# Post Genomics – Fall 2025
# Homework 6
# Due: Thursday, October 16, 2025

**Part 1. Manipulating Data. (Programming - Python)** *(50 points)*

1.1 Using the Normal and Tumor CSV files from Homework 3. Subset the two CSV files with only the columns, ["chrom", "left", "ref_seq", "alt_seq", "Patient_ID", 'VCF_ID"]

1.1.1 How many unique normal patients do we have?

4 normal patients

1.1.2 How many unique tumor patients do we have?

5 tumor patients

1.1.3 Group by variant info, chrom, left, ref_seq, and alt_seq, let the other columns turn into list.

Done. Check output of the code. (Also Available at the end of this document)

1.1.4 Create a new column with the number of patients per variant on both the normal and tumor (name the column, N# and T#, respectively).

Done. Check output of the code. (Also Available at the end of this document)

1.1.5 Rename the columns, Patient_ID and VCF_ID, to have, _Normal or _Tumor, added depending which file you are working with.

Done. Check output of the code. (Also Available at the end of this document)

1.2 Using the output from part A, merge (how = outer) the Normal and Tumor together based on the columns [chrom, left, ref_seq, alt_seq] into a single CSV file named AML.

1.2.1 How many unique normal variants?

Zero (0).

1.2.2 How many unique tumor variants?

1408

1.2.3 How many variants are shared between normal and tumor (common)?

165

1.3 Using the Normal and Tumor files from Homework 3, concatenate these files along the axis = 0, with this Expand/Explode the rows based on the CSQ columns and save this file as AML_Expand.csv. Remove duplicate rows.

Saved and uploaded.

1.3.1 How many rows are in this file?

10,234

1.3.2 Create two new CSVs:

1. Subset of expianded with only the columns, ["SYMBOL", "Gene", "Feature"], name this AML_gene.csv.

   (Completed)

2. Subset of expanded CSV with only the columns, ["chrom", "left", "right", "ref_seq", "alt_seq", "Feature", "cDNA_position", "BIOTYPE"], name this AML_tx.csv.

   (Completed)

**Part 2 (Random Forest)** *(25 points)*

The Iris Dataset is a useful example set for machine learning classification problems. Work through the tutorial (https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/), and answer the questions below:

2.1 What was the accuracy of the model you built?

100% accuracy

2.2 What order were the important features ranked?

1. Petal length

2. Petal width

3. sepal length

4. sepal width

2.3 Change 2 of the parameters and repeat the model generation. What 2 parameters did you choose, what effects did they have on the model, and why do you think that was the case?

I’ve chosen test_size = 0.45 and n_stimators = 10. Accuracy decreased to 98.53%, the top of most relevant characteristics also changed, now being width the most relevant feature, also confusion matrix, the numbers in the diagonal squares are larger compared to previous and the value [3,2] of the matrix changed to one. All of these happened because of the reduction of the training set, since there was lesser data to practice, the model was not so optimized, that’s why there was a misidentification.

Images of both outputs at the end.

**Part 3 (K-Means Clustering)** *(25 points)*

Work through the tutorial (https://scikitlearn.org/stable/auto_examples/cluster/plot_cluster_iris.html#sphx-glr-auto-examples-cluster-plot-cluster-iris-py), and answer the questions below:

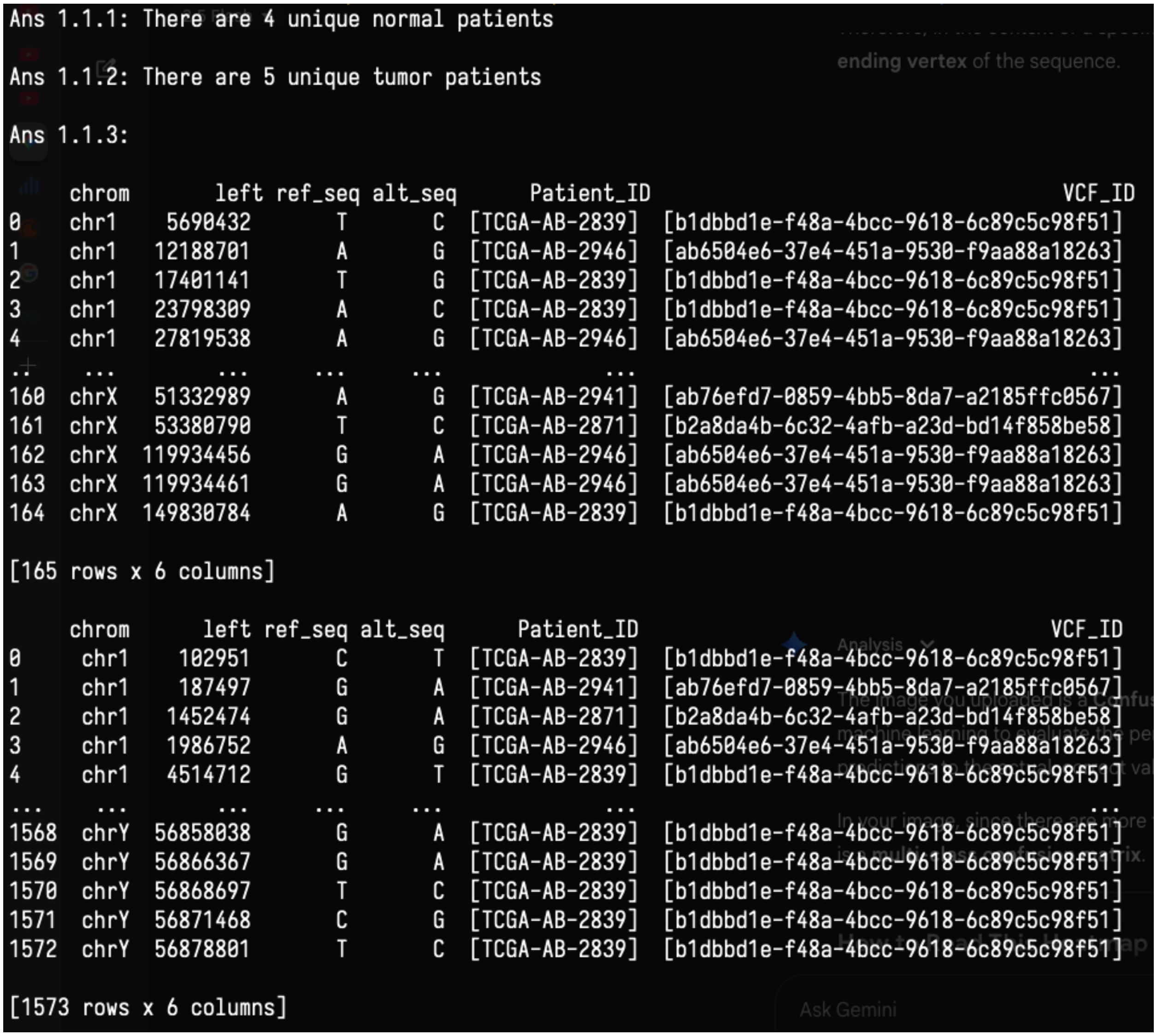
3.1 What are some conclusions you can draw about the clustering analysis?

The original clustering results really nail home how crucial choosing the right k is. Setting k=3, which matches the Iris flowers' actual species count, gives you a result that's nearly identical to the Ground Truth. The distinct Setosa group is always perfectly isolated, but the remaining clusters show K-Means' flaws: k=8 massively shreds the data into too many pieces, and trying k=3 with a bad random start messes up the separation of the two overlapping species, proving that good initialization isn't just nice to have it's essential to avoid getting stuck in a bad solution.

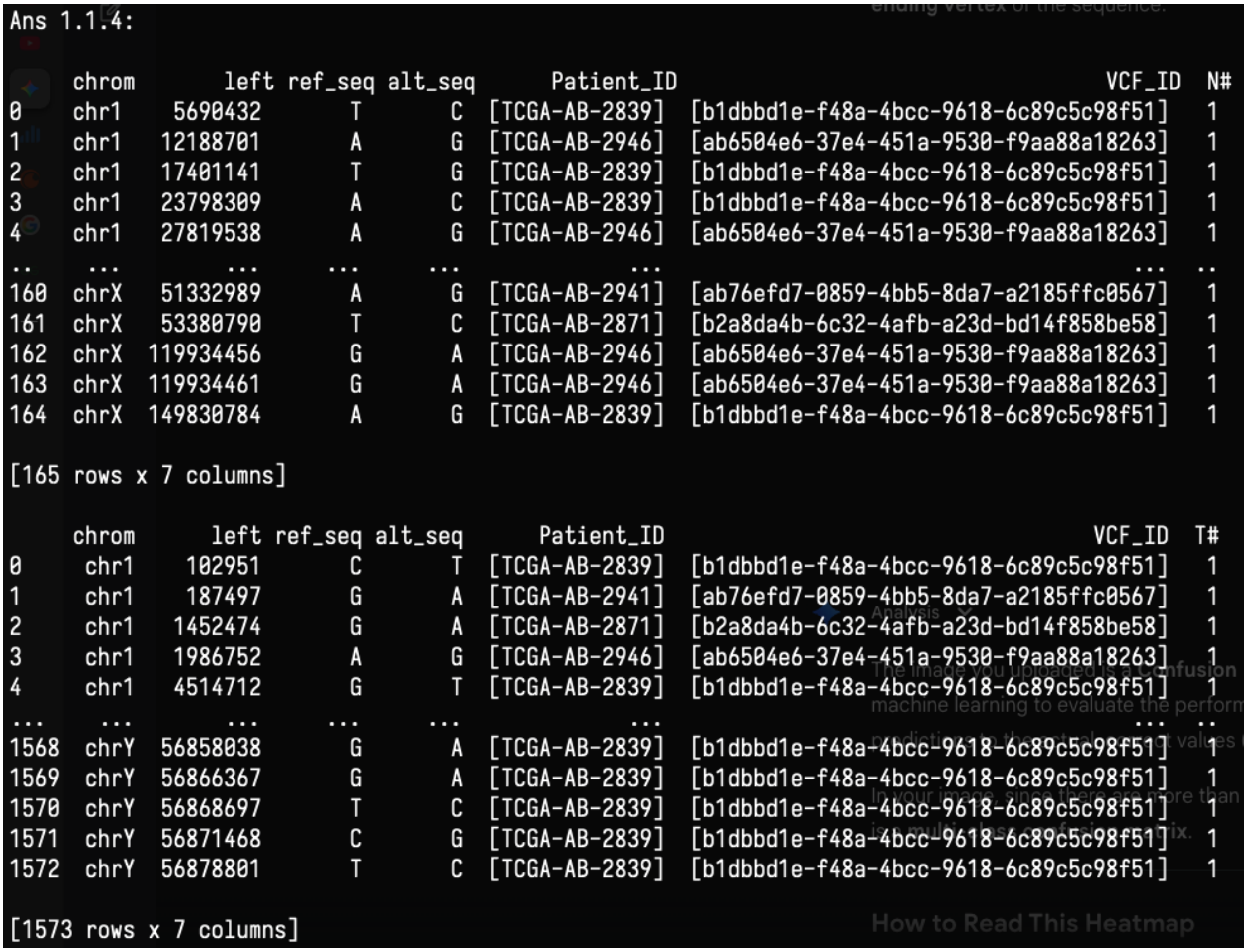
3.2 Repeat the process, except change out the cluster number from 8 to 4, and 3 to 2. How does the clustering change?

When we changed the numbers, the results changed predictably: k=4 is a bit of an **overkill**, creating a fourth cluster by splitting an existing group, but it's much better *than k=8. On the flip side, k=2 is a classic case of* ***under-segmentation****, where the model accurately finds the isolated Setosa, but is forced to* ***lump the other two species together*** *into a single, big group. Just like before, using a bad random start with k=2 makes things worse, proving the initialization problem is a persistent weak spot, regardless of whether you're over- or under-clustering the data.*

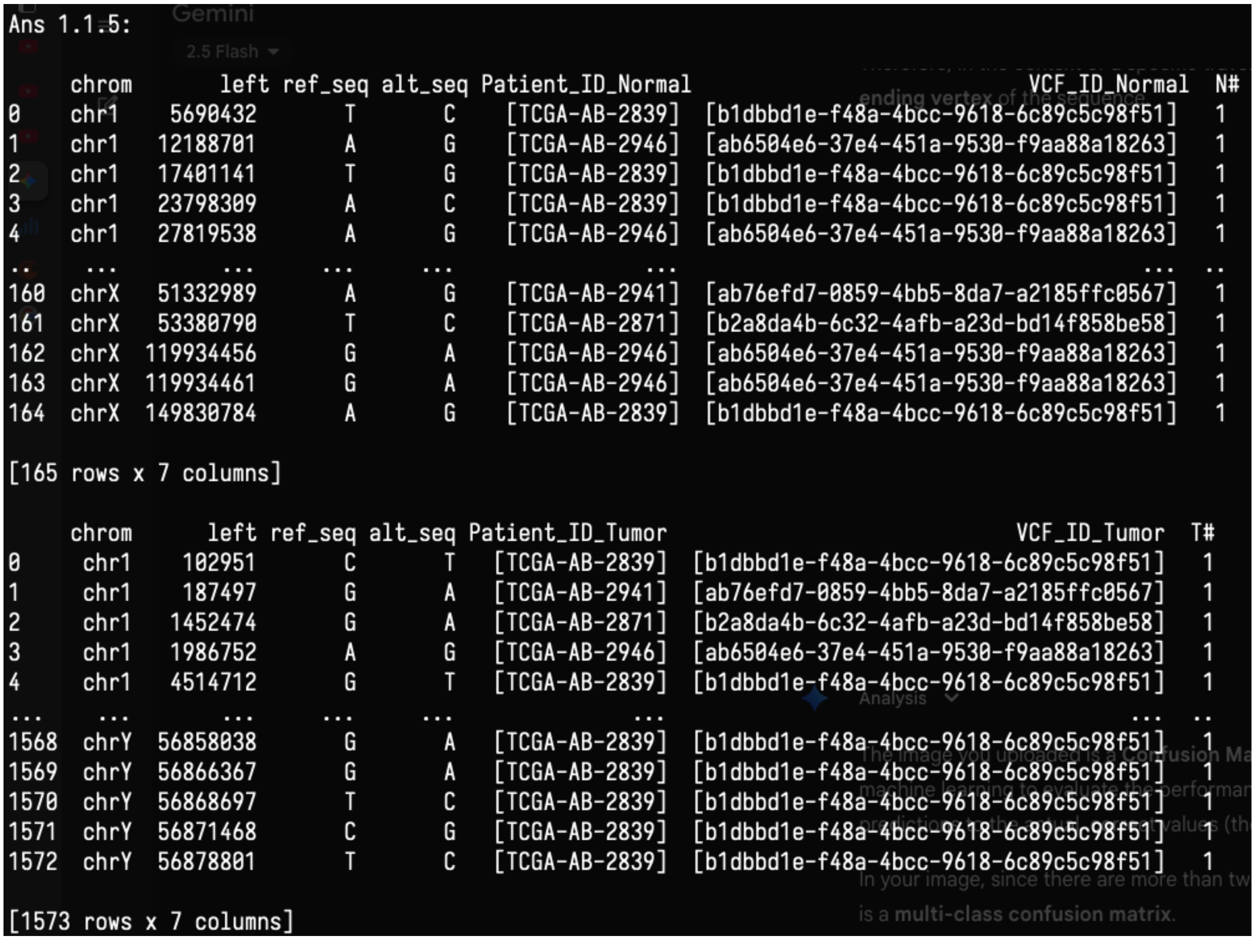
Outputs:



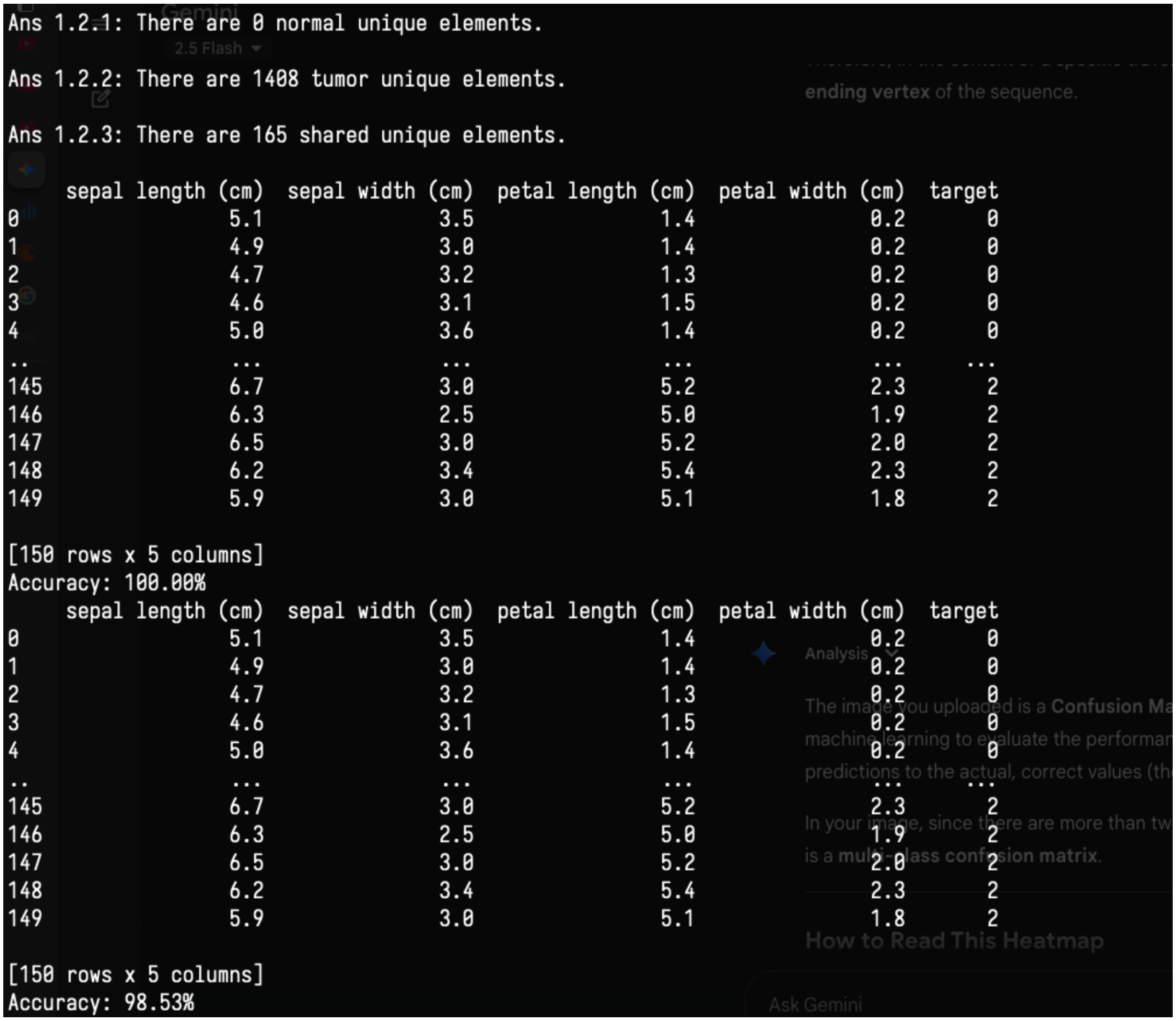
```
Ans 1.1.1: There are 4 unique normal patients

Ans 1.1.2: There are 5 unique tumor patients

Ans 1.1.3:

    chrom       left ref_seq alt_seq      Patient_ID                                  VCF_ID
0    chr1    5690432       T       C  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]
1    chr1   12188701       A       G  [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]
2    chr1   17401141       T       G  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]
3    chr1   23798309       A       C  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]
4    chr1   27819538       A       G  [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]
..    ...        ...     ...     ...             ...                                     ...
160  chrX   51332989       A       G  [TCGA-AB-2941]  [ab76efd7-0859-4bb5-8da7-a2185ffc0567]
161  chrX   53380790       T       C  [TCGA-AB-2871]  [b2a8da4b-6c32-4afb-a23d-bd14f858be58]
162  chrX  119934456       G       A  [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]
163  chrX  119934461       G       A  [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]
164  chrX  149830784       A       G  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]

[165 rows x 6 columns]

     chrom      left ref_seq alt_seq      Patient_ID                                  VCF_ID
0     chr1    102951       C       T  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]
1     chr1    187497       G       A  [TCGA-AB-2941]  [ab76efd7-0859-4bb5-8da7-a2185ffc0567]
2     chr1   1452474       G       A  [TCGA-AB-2871]  [b2a8da4b-6c32-4afb-a23d-bd14f858be58]
3     chr1   1986752       A       G  [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]
4     chr1   4514712       G       T  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]
...    ...       ...     ...     ...             ...                                     ...
1568  chrY  56858038       G       A  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]
1569  chrY  56866367       G       A  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]
1570  chrY  56868697       T       C  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]
1571  chrY  56871468       C       G  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]
1572  chrY  56878801       T       C  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]

[1573 rows x 6 columns]
```



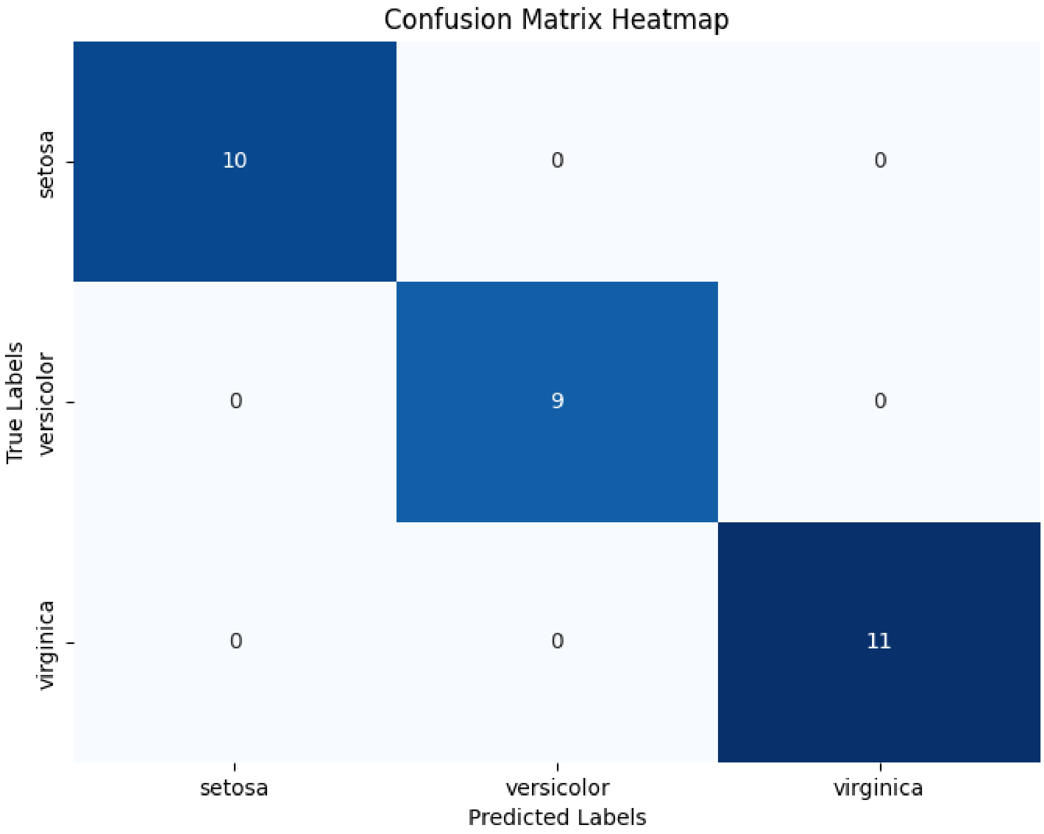
```
Ans 1.1.4:

     chrom       left ref_seq alt_seq      Patient_ID                                  VCF_ID  N#
0    chr1     5690432       T       C  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
1    chr1    12188701       A       G  [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]   1
2    chr1    17401141       T       G  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
3    chr1    23798309       A       C  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
4    chr1    27819538       A       G  [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]   1
..    ...         ...     ...     ...             ...                                     ...  ..
160  chrX    51332989       A       G  [TCGA-AB-2941]  [ab76efd7-0859-4bb5-8da7-a2185ffc0567]   1
161  chrX    53380790       T       C  [TCGA-AB-2871]  [b2a8da4b-6c32-4afb-a23d-bd14f858be58]   1
162  chrX   119934456       G       A  [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]   1
163  chrX   119934461       G       A  [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]   1
164  chrX   149830784       A       G  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1

[165 rows x 7 columns]

     chrom      left ref_seq alt_seq      Patient_ID                                  VCF_ID  T#
0     chr1    102951       C       T  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
1     chr1    187497       G       A  [TCGA-AB-2941]  [ab76efd7-0859-4bb5-8da7-a2185ffc0567]   1
2     chr1   1452474       G       A  [TCGA-AB-2871]  [b2a8da4b-6c32-4afb-a23d-bd14f858be58]   1
3     chr1   1986752       A       G  [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]   1
4     chr1   4514712       G       T  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
...    ...       ...     ...     ...             ...                                     ...  ..
1568  chrY  56858038       G       A  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
1569  chrY  56866367       G       A  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
1570  chrY  56868697       T       C  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
1571  chrY  56871468       C       G  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
1572  chrY  56878801       T       C  [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1

[1573 rows x 7 columns]
```



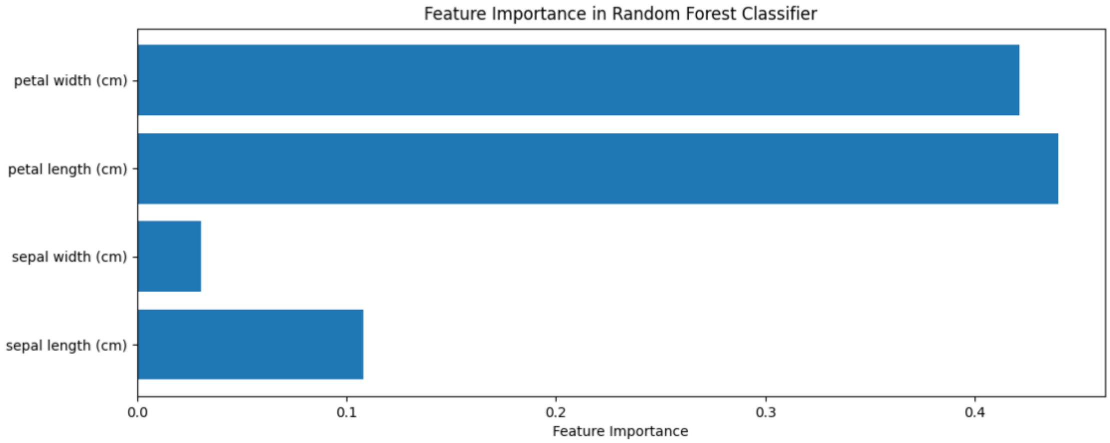
```
Ans 1.1.5:

     chrom       left ref_seq alt_seq Patient_ID_Normal                           VCF_ID_Normal  N#
0     chr1    5690432       T       C    [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
1     chr1   12188701       A       G    [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]   1
2     chr1   17401141       T       G    [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
3     chr1   23798309       A       C    [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
4     chr1   27819538       A       G    [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]   1
..     ...        ...     ...     ...               ...                                     ...  ..
160   chrX   51332989       A       G    [TCGA-AB-2941]  [ab76efd7-0859-4bb5-8da7-a2185ffc0567]   1
161   chrX   53380790       T       C    [TCGA-AB-2871]  [b2a8da4b-6c32-4afb-a23d-bd14f858be58]   1
162   chrX  119934456       G       A    [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]   1
163   chrX  119934461       G       A    [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]   1
164   chrX  149830784       A       G    [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1

[165 rows x 7 columns]

     chrom      left ref_seq alt_seq Patient_ID_Tumor                            VCF_ID_Tumor  T#
0      chr1    102951       C       T   [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
1      chr1    187497       G       A   [TCGA-AB-2941]  [ab76efd7-0859-4bb5-8da7-a2185ffc0567]   1
2      chr1   1452474       G       A   [TCGA-AB-2871]  [b2a8da4b-6c32-4afb-a23d-bd14f858be58]   1
3      chr1   1986752       A       G   [TCGA-AB-2946]  [ab6504e6-37e4-451a-9530-f9aa88a18263]   1
4      chr1   4514712       G       T   [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
...     ...       ...     ...     ...              ...                                     ...  ..
1568   chrY  56858038       G       A   [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
1569   chrY  56866367       G       A   [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
1570   chrY  56868697       T       C   [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
1571   chrY  56871468       C       G   [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1
1572   chrY  56878801       T       C   [TCGA-AB-2839]  [b1dbbd1e-f48a-4bcc-9618-6c89c5c98f51]   1

[1573 rows x 7 columns]
```



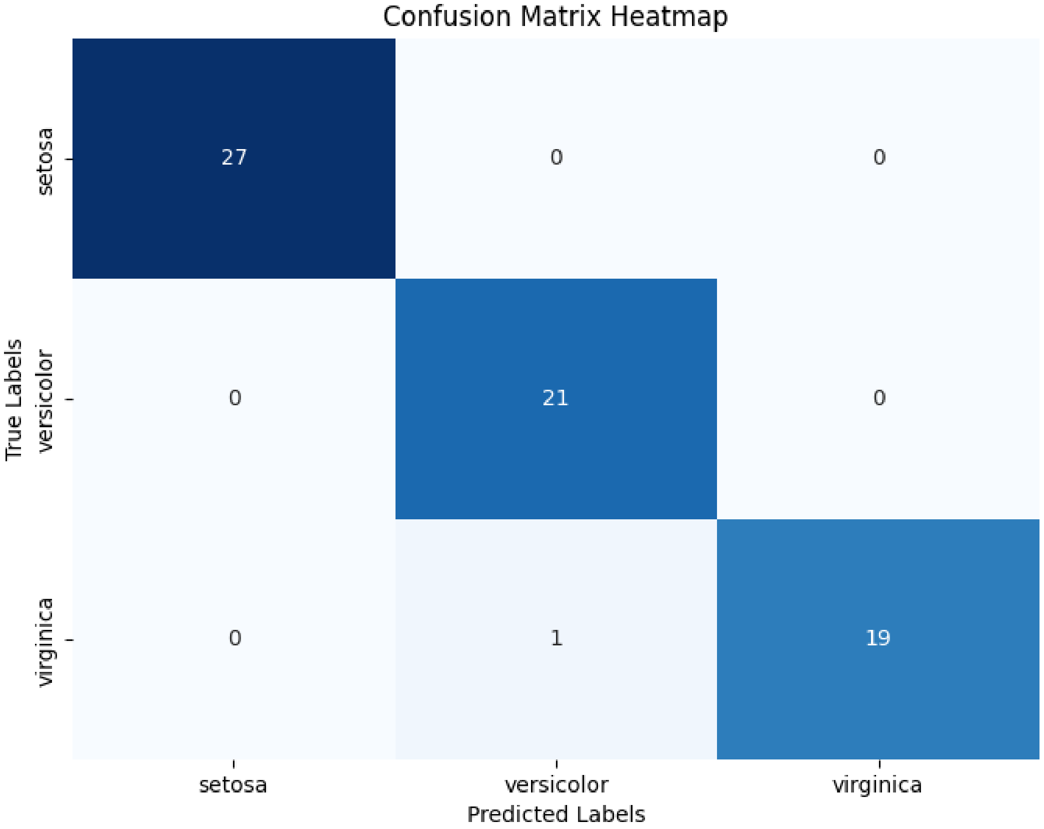
```
Ans 1.2.1: There are 0 normal unique elements.

Ans 1.2.2: There are 1408 tumor unique elements.

Ans 1.2.3: There are 165 shared unique elements.

     sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)  target
0                  5.1               3.5                1.4               0.2       0
1                  4.9               3.0                1.4               0.2       0
2                  4.7               3.2                1.3               0.2       0
3                  4.6               3.1                1.5               0.2       0
4                  5.0               3.6                1.4               0.2       0
..                 ...               ...                ...               ...     ...
145                6.7               3.0                5.2               2.3       2
146                6.3               2.5                5.0               1.9       2
147                6.5               3.0                5.2               2.0       2
148                6.2               3.4                5.4               2.3       2
149                5.9               3.0                5.1               1.8       2

[150 rows x 5 columns]
Accuracy: 100.00%
     sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)  target
0                  5.1               3.5                1.4               0.2       0
1                  4.9               3.0                1.4               0.2       0
2                  4.7               3.2                1.3               0.2       0
3                  4.6               3.1                1.5               0.2       0
4                  5.0               3.6                1.4               0.2       0
..                 ...               ...                ...               ...     ...
145                6.7               3.0                5.2               2.3       2
146                6.3               2.5                5.0               1.9       2
147                6.5               3.0                5.2               2.0       2
148                6.2               3.4                5.4               2.3       2
149                5.9               3.0                5.1               1.8       2

[150 rows x 5 columns]
Accuracy: 98.53%
```

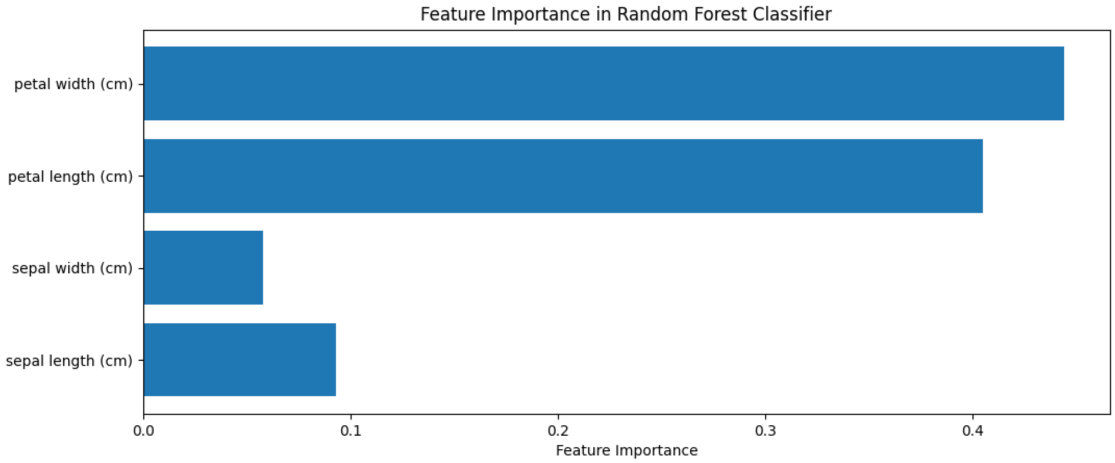


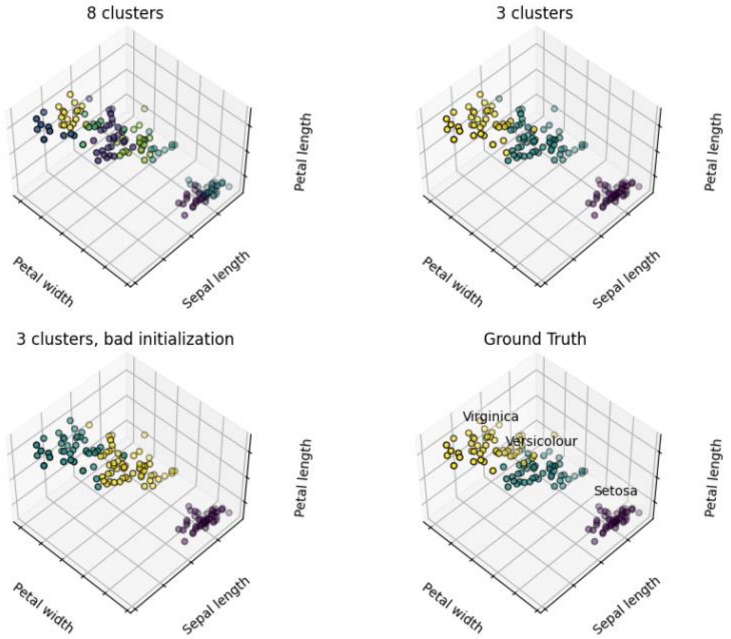
Confusion Matrix Heatmap



Feature Importance in Random Forest Classifier


Confusion Matrix Heatmap
True Labels
setosa
versicolor
virginica
27
0
0
0
21
0
0
1
19
setosa
versicolor
virginica
Predicted Labels


Feature Importance in Random Forest Classifier
petal width (cm)
petal length (cm)
sepal width (cm)
sepal length (cm)
0.0
0.1
0.2
0.3
0.4
Feature Importance

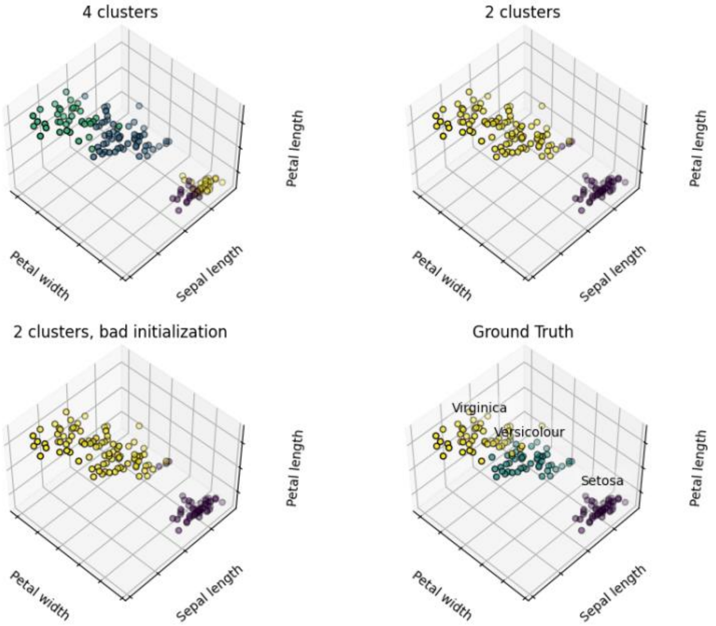


8 clusters

3 clusters

3 clusters, bad initialization

Ground Truth



4 clusters

2 clusters

2 clusters, bad initialization

Ground Truth

**What to Submit:**

1) **A single PDF with responses and screen shots from Part 1,2, and 3. (Your Last Name) _HW5.pdf *****Submit on Blackboard*********

2) **A single Python file with code for Part 1, 2, and 3. (Your Last Name) _HW5.py** ***** Submit on GitHub *****

3) **The two CSV files from Part 1. AML_gene.csv and AML_tx.csv **** ***Submit on Blackboard *****