   10   0    0   0    0
5                0    0    1  12    3    0    1   1   33   4    0   5    0
6                0    0    0   4    0    1    1   0  393   2    0  60    1
7                0    0    0   0    0    0    0   0    0   0    0   0    0
```



```
Pred50          1  12  13  15  16  17   2  25    3  34  35  36    4  43  46  \
True_Labels
1             560   0   1   1   0   1  11   0    4   0   0   0    0   0   0
2             496   2   0  10   1   0  35   2   20   0   1   0    3   0   0
3               5   0   0   0   0   0   2   0  264   0   0   3  235   2   2
4               0   0   0   0   0   0   0   0    1   1   0   0  795   0   0
5             194   5   0  12   3   0  20   0   47   0   0   1    0   0   0
6              13   0   0   0   0   0   0   0   56   1   0   5  187   0   6
7              92   0   0   0   0   0   3   0    0   0   0   0    0   0   0

Pred50          5  51  52  53  56  57    6  61  63  64  65    7  71
True_Labels
1              33   1   1   0   1   0    7   1   0   0   2  205   2
2              94   4   2   4   3   0   67   2   0   0   4   49   2
3              41   0   0   2   2   0  251   0   1   6   4    0   0
4               0   0   0   0   0   0   19   0   0   1   0    0   0
5             428  14   0   4   5   1   49   0   3   0  10    8   0
6              20   2   0   0   2   0  490   0   2   3   2    0   0
7              12   0   0   0   0   0    0   0   0   0   0  701   0
```



```
Pred80        635  64  643  65  651  652  653    7  71  713  75
True_Labels
1               0   0    0   1    2    0    0  143  63    0   1
2               0   1    0  10    1    0    1   28  22    1   0
3               0  21    2   3    0    1    2    0   0    0   0
4               0  10    0   0    0    0    0    0   0    0   0
5               3   0    0  11    4    0    2    4   2    0   2
6               1  33    1   6    0    0    0    0   0    0   0
7               0   0    0   0    0    0    0  671  30    0   0
```

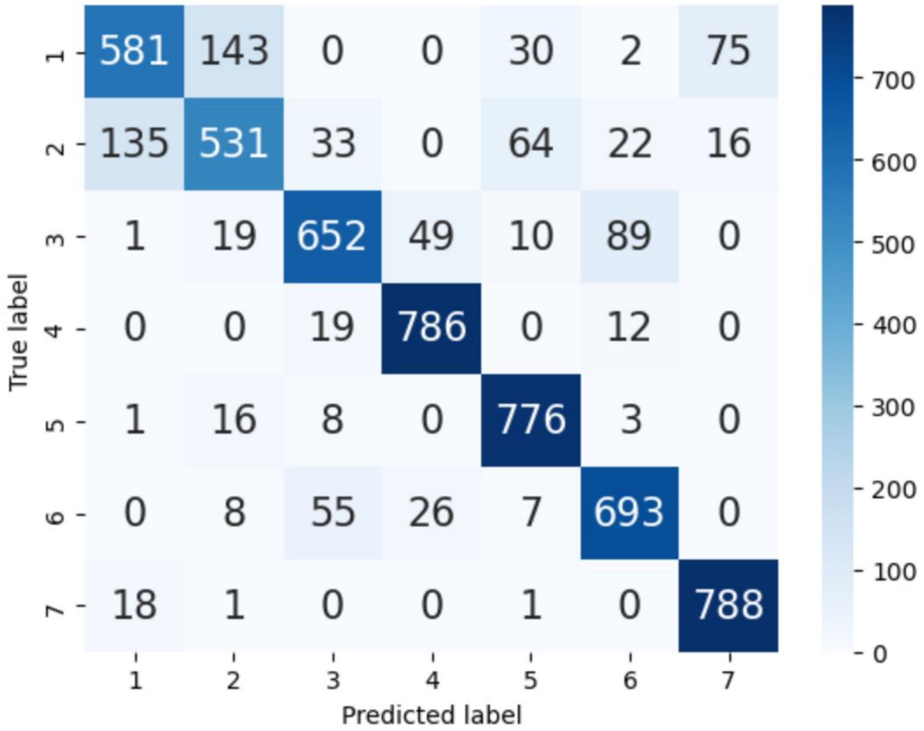
Figure 9. Classification Table based on 80% PIs

KNN Classifier – All features

CV	fit_time	accuracy
1	0.01	85.8%
2	0.008	85.9%
3	0.006	85.4%
4	0.006	85.0%
5	0.006	86.0%
Average	0.008	85.6%

Table 5. 5-Fold Cross-Validation Performance

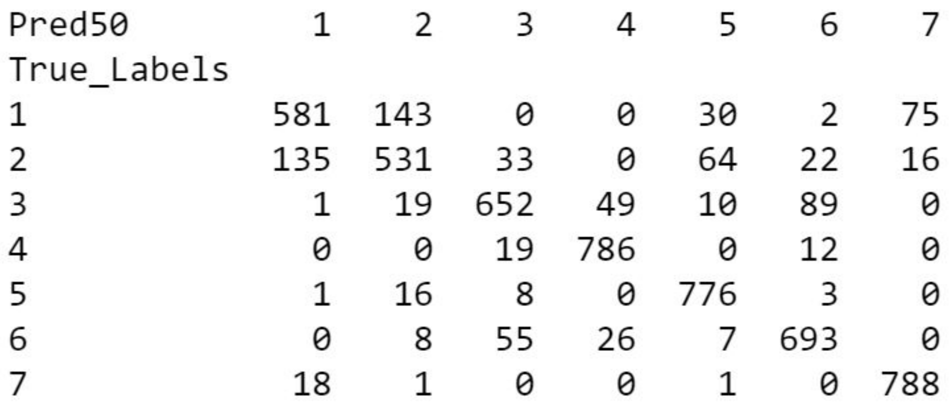
- The average of 5-fold cross-validation accuracy is 85.6%.
- The overall out-of-sample accuracy is 84.8%.



True label \ Predicted label	1	2	3	4	5	6	7
1	581	143	0	0	30	2	75
2	135	531	33	0	64	22	16
3	1	19	652	49	10	89	0
4	0	0	19	786	0	12	0
5	1	16	8	0	776	3	0
6	0	8	55	26	7	693	0
7	18	1	0	0	1	0	788

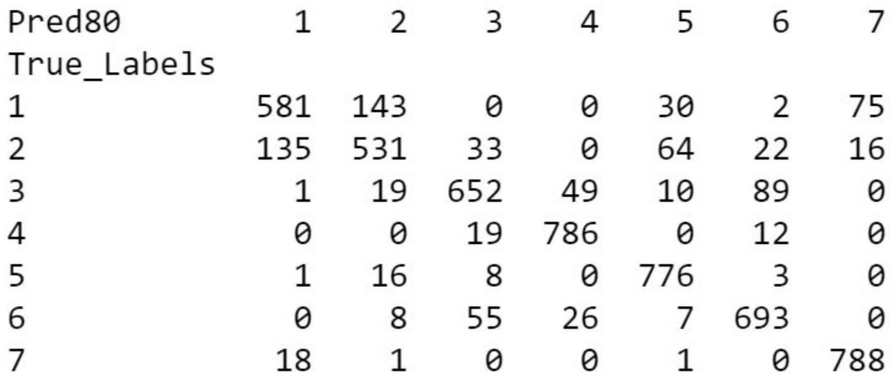
Figure 10. Out-of-sample Confusion Matrix

KNN Classifier – All features



```
Pred50          1    2    3    4    5    6    7
True_Labels
1             581  143    0    0   30    2   75
2             135  531   33    0   64   22   16
3               1   19  652   49   10   89    0
4               0    0   19  786    0   12    0
5               1   16    8    0  776    3    0
6               0    8   55   26    7  693    0
7              18    1    0    0    1    0  788
```

Figure 11. Classification Table based on 50% PIs



```
Pred80          1    2    3    4    5    6    7
True_Labels
1             581  143    0    0   30    2   75
2             135  531   33    0   64   22   16
3               1   19  652   49   10   89    0
4               0    0   19  786    0   12    0
5               1   16    8    0  776    3    0
6               0    8   55   26    7  693    0
7              18    1    0    0    1    0  788
```

Figure 12. Classification Table based on 80% PIs

- Based on 50% PIs, the accuracy is 84.8%. Cover type 1 and 2 are harder to discriminate.
- Based on 80% Pis, the accuracy is 84.8%. Cover types 1 and 2 are harder to discriminate.

Feature Importance

Variable	Elevation	HDTR	HDTF	HDTH	VDTH	H9	WA4	TA
Random forest importance	0.3	0.1	0.08	0.07	0.06	0.05	0.05	0.05

Variable	H3	HN	Aspect	Slope	S10	WA3	S3
Random forest importance	0.04	0.04	0.04	0.03	0.02	0.02	0.02

Table 6. Random Forest Feature Importance

Next Steps

- Perform models with a subset of features, based on feature importance
- Confirm the best model (Random Forest) with larger sample sizes (the original dataset)
- Investigate the reason why 50% and 80% PI gave the same output for some models

Appendix

Variable	Explanation	Variable	Explanation
Elevation	Elevation in meters	W3	Wilderness Area 3
Slope	Slope in degrees	W4	Wilderness Area 4
HDTH	Horz Dist to nearest surface water features	S3	Soil type 3
VDTH	Vert Dist to nearest surface water features	S4	Soil type 4
HDTR	Horz Dist to nearest roadway	S10	Soil type 10
H9	Hillshade index at 9am, summer solstice	S23	Soil type 23
HN	Hillshade index at noon, summer solstice	S29	Soil type 29
H3	Hillshade index at 3pm, summer solstice	S30	Soil type 30
HDTF	Horz Dist to nearest wildfire ignition points	S32	Soil type 32
TA	Transformed Aspect (cosine(radians(aspect)))	CT	Cover Type

Table 2. Explanations of features