CS 334 - Machine Learning
Fall 2021 Midterm Exam

Instructions:

- This exam is open book and open notes (you can access textbooks, lecture notes, homeworks).
- The exam is governed by **Emory Honor Code**.
- You cannot consult anyone for answers. You cannot save or share any content of the exam with anyone.
- Make sure you justify your answer, *no points* will be given if there is no explanation or it is incorrect. Write any assumptions down if you feel there is any ambiguity.
- Please submit your answers as a single PDF.
- Make sure your writing, if any, is legible.
- The total points are 105 including 5 bonus points.

Name: Elijah Chou

Emory Network ID: echou4

I will abide by the Emory College Honor Code:

Signature: 

Point Distribution

Question	Points
k-NN and Model Selection	17
Decision Tree	14
Linear Regression Regularization	16
Naive Bayes	16
Model Evaluation	9
Bias/Variance Trade-off	9
Classification and Feature Representation	24
Total:	105

Question 1: k-NN and Model Selection (17 points)

Imagine that you are given the following set of training examples. Each feature can take on one of three nominal values: a, b, or c. We will use *leave-one-out cross validation* to select the best 'k'.

Sample	F_1	F_2	F_3	Class
1	c	b	c	+
2	a	c	a	+
3	c	a	c	+
4	b	a	c	-
5	b	c	a	-
6	b	b	b	-

(a) (3 points) What is a good distance metric to use for this problem and why?

A good distance metric to use for this problem is Manhattan distance or Hamming distance which measures how many attributes are the same for two records in this case.

(b) (6 points) What is the estimated leave-one-out cross-validation error of 1-nearest neighbor on this dataset? In the case of ties, use the worst case scenario.

△ error = 0 if class prediction is correct

1. Sample 1 nearest to #3 → error = 0
2. sample 2 " " #5 → error = 1
3. " 3 " " #1 & #4 → worst case is #4 → error = 1
4. " 4 " " #3 → error = 1
5. " 5 " " #2 → error = 1
6. " 6 " " #1 in worst case → error = 1

$$CV\ error = \frac{1+1+1+1+1+0}{6} = \frac{5}{6} \approx 83\%$$

(c) (6 points) What is the estimated leave-one-out cross validation error of 3-nearest neighbor on this dataset? In the case of ties, use the worst case scenario.

1. sample 1 nearest to #3,4,6 → error = 1
2. sample 2 " " #5,4,6 in worst case → error = 1
3. sample 3 " " #1,4,5 in " " → error = 1
4. " 4 " " #3,1,5 " " " → error = 1
5. " 5 " " #2,4,6 → error = 0
6. " 6 " " #1,4,5 → error = 0

$$\text{CV error} = \frac{1+1+1+1+0+0}{6} = \frac{4}{6} \approx 67\%$$

(d) (2 points) What would you choose for k? Justify your reason.

I would choose k=3 because the average CV error for leave-one-out CV is lower in k=3 than k=1.

Question 2: Decision Tree (14 points)

The following dataset captures if people will be hired by Google (Y) or not (N) based on their CS 334 grade (high or low), their GPA (high or low), and whether or not they did an internship during their undergraduate studies.

ML grade	GPA	Internship	Hired
L	H	Y	Y
L	L	N	N
L	L	Y	N
L	L	N	N
H	H	Y	Y
H	L	Y	Y
H	H	N	Y
H	L	N	Y

(a) (3 points) What is the entropy H(Whether Hired | Internship = N)? You can keep the number in fractions. Briefly justify.

$H(\text{Whether Hired} \mid \text{Internship} = N) = -\frac{2}{4} \cdot \log\left(\frac{2}{4}\right) - \frac{2}{4} \cdot \log\left(\frac{2}{4}\right)$

$= 1$

This is because given Internship = N, half of them are hired & the other half are not hired. Therefore entropy is at its maximum of 1.

(b) (3 points) What is the entropy H(Whether Hired | GPA = H)? Briefly justify.

$H(\text{Whether Hired} \mid \text{GPA} = H) = -\frac{3}{3} \cdot \log_2\left(\frac{3}{3}\right) - \frac{0}{3} \cdot \log_2\left(\frac{0}{3}\right)$

$= 1 \cdot 0 - 0$

$= 0$

Given GPA = H, then we know for sure that they are hired. No probability of predicting wrong here.

(c) (8 points) Draw the full decision tree learned for this data using entropy and assuming no pruning. Show your work. You don't have to compute every entropy values if you can justify your answers.

Before first split: (YNNNYYYY)

Split on ML Grade: (YNNN, YYYY); on GPA: (YYY, NNNY)

on Internship: (YNYY, NNYY)

$\text{Info Gain}_A = \text{Info}(D) - \text{Info}_A(D)$

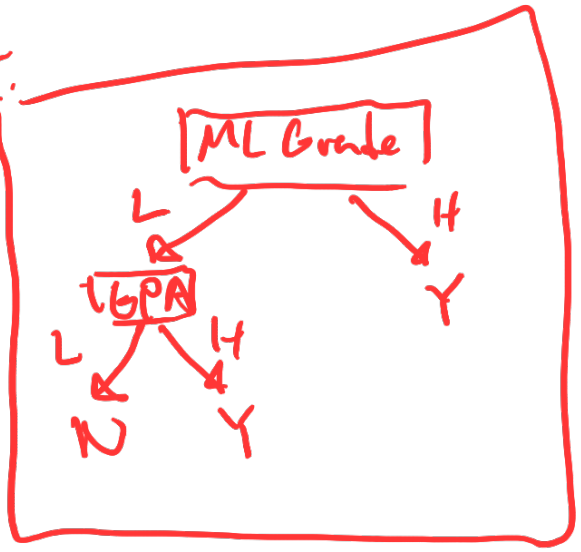
$\text{Info}_{ML}(D) = \frac{4}{8}\left(-\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right) + 0 \simeq 0.405$

$\text{Info}_{GPA}(D) = \frac{5}{8}\left(-\frac{3}{5}\log_2\left(\frac{1}{4}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right) + 0 \simeq 0.607$

$\text{Info}_{Intern}(D) = \frac{4}{8}\left(-\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right) + \frac{4}{8}\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) \simeq 0.906$

$\text{Info}_{Intern}(D) > \text{Info}_{GPA}(D) > \text{Info}_{ML}(D)$

⇒ Select ML as first split.



For ML = L, before second split: (YNNN)

Split on GPA: (Y, NNN); on Internship: (YN, NN)

∴ Split on GPA since it splits into 100% pure leaf nodes

Question 3: Linear Regression Regularization (16 points)

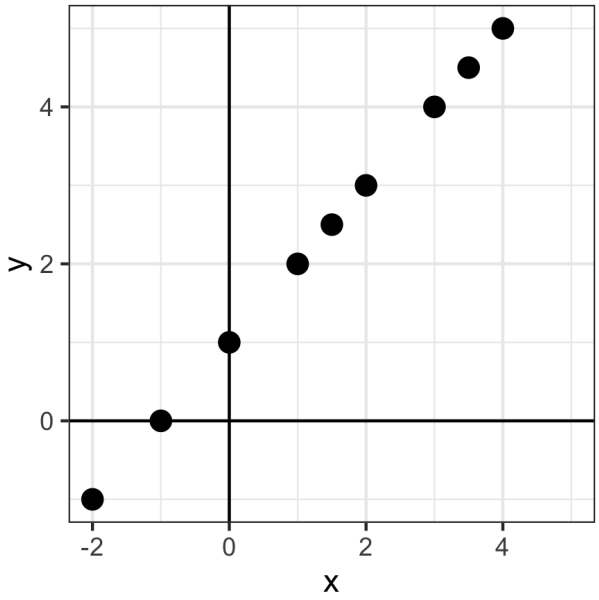


Figure 1: A 2-D dataset with the feature x and the target y.

We're attempting to solve the regression task depicted in Figure 1 with the regularized linear regression model, where the target is on the y-axis and the feature is on the x-axis. The loss function for the regularization model is:

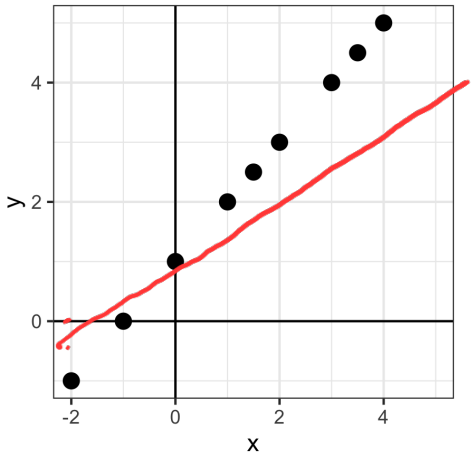
$$\min \frac{1}{2n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x))^2 + \lambda_0 |\beta_0| + \lambda_1 |\beta_1|$$

for either a zero or very large value of λ_i. This problem is similar to the LASSO regularization except for the fact that each coefficient has a different regularization parameter, and β_0 is being penalized explicitly. Observe that without any regularization on the standard linear regression model, the residual sum of squares is exactly zero.

(a) Given the training data in Figure 1, how does the residual sum of errors change with regularization of each parameter β_j? For this part, we will restrict ourselves to *only regularizing a single parameter at a given time* (i.e., either $\lambda_0 = 0$ and $\lambda_1 >> 1$ or $\lambda_0 >> 1$ and $\lambda_1 = 0$). State whether the RSS increases, stays the same (zero), or decreases for each β_j for a very large λ. Draw the resulting model (the figure is repeated in each part) that will be obtained from a high value of the regularization parameter and provide a brief justification for your answers. **No credit will be awarded for an incorrect explanation**.

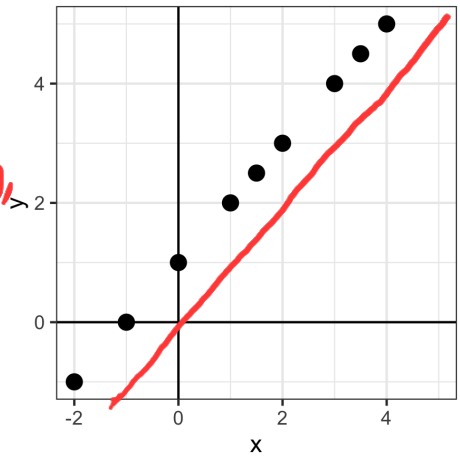
i. (6 points) Regularization of β_1 (i.e., $\lambda_1 >> 1; \lambda_0 = 0$)

The resulting model will be as follows because by penalizing B_1 explicitly, B_1 will decrease (smaller slope) but B_0, the intercept, will stay the same.



ii. (6 points) Regularization of β_0 (i.e., $\lambda_0 >> 1; \lambda_1 = 0$)

The resulting model will be as shown because by penalizing β_0 explicitly, β_0 will decrease (lower intercept) but β_1, the slope, will still stay the same.



(b) (4 points) Assume that we've set the regularization parameters to be the same: $\lambda_0 = \lambda_1 = \lambda$. Plot what you expect the coefficient path for both β_0, β_1 to be as a function of λ for $\lambda \geq 0$. Make sure to clearly mark the values of λ on your graph (in terms of 0 and ∞).

Question 4: Naive Bayes (16 points)

You have been hired by Microsoft to improve the spam detection in Outlook. Your goal is to estimate the probability a new email m is spam using the terms $(w_1, w_2, \cdots, w_d)$ in the email. For example, the terms might be "money", "!", "act now", "bargain", "click now", etc.

(a) (5 points) Your boss tells you the current system estimates the probability of spam given the new email message as follows:

$$P(\text{spam} = 1 | \text{new email } m) = \frac{\#\text{ of spam e-mails with terms } w_1, w_2, \cdots, w_d \text{ in training}}{\#\text{ of total e-mails with terms } w_1, w_2, \cdots, w_d \text{ in training}}$$

Explain why the current spam detection system might have poor performance based on the above calculation.

It might have poor performance because there may be too many features in the model, which will make H too complex & therefore have bad performance.

(b) (8 points) Describe how to estimate the probability an email is spam using Naive Bayes. Your answer should include multiple formulas.

$$P(\text{spam}=1 | w_1, w_2, \ldots, w_d) \propto P(w_1, w_2, \ldots w_d | \text{spam}=1) P(\text{spam}=1)$$

$$= P(w_1|\text{spam}=1) * P(w_2|\text{spam}=1) \times \ldots$$

$$\times P(w_d|\text{spam}=1) \cdot P(\text{spam}=1)$$

$$P(\text{spam}=0 | w_1, w_2, \ldots, w_d) \propto P(w_1, w_2, \ldots, w_d | \text{spam}=0) P(\text{spam}=0)$$

$$= P(w_1|\text{spam}=0) \times P(w_2|\text{spam}=0) \times \ldots$$

$$\times P(w_d|\text{spam}=0) P(\text{spam}=0).$$

A classification of spam depends on which probability is larger.

Part (b) continued.

(c) (3 points) Why is the Naive Bayes classifier in (b) better than the current Outlook spam detection in (a)?

the classifier in (b) is better than the current one in (a) because it is much faster to train (in a single scan) & classify, it isn't sensitive to irrelevant features, and it handles all types of data well.

Question 5: Model Evaluation (9 points)

You are a reviewer for the International Conference on Awesome Machine Learning Algorithms, and you are assigned the following papers that have the following statements. Would you accept or reject the paper? You *must justify your answer* with at least 1-2 sentences.

(a) (3 points) "My algorithm achieves the best model amongst these other 5 baselines on the task of wine quality prediction. There is an improvement of almost 2% in terms of accuracy measured on the training data."

Reject the paper because better performance in accuracy on training data does not reflect how robust or how good it performs with new, unseen data, thereby not necessarily making the model better than the 5 baselines.

(b) (3 points) "Our new algorithm is awesome at detecting early-stage prostate cancer. On a large cancer dataset with over 1 million samples and a prevalence (rate) of prostate cancer of 0.01% (1 out of every 10,000), my model achieves an accuracy rate of 99.5% compared to 90.0% for a logistic regression model."

Reject because they base their conclusion on accuracy alone, which is only really helpful/valid if the data is balanced. However, since the data is imbalanced as depicted by the low prevalence rate of 0.01%, this conclusion cannot be made due to the inflation of accuracy caused by data imbalance.

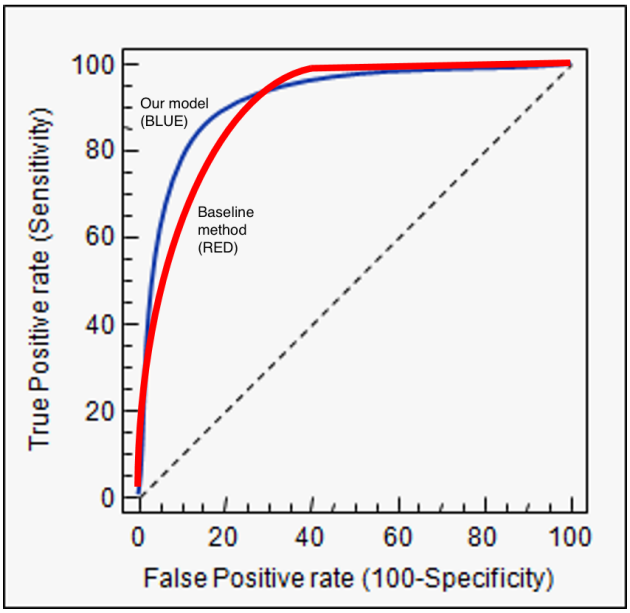


Figure 2: ROC curve for paper (d)

(c) (3 points) "Our model always outperforms the state-of-the-art method. Figure 2 illustrates the ROC curve for our model (blue) compared to the state-of-the-art method (red) on the test dataset. As can be seen from figure, our area under the ROC curve (AUC) is significantly higher than the baseline (red) method."

Reject the paper because the AUC was not calculated/shown so we cannot say for sure if the AUC was really greater. Also, since the ROC curves intersect, it also means that their model doesn't ALWAYS outperform the baseline method.

Question 6: Bias/Variance Trade-off (9 points)

For the following questions, explain what effect (i.e., increase, no change, decrease) the following operations will have on *both the bias and variance of your resulting model.* You *must* justify your answers. No points will be given if there is no explanation or it is incorrect.

(a) (3 points) You increase K for KNN classifier on the same training dataset.

The bias will increase but variance will decrease because the model becomes less overfitted to the data & thus becomes more robust in its prediction.

(b) (3 points) You set the regularization parameter (λ) in ridge-regularized linear regression from 1 to 100 for training on the same dataset.

The bias will increase but variance decreases because by penalizing the coefficients more in regularization, the model becomes more generalized to better fit unseen data.

(c) (3 points) You train a decision tree with the same hyperparameters on a random subset of the features for the same dataset.

There will be no change on either bias or variance because if you don't change the hyperparameters but still train on same dataset, then performance should be the same.

Question 7: Classification and Feature Representation (24 points)

Consider the following binary classification problem (A = circles vs B = triangles) shown in Figure 3.

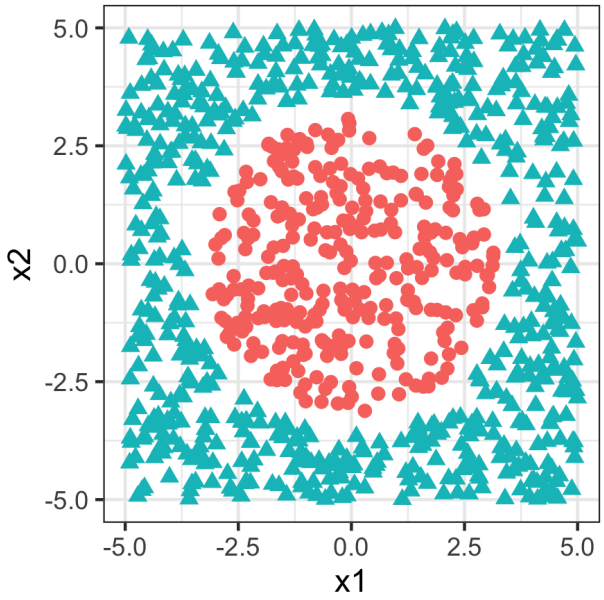
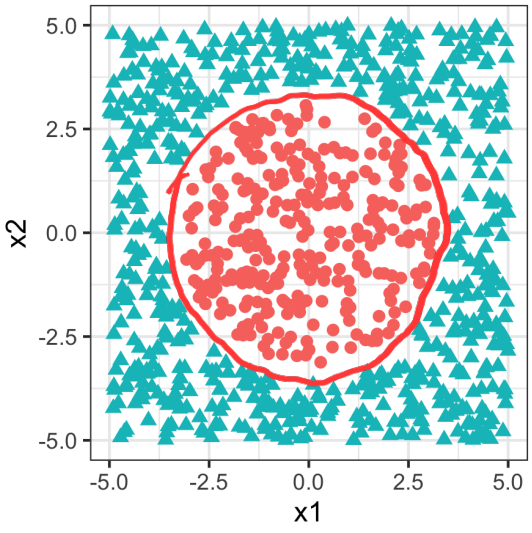


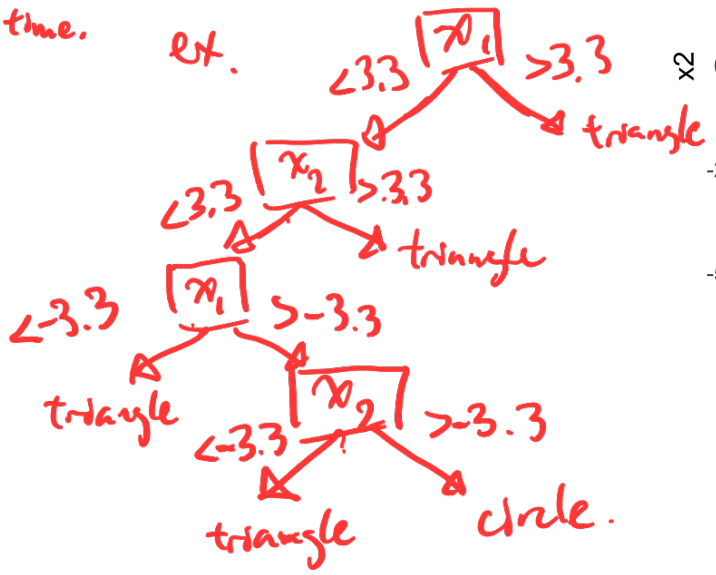
Figure 3: A binary classification problem with A = circles and B = triangles.

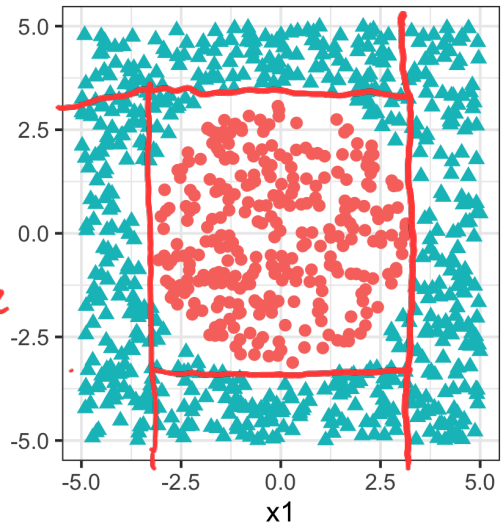
(a) (2 points) If you were given this dataset, what should the boundary be? Draw the boundary *you* would use to determine if it was class A or class B.



(b) (6 points) You pass the dataset to a decision tree to learn the boundary. Draw the boundary that the *decision tree* would achieve. Explain how you arrived at this conclusion (you can do this by drawing the decision tree).

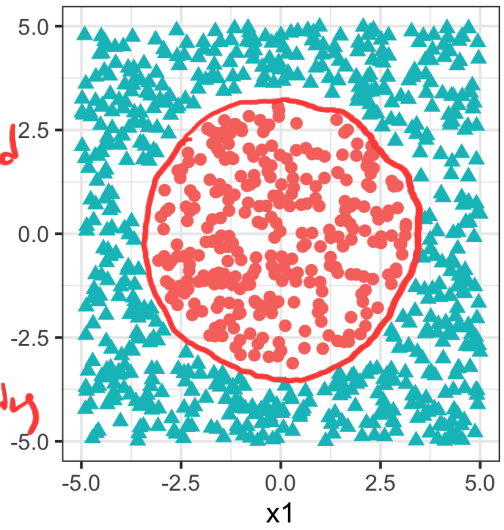
In decision trees, vertical/horizontal boundaries can only be made since they split on one feature at a time.





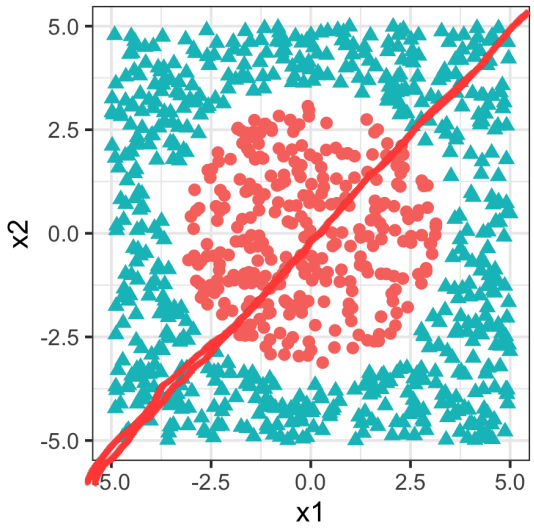
(c) (6 points) You pass the dataset to a 1-NN to learn the boundary. Draw the boundary that the *1-NN* would achieve. Explain how you arrived at this conclusion.

The 1-NN boundary would look like this because the classifier will predict based on the closest neighbor, which means that everything in the boundary will be closest to a circle while everything outside the boundary will be closest to a triangle.



(d) (5 points) You pass the dataset to a logistic regression model to learn the boundary. Draw the boundary that the *logistic regression* model would achieve. Explain how you arrived at this conclusion.

Since logistic regression is a linear predictor, it will choose this particular boundary, which does its best to produce a boundary as a linear boundary.



(e) (5 points) Are there new features you can introduce (i.e., new features that you can derive from the existing features) that could allow you to achieve better results across all three of the discussed models (decision tree, 1-NN, logistic regression). If yes, then what are the new features and why would they work. If no, why not?

Logistic regression: enlarge space of features by including transformations (ex. x_1^2, x_1^3, x_1x_2, $x_1x_2^2$). By adding these new features, non-linear boundaries are possible & make it easier for logistic regression to improve

1-NN: no, because it is already the best and adding more features may complicate the model more

decision tree: no, because there are no new features that can be extracted from the current features to allow the decision tree to create non-linear boundaries.

(Page left intentionally blank)

(Page left intentionally blank)