Assignment - 6
Siddharth Shah – sas151830

ASSIGNMENT 6 - PART I

I. Clustering without PCA:

1. k-means with k = 2: SSE = 12598.19

Confusion matrix (Predicted vs Actual)

	Actual Class		
Predicted Class		ALL	AML
	ALL	26	7
	AML	1	4

2. k-means with k = 3: SSE = 11186.92

Confusion matrix (Predicted vs Actual)

	Actual Class			
Predicted Class		ALL-T	AML	ALL-B
	ALL-T	5	7	8
	AML	0	3	0
	ALL-B	3	1	11

3. Hierarchical clustering with k = 2

Confusion matrix (Predicted vs Actual)

	Actual Class		
Predicted Class		ALL	AML
	ALL	26	11
	AML	0	1

4. Hierarchical clustering with k = 3
Confusion matrix (Predicted vs Actual)

	Actual Class			
Predicted Class		ALL-T	AML	ALL-B
	ALL-T	5	7	8

	AML	0	3	0
	ALL-B	3	1	11

II. Clustering after PCA:

1. k-means with k = 2: SSE = 114.35

Confusion matrix (Predicted vs Actual)

	Actual Class		
Predicted Class		AML	ALL
	AML	5	10
	ALL	6	17

2. k-means with k = 3: SSE = 109.84

Confusion matrix (Predicted vs Actual)

	Actual Class			
Predicted Class		ALL-T	ALL-B	AML
	ALL-T	4	6	4
	ALL-B	3	10	4
	AML	1	3	3

3. Hierarchical clustering with k = 2

Confusion matrix (Predicted vs Actual)

	Actual Class		
Predicted Class		ALL	AML
	ALL	12	4
	AML	15	7

4. Hierarchical clustering with k = 3

Confusion matrix (Predicted vs Actual)

	Actual Class			
Predicted Class		ALL-T	ALL-B	AML
	ALL-T	4	8	4
	ALL-B	4	11	4
	AML	0	0	3

III. Classification (80 points)

Accuracy on train data

algorithms	parameter	Training accuracy
J48decision tree	seed=1	44.73
random forest	seed=3,numtree=100	100
naivebayes		100
bagging	seed=3,numtree=100	81.57

Accuracy on test data

algorithms	parameter	Test accuracy
J48decision tree	seed=1	2.98
random forest	seed=3,numtree=100	0
naivebayes		2.63
bagging	seed=3,numtree=100	7.89