Exam

Graded

Student

Tagore Kosireddy

Total Points

65.5 / 116 pts

Question 1

Question 1 — **9** / 10 pts

1.1 **Q1(a)** — **2** / 2 pts

✔ **– 0 pts** Correct

1.2 **Q1(b)** — **2** / 2 pts

✔ **– 0 pts** Answer is more specific than desired response: *classifier*

1.3 **Q1(c)** — **3** / 3 pts

✔ **– 0 pts** Correct

1.4 **Q1(d)** — **2** / 3 pts

✔ **– 1 pt** Incorrect type for *Calories*

Question 2

Question 2 — **5** / 10 pts

2.1 **Q2(a)** — **1** / 2 pts

✔ **– 1 pt** Answer should not include bar plot

2.2 **Q2(b)** — **1** / 2 pts

✔ **– 1 pt** Answer should not include histograms

2.3 **Q2(c)** — **3** / 3 pts

✔ **– 0 pts** Correct

2.4 **Q2(d)** — **0** / 3 pts

✔ **– 3 pts** Incorrect - D

Question 3

Question 3 **18** / 18 pts

3.1 **Q3 - initial** **5** / 5 pts

✔ **– 0 pts** Correct

3.2 **Step1** **2** / 2 pts

✔ **– 0 pts** Correct

3.3 **Step1 - matrix** **3** / 3 pts

✔ **– 0 pts** Correct

3.4 **Step2** **2** / 2 pts

✔ **– 0 pts** Correct

3.5 **Step2 - matrix** **1.5** / 1.5 pts

✔ **– 0 pts** Correct

3.6 **Step3** **2** / 2 pts

✔ **– 0 pts** Correct

3.7 **Step3 - matrix** **0.5** / 0.5 pts

✔ **– 0 pts** distance value in matrix is incorrect, should be updated.

3.8 **Step4** **2** / 2 pts

✔ **– 0 pts** Correct / correct given error from above

Question 4

Question 4 **4.5** / 24 pts

4.1 **Q4(a)** **0.5** / 3 pts

✔ **– 2.5 pts** incorrect

4.2 **Q4(b)** **1.5** / 9 pts

✔ **– 7.5 pts** Incorrect:

$$H_{orig} - \frac{4}{10}(-\frac{1}{4} * \log_2 \frac{1}{4} - \frac{3}{4} * \log_2 \frac{3}{4}) - \frac{6}{10}(-\frac{3}{6} * \log_2 \frac{3}{6} - \frac{3}{6} * \log_2 \frac{3}{6})$$

4.3 **Q4(c)** **0.5** / 3 pts

✔ **– 2.5 pts** Incorrect:

$$GINI_0 = 1 - (\frac{4}{10})^2 - (\frac{6}{10})^2$$

4.4 **Q4(d)** **2** / 9 pts

✔ **– 7 pts** Incorrect, some connection to GINI formula

$$Gain_{GINI} = GINI_0 - \frac{5}{10}(1 - \frac{4}{5}^2 - \frac{1}{5}^2) - \frac{5}{10}(1 - \frac{2}{5}^2 - \frac{3}{5}^2)$$

Question 5

Question 5 **18.5** / 22 pts

5.1 **Q5(a)i** **0.75** / 1 pt

- ✔ **– 0 pts** Correct
- ✔ **– 0.25 pts** Answer should be unreduced

5.2 **Q5(a)ii** **0.75** / 1 pt

- ✔ **– 0 pts** Correct
- ✔ **– 0.25 pts** Answer should be unreduced

5.3 **Q5(a)iii** **1** / 1 pt

- ✔ **– 0 pts** Correct

5.4 **Q5(a)iv** **1** / 1 pt

- ✔ **– 0 pts** Correct

5.5 **Q5(a)v** **1** / 1 pt

- ✔ **– 0 pts** Correct

5.6 **Q5(a)vi** **1** / 1 pt

- ✔ **– 0 pts** Correct

5.7 **Q5(b)** **13** / 16 pts

- ✔ **– 2 pts** using Multinomial approach and Bernoulli cond. probs
- 💬 **– 1 pt** some mistakes in cond. probabilities

Question 6

Question 6 — **0** / 10 pts

6.1 **Q6(a)** — **0** / 1 pt

✔ **– 1 pt** Question for other section

6.2 **Q6(b)** — **0** / 1 pt

✔ **– 1 pt** Question for other section

6.3 **Q6(c)** — **0** / 1 pt

✔ **– 1 pt** Question for other section

6.4 **Q6(d)** — **0** / 1 pt

✔ **– 1 pt** Question for other section

6.5 **Q6(e)** — **0** / 1 pt

✔ **– 1 pt** Question for other section

6.6 **Q6(f)** — **0** / 1 pt

✔ **– 1 pt** Question for other section

6.7 **Q6(g)** — **0** / 1 pt

✔ **– 1 pt** Question for other section

6.8 **Q6(h)** — **0** / 1 pt

✔ **– 1 pt** Question for other section

6.9 **Q6(i)** — **0** / 2 pts

✔ **– 2 pts** Question for other section

Question 7

Question 7 — **0** / 6 pts

7.1 **Q7(a)** — **0** / 2 pts

✔ **– 2 pts** Question for other section

7.2 **Q7(b)** — **0** / 4 pts

✔ **– 4 pts** Question for other section

Question 8

Question 8 **7** / 10 pts

8.1 Q8(a) **5** / 8 pts

- ✔ **– 1 pt** Missing the following issue:

  Line: 8
  Bug: Data may be imbalanced
  Soln: Use `StratifiedKFold`

- ✔ **– 1 pt** Missing the following issue:

  Line: 13
  Bug: wrong name for "scalar"
  Soln: code should be `X_test_sc = scaler.transform(X_test)`

  Also this should be done only at the very end after found the best hyperparameters.

- ✔ **– 1 pt** Missing the following issue:

  Line: 27
  Bug: wrong sklearn class "metics"
  Soln: code should have `acc = metrics.accuracy_score(...)`

8.2 Q8(b) **2** / 2 pts

- ✔ **– 0 pts** Correct / close to correct

Question 9

Question 9 **3.5** / 6 pts

9.1 Q9(a) **0.5** / 3 pts

- ✔ **– 2.5 pts** Incorrect

9.2 Q9(b) **3** / 3 pts

- ✔ **– 0 pts** Correct

19

Name: First Last	Tagore Kosireddy

ONLY COMPLETE THE QUESTIONS FOR YOUR CLASS.
See labels of **4821 + 5831**, **4821 ONLY**, and **5831 ONLY**.

Question	Points	Score
1 : **4821 + 5831** : Problem Formulation, Data	10	
2 : **4821 + 5831** : Visualization	10	
3 : **4821 + 5831** : Clustering	18	
4 : **4821 + 5831** : Classification I	24	
5 : **4821 + 5831** : Text Mining, Classification II	22	
6 : **4821 ONLY** : Evaluation	10	
7 : **4821 ONLY** : Cross Validation	6	
8 : **5831 ONLY** : Cross Validation	10	
9 : **5831 ONLY** : PCA	6	
Total	100	

Honor Agreement:

- I attest that by signing below, I am the person taking the exam.
- I understand that I am on my honor to do my own work without any assistance from others or outside sources not allowed by my instructor.
- I acknowledge that I can use the following **allowable aids**:
  A single 8.5 × 11" (front and back) notes sheet and blank scrap paper
- I acknowledge that all other aids are **not allowed**
  This includes notes, books, cell phones, smart watches, computers, pdas, etc.
- If I am discovered to have compromised this exam in any way, I understand that I will receive a zero for the exam and be referred to the Office of Student Affairs for violation of university policies regarding academic integrity.
- I understand that my failure to accept this nondisclosure agreement will result in a denial of the early exam, but I will be able to take the exam with the rest of my classmates at the regularly scheduled time and place.

I have read, understood, and agree to this honor agreement.

Student's signature: Tagore Kosireddy

Table 1: Cerial Data

	Name	Company	Type	Calories	Protein	Fat	Fiber	Carbs	Sugar	Used_in_day
0	All Bran	Kelloggs	cold	low	4	1	8.8	7.0	5	0
1	100% Bran	Nabisco	cold	low	4	1	10.3	5.0	6	0
2	Oats	Quaker	hot	low	3	0	2.5	12	1	0
3	Cheerios	General Mills	cold	low	2.5	1.25	2	14.5	1	1
4	Honey Nut Cheerios	General Mills	cold	med	1	0.75	1	11	4.5	1
5	Frosted Flakes	Kelloggs	cold	high	2	0	1	24	8	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

1. (10 points) **4821 + 5831** : Problem Formulation, Data

   You want to build a model to see if your favorite cereal will run out in the dorms in one day using data in Table 1.

   (a) (2 points) What type of general learning problem is considered? *Be brief, less than 5 words.*

   supervised learning, features, target

   (b) (2 points) What general type of model could be built to predict this outcome? *Be brief, less than 8 words.*

   classification model, e.g. - Random Forest

   (c) (3 points) For each variable, how would it be used in your problem formulation? Which should be **I**gnored? a target **C**lass? a **P**redictor?

Variable	Type (I/C/P)	Variable	Type (I/C/P)
Name	I	Used_in_day	C
Calories	P	Sugar	P
Fat	P	Company	I

   (d) (3 points) For each variable below, state it's type: nominal, ordinal, interval, ratio

Type	● Nominal	○ Ordinal	○ Interval	○ Ratio
Fiber	○ Nominal	○ Ordinal	○ Interval	● Ratio
Calories	○ Nominal	○ Ordinal	○ Interval	● Ratio

2. (10 points) **4821 + 5831** : Visualization

Suppose we have rental housing data; see sample below.

	address	rent_price	num_bedrooms	sq_feet	utilities	parking
0	500 College	850	2	1100	no	1
1	1880 Townsend	900	3	1300	yes	2
2	80 Agate	1600	4	1950	yes	2
3	450 Walnut	1400	2	1250	no	2

(a) (2 points) Fill in all suitable visualizations for comparing the distribution of rent price for 2, 3, and 4 bedroom units.

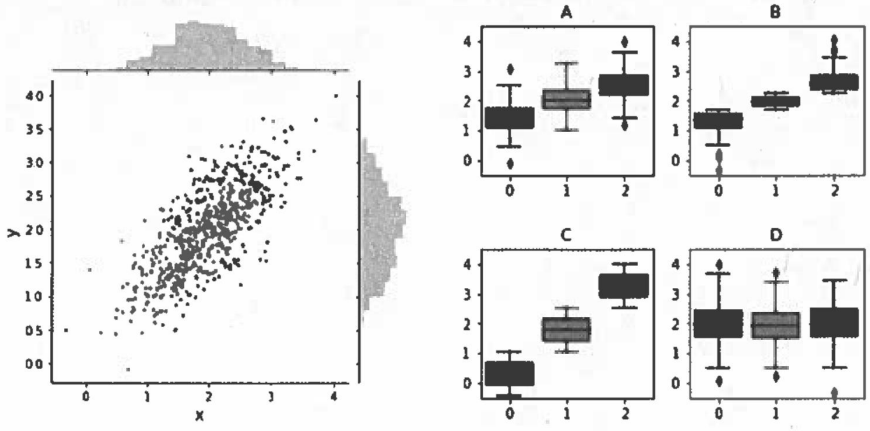
● Bar plot ● Violin plot ○ Scatter plot ● Box Plot ○ None of these

(b) (2 points) Fill in all suitable visualizations for examining how rent price and square footage relate.

○ Density plots ● Scatter plot ● Histograms ○ Bar Plots ○ None of these

The scatterplot generated by `sns.jointplot` below to the left visualizes a dataset containing 600 (x, y) pairs. You partition the pairs by their x value into three groups:

- The 200 with the lowest x values are in group 0.
- The 200 with intermediate x values are in group 1.
- The 200 with the highest x values are in group 2.



(c) (3 points) Fill in the letter of the boxplot that shows the distribution of x values in each of the three groups.

○ A ● B ○ C ○ D

(d) (3 points) Fill in the letter of the boxplot that shows the distribution of y values in each of the three groups.

○ A ○ B ○ C ● D

3. (18 points) **4821 + 5831** : Hierarchical Clustering

Consider the data set presented in Table 2 of 5 samples (A-E) over 2 variables (X1 and X2).

Table 2: Cluster Data

Sample	X1	X2
A	1	1
B	3	2
C	2	7
D	1	6
E	4	2

Perform agglomerative hierarchical clustering with **complete linkage** using the Manhattan distance metric.

For each step of the clustering, say what are the next two items (data samples or clusters) to be combined and report the updated distance matrix. *Only report the lower triangular part of the distance matrix and order samples and groups alphabetically*

The initial distance matrix for the data is:

	A	B	C	D	E
A	0	-	-	-	-
B	3	0	-	-	-
C	7	6	0	-	-
D	5	6	2	0	-
E	4	1	7	7	0

Step 1: Combine B and E.
Report new distance matrix between groups.

	A	BE	C	D
A	0	-	-	-
BE	4	0	-	-
C	7	7	0	-
D	5	7	2	0

Step 2: Combine C and D.
Report new distance matrix between groups.

	A	BE	CD
A	0	-	-
BE	4	0	-
CD	7	7	0

Step 3: Combine A and BE.
Report new distance matrix.

	ABE	CD
ABE	0	-
CD	7	0

Step 4: Combine ABE and CD.

4. (24 points) **4821 + 5831** : Classification I

Consider the 10 training examples, given in Table 3 over 2 variables (A and B) and the target variable. *For each part of this question, you do not have to calculate the final value, but instead show the formula, e.g.,*

$$H_0 = \log_2 \frac{2}{5} + 2^3 - \gamma + \sqrt{3} - \log_{10} \frac{3}{5}^2$$

(a) (3 points) What is the overall entropy of the samples before splitting? Call this quantity H_0.

$H_0 = -\sum_{i=1}^{k} p_{M,i} \log_2 p_{M,i}$

$H_0(A) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{4}\log_2\frac{3}{4}$

$H_0(B) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{4}{5}\log_2\frac{4}{5}$

$H_0 = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{4}\log_2\frac{3}{4} - \frac{3}{5}\log_2\frac{3}{5} - \frac{4}{5}\log_2\frac{4}{5}$

Instance	A	B	Target
1	F	T	+
2	F	T	-
3	F	T	+
4	F	T	+
5	F	F	-
6	T	F	+
7	F	F	-
8	T	F	+
9	T	F	-
10	T	T	+

Table 3: Classification Data

(b) (9 points) Calculate the information gain when splitting on A. *You can use H_0 above as a variable in your solution here.*

$\text{Information gain} = H_0 - \sum_{l=1}^{2} \frac{n_l}{n_m} H(X_{m,l})$

$= H_0 - \frac{10}{20}\left(-\frac{3}{10}\log_2\frac{3}{10} - \frac{3}{10}\log_2\frac{3}{10}\right)$

(c) (3 points) Calculate the gini of the samples before splitting. Call this quantity $GINI_0$.

$$GINI(x_m) = 1 - \sum_{j=1}^{k} p_{mj}^2$$

$$GINI_0 = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{4}\right)^2 - \left(\frac{3}{5}\right)^2\left(\frac{4}{5}\right)^2$$

Instance	A	B	Target
1	F	T	+
2	F	T	-
3	F	T	+
4	F	T	+
5	F	F	-
6	T	F	+
7	F	F	-
8	T	F	+
9	T	F	-
10	T	T	+

Table 3: Classification Data

(d) (9 points) Calculate the gain in gini when splitting on B. *Provide formula, you do not have to calculate final value. You can use $GINI_0$ as a variable in the formula.*

$$\text{gain in gini} = GINI_0 - \sum_{l=1}^{2} \frac{n_l}{n} GINI(x_{ml})$$

$$= GINI_0 - \frac{10}{20}\left[\left(1 - \left(\frac{4}{10}\right)^2\left(\frac{3}{10}\right)^2\right]\right.$$

5. (22 points) **4821 + 5831** : Text Mining, Classification II

Consider the problem of classifying the following documents:

ID	Document	Class
1	CMX CMX MQT MSP	+
2	DTW ESC ORD MSP	-
3	DTW DTW ORD	-
4	ESC ORD MSP	+
5	DTW CMX ESC	-

(a) (6 points) Estimate the following conditional probabilities (use laplace smoothing). *Report as unreduced fraction, e.g.,* $\frac{1}{3}, \frac{2}{10}, \frac{0}{5}$.

Bernoulli Model

Cond.	Prob.
$P(CMX \mid +)$	$\frac{1}{2}$
$P(ORD \mid +)$	$\frac{1}{2}$
$P(MSP \mid -)$	$\frac{2}{5}$

Multinomial Model

Cond.	Prob.
$P(MQT \mid -)$	$\frac{1}{16}$
$P(CMX \mid +)$	$\frac{3}{13}$
$P(DTW \mid -)$	$\frac{5}{16}$

(b) (16 points) Show the calculation for predicting the class of a test document "MSP DTW ORD" using the Bernoulli Naive Bayes model. Use Laplace smoothing for conditional probabilities. *Provide formula, you do not have to calculate final value.*

$$\hat{P}(c) = \frac{N_c}{N}, \quad \hat{P}(t|c) = \frac{N_{ct} + 1}{N_c + 2}$$

$$P(+ \mid MSP, DTW, ORD) = P(+) P(MSP|+) P(DTW|+) P(ORD|+)$$

$$P(- \mid MSP, DTW, ORD) = P(-) P(MSP|-) P(DTW|-) P(ORD|-)$$

$$P(+) = \frac{2}{5}, \quad P(-) = \frac{3}{5},$$

$$P(+ \mid MSP, DTW, ORD) = \frac{2}{5} \times \frac{3}{4} \times \frac{1}{4} \times \frac{1}{2}$$

$$P(- \mid MSP, DTW, ORD) = \frac{3}{5} \times \frac{2}{5} \times \frac{2}{3} \times \frac{1}{2}$$

6. (10 points) **4821 ONLY** : Evaluation

Suppose we train a model to predict whether an email is Spam or Not Spam. After training, we apply it to test set of 500 new emails and the model produces the following confusion matrix. Assume Spam is the positive class.

		Actual Class	
		Spam	Not Spam
Predicted Class	Spam	50	25
	Not Spam	75	350

Report as unreduced fraction, e.g., $\frac{30}{70}$, $\frac{70}{500}$, $\frac{200}{380}$

(a) (1 point) Compute the precision of the model.

(a) _____

(b) (1 point) Compute the recall of the model.

(b) _____

(c) (1 point) Compute the sensitivity of the model.

(c) _____

(d) (1 point) Compute the specificity of the model.

(d) _____

(e) (1 point) Compute the positive predictive value of the model.

(e) _____

(f) (1 point) Compute the negative predictive value of the model.

(f) _____

(g) (1 point) Compute the accuracy of the model.

(g) _____

(h) (1 point) Compute the error rate of the model.

(h) _____

(i) (2 points) For this problem, which metric is better at judging model performance:
○ Accuracy ○ Error Rate ○ AUC

7. (6 points) **4821 ONLY** : Cross Validation

   Suppose we have a training data set of 400 samples, and a test set of 100 samples. We want to explore the best hyperparameters for an SVM. The candidate hyperparameters are C = [0.01, 0.1, 1, 10] and the linear and polynomial kernels with degree = [2, 3].

   (a) (2 points) A student suggests performing 5-fold cross validation with grid search to find the optimal hyperparameters. Is this choice of 5-fold cross-validation reasonable?

   - ○ Yes
   - ○ No, since with have 7 possible hyperparameters we should use 7-fold cross validation.
   - ○ No, since we have 100 test samples, we should use 10-fold cross validation.
   - ○ No, cross validation should never be used to select hyperparameters.

   (b) (4 points) How many different SVM models will be trained in the 5-fold cross validation with grid search to select the best hyperparameters?
   *Your answer should be left as a formula, e.g., "3 * 7 + 10 * 30"*

8. (6 points) **5831 ONLY** : Cross Validation

Below is some code that attempts to implement a cross-validation procedure to find the optimal hyperparameters for a Random Forest Classifier.

We want to perform 5-fold cross validation, with 2 hyperparameters (5 values for hyperparameter 1 and 3 values for hyperparameter 2).

Assume:

- An initial split of the data created the Data Frames `X_trainval`, `X_test` and Data Series `y_trainval`, `y_test` was already performed.
- The candidate values for each hyperparameters are in the lists: `hparams1`, `hparams2`.
- The random forest is constructed with the call `ensemble.RandomForestClassifier()` where the appropriate hyperparameter values are set.

```
from sklearn.model_selection import KFold
from sklearn import ensemble
from sklearn import metrics
from sklearn import preprocessing
import numpy as np
import pandas as pd

kf = KFold(n_splits = 5)
cv_scores = []

scaler = preprocessing.StandardScaler().fit(X_trainval)
X_trval_sc = scaler.transform(X_trainval)
X_test_sc = scalar.transform(X_test)

for hp1 in hparams1:
    for hp2 in hparams2:

        valid_scores = []
        for tr_indx, val_indx in kf.split(X_trainval):
            X_train, X_valid = X_trval_sc.iloc[tr_indx], X_trval_sc.iloc[val_indx]
            y_train, y_valid = y_trainval.iloc[tr_indx], y_trainval.iloc[val_indx]

            model = ensemble.RandomForestClassifier(hyperp1 = hp1, hyperp2 = hp2)
            model.fit(X_trainval, y_trainval)
            yhat = model.predict(X_valid)

            acc = metics.accuracy_score(y_valid, yhat)
            valid_scores.append(acc)

        cv_scores.append(np.mean(valid_scores))

```

(a) (8 points) Describe any bugs (if they exist) in the code. *A bug can be invalid Python syntax as well as improper operations to achieve the goal of the code*

In the table below, list the line number and a brief description of the bug and a solution.

Note, you may not need all rows of the table. Your response should be clear and succinct. Lengthy, rambling answers will be penalized.

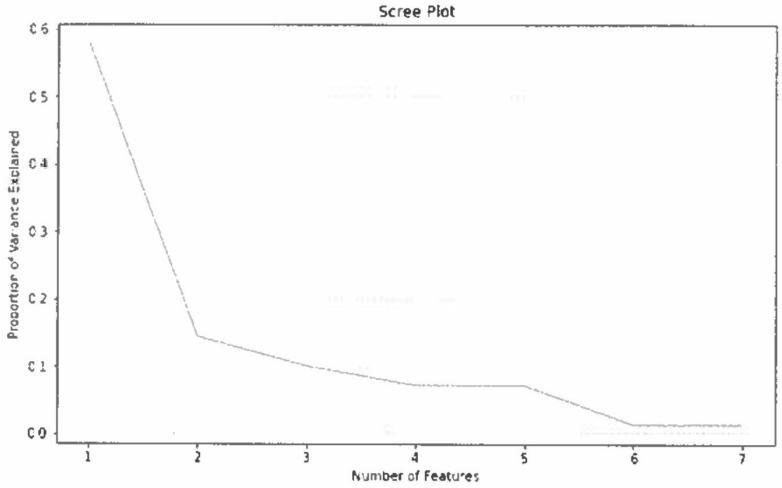
Line Num(s)	Bug / Solution
	Bug: hyperparameter are not defined ie values Soln: hparam1 :- {x, y}, hparam2 :- {A, B}
	Bug: performing scaling before cross validation Soln: scaling must be done inside the for loop
	Bug: fitting x_trainval, Y_trainval to model Soln: x_train, Y_train should be used.
	Bug: Soln:
	Bug: Soln:
	Bug: Soln:

(b) (2 points) Explain what is data leakage. *Be clear and succinct, less than 16 words.*

Exposing the validation or test data to the model in the training.

9. (6 points) **5831 ONLY** : PCA

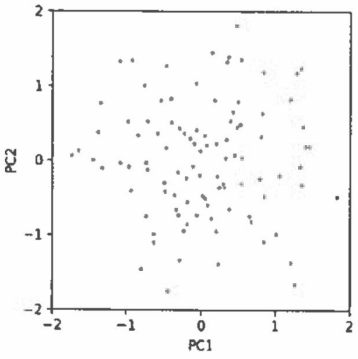
A student successfully applies PCA to a data set and makes a plot showing the proportion of variation explained for each principal component.



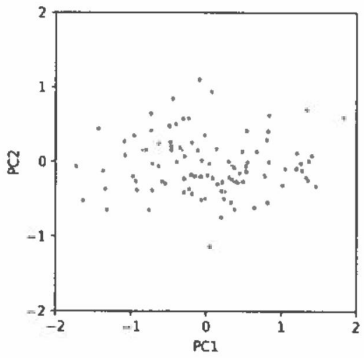
(a) What is the minimum number of principal components to capture 80% of the variance?

● 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ All

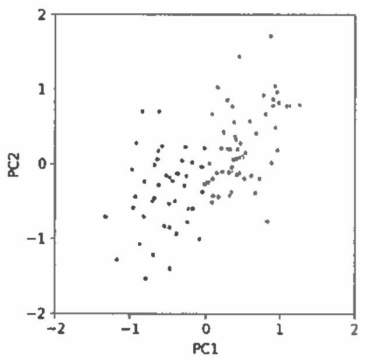
The student now plots the first two principal components in a scatter plot.



Plot A



Plot B



Plot C

(b) Which of the three plots could potentially display the first two principal components?

○ Plot A ● Plot B ○ Plot C