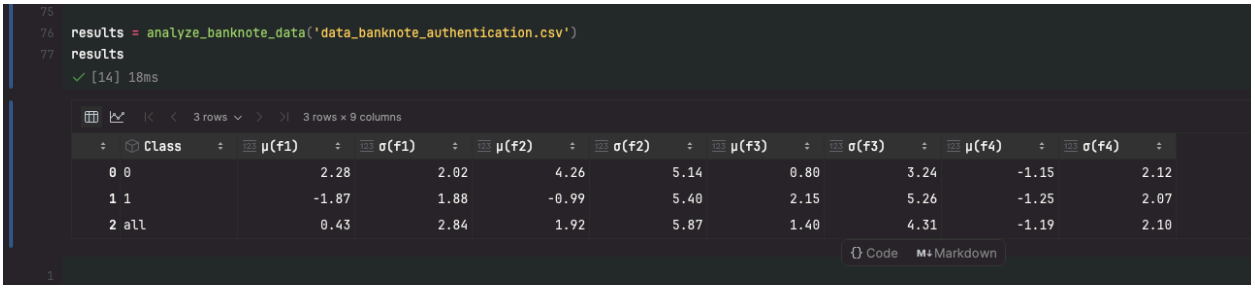
Ananya Singh

Assignment #3

1) Solution:



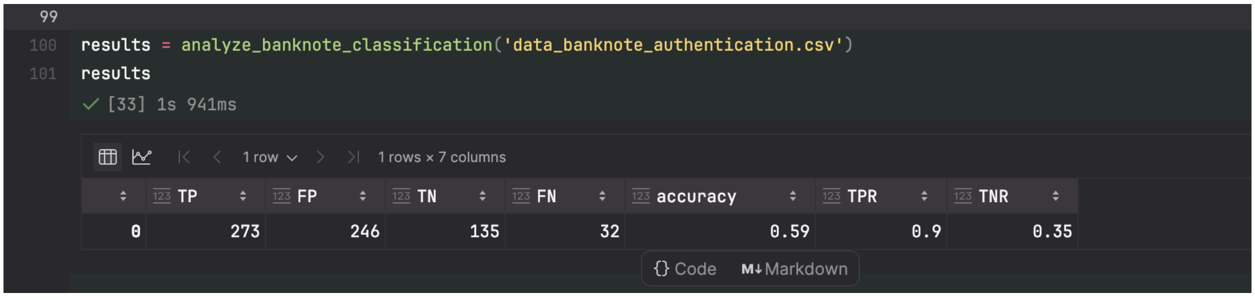
```
results = analyze_banknote_data('data_banknote_authentication.csv')
results
✓ [14] 18ms
```

	Class	μ(f1)	σ(f1)	μ(f2)	σ(f2)	μ(f3)	σ(f3)	μ(f4)	σ(f4)
0	0	2.28	2.02	4.26	5.14	0.80	3.24	-1.15	2.12
1	1	-1.87	1.88	-0.99	5.40	2.15	5.26	-1.25	2.07
2	all	0.43	2.84	1.92	5.87	1.40	4.31	-1.19	2.10

Analysis:

With analyzing the data above, you can see that f1 and f2 show the most obvious pattern which can be used to separate real and fake banknotes effectively. f4 is consistent but there isn't much variation. f3 has some separability which can make it somewhat useful.

2) Solution:



```
results = analyze_banknote_classification('data_banknote_authentication.csv')
results
✓ [33] 1s 941ms
```

	TP	FP	TN	FN	accuracy	TPR	TNR
0	273	246	135	32	0.59	0.9	0.35

Part 6:

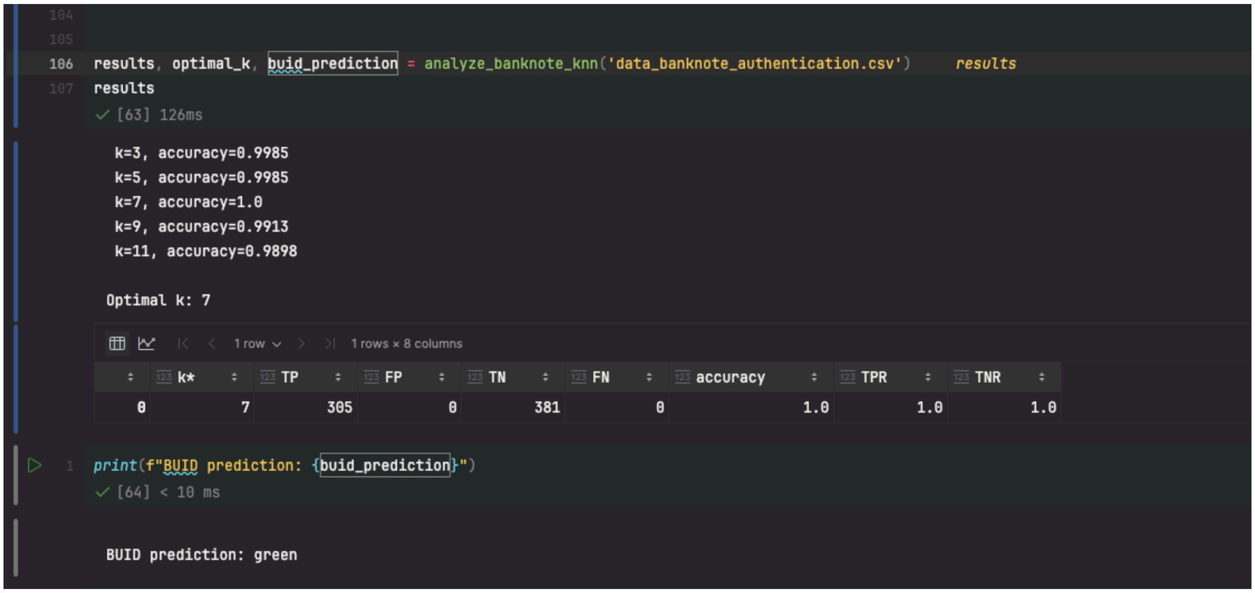
Does your simple classifier give you higher accuracy on identifying ”fake” bills or ”real” bills”?

Answer: The simple classifier is much better at identifying fake bills (TPR = 90%) than identifying actual bills (TNR = 35%).

Is your accuracy better than 50% (”coin” flipping)?

Answer: Yes, overall accuracy is at 59% which is way better than 50%.

3) Solution (answers to part 1, 2, 3, and 5):



```
results, optimal_k, buid_prediction = analyze_banknote_knn('data_banknote_authentication.csv')     results
results
✓ [63] 126ms
```

```
  k=3, accuracy=0.9985
  k=5, accuracy=0.9985
  k=7, accuracy=1.0
  k=9, accuracy=0.9913
  k=11, accuracy=0.9898

 Optimal k: 7
```

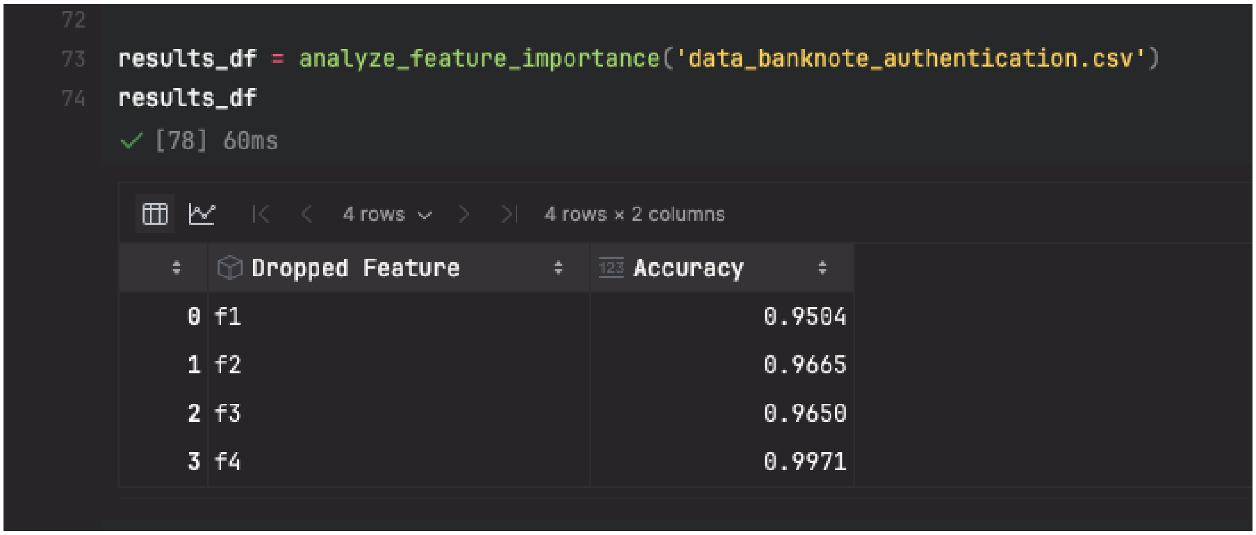
	k*	TP	FP	TN	FN	accuracy	TPR	TNR
0	7	305	0	381	0	1.0	1.0	1.0

```
print(f"BUID prediction: {buid_prediction}")
✓ [64] < 10 ms
```

```
 BUID prediction: green
```

Question: is your k-NN classifier better than your simple classifier for any of the measures from the previous table?’
Answer: Yes evidently. It correctly identifies all real and fake banknotes from the test set whereas the simple classifier had more than 246 incorrectly identified.

4) Solution:



```
results_df = analyze_feature_importance('data_banknote_authentication.csv')
results_df
✓ [78] 60ms
```

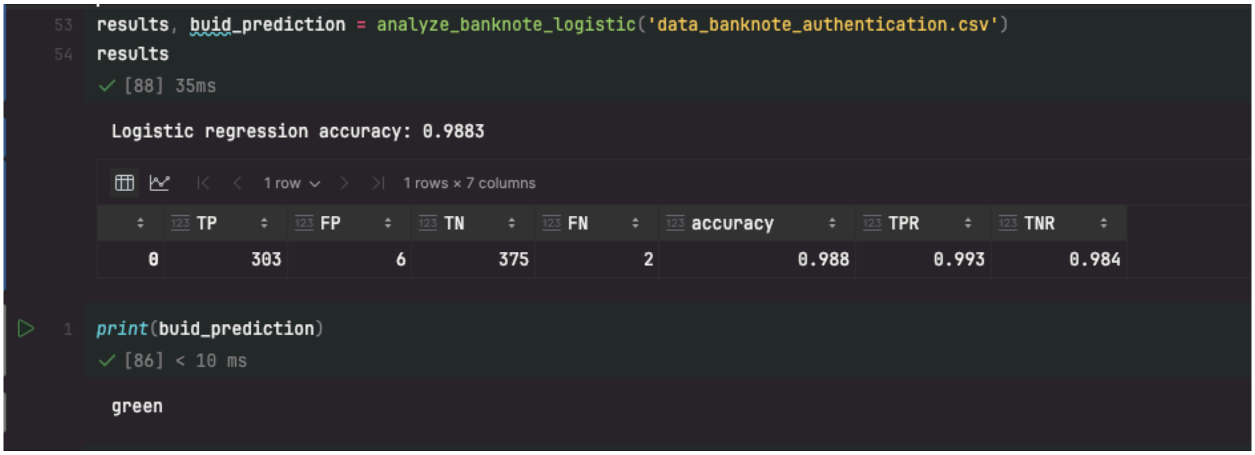
	Dropped Feature	Accuracy
0	f1	0.9504
1	f2	0.9665
2	f3	0.9650
3	f4	0.9971

Question: did accuracy increase in any of the 4 cases compared with accuracy when all 4 features are used?
Answer: Technically no but that's because we achieved 100% accuracy. This, however, means that keeping all features results in perfect accuracy.

Question: which feature, when removed, contributed the most to loss of accuracy?
Answer: Removing f1 caused the accuracy to drop to about 95% which is the lowest which means f1 is the most important feature.

Question: which feature, when removed, contributed the least to loss of accuracy?
Answer: Removing f4 caused the accuracy to drop to 99% which is the smallest accuracy drop. This means f4 is the least important feature.

5) Solution:



```
results, buid_prediction = analyze_banknote_logistic('data_banknote_authentication.csv')
results
✓ [88] 35ms
```

```
Logistic regression accuracy: 0.9883
```

	TP	FP	TN	FN	accuracy	TPR	TNR
0	303	6	375	2	0.988	0.993	0.984

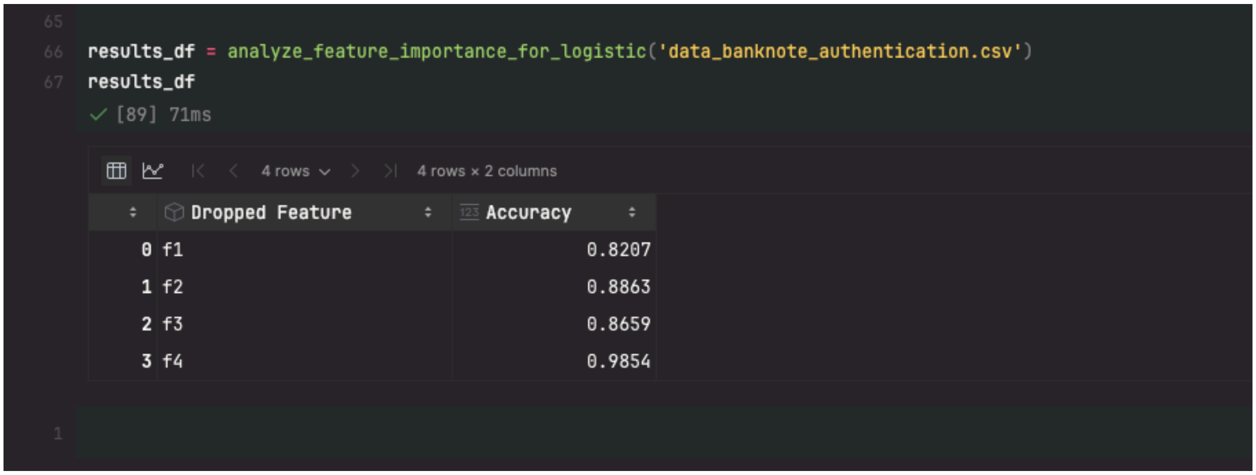
```
print(buid_prediction)
✓ [86] < 10 ms
```

```
green
```

Question: is your logistic regression better than your simple classifier for any of the measures from the previous table?
Answer: Comparing both tables, logistic regression is evidently better than simple classifier for pretty much all measures. It outperforms the simple classifier (98.4% vs. 35%) in correctly identifying real bills. Accuracy, TPR and TNR are significantly better.

Question: is your logistic regression better than your k-NN classifier (using the best k∗) for any of the measures from the previous table?
Answer: kNN had perfect scores throughout with 100% accuracy with the best k*. Therefore automatically, it is better than logistic regression.

6) Solution:



```
results_df = analyze_feature_importance_for_logistic('data_banknote_authentication.csv')
results_df
```

	Dropped Feature	Accuracy
0	f1	0.8207
1	f2	0.8863
2	f3	0.8659
3	f4	0.9854

Question: did accuracy increase in any of the 4 cases compared with accuracy when all 4 features are used?

Answer: No since accuracy with all features us higher than any accuracy when a feature is dropped.

Question: which feature, when removed, contributed the most to loss of accuracy?

Answer: The feature when removed caused the biggest accuracy drop is f1 which has the lowest accuracy out of all.

Question: which feature, when removed, contributed the least to loss of accuracy?

Answer: f4 contributed the least to the loss of accuracy with about ~0.30% loss from all features accuracy.

Question: is relative significance of features the same as you obtained using k-NN?

Answer: After comparing k-NN and logistic regression significance of features tables that I formulated, both tables/analysis agree that the most important feature is f1 and f4 is the least significant feature. So yes there is relative significance of features.