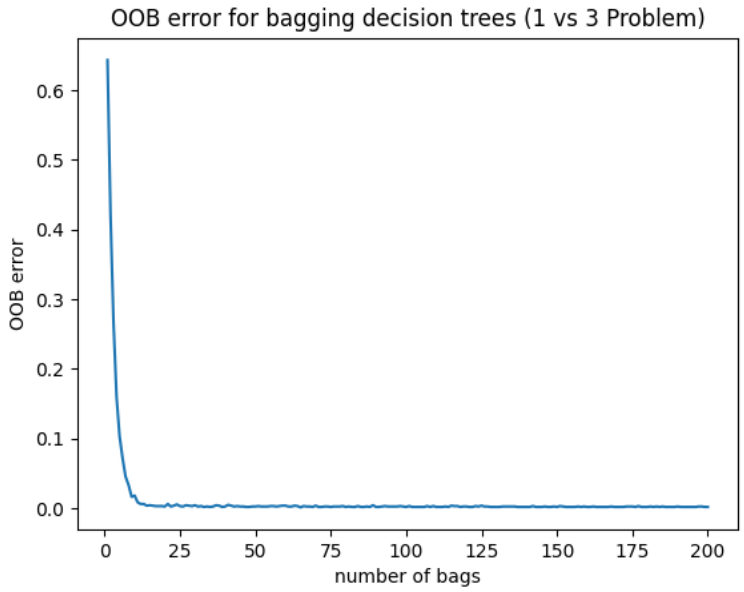
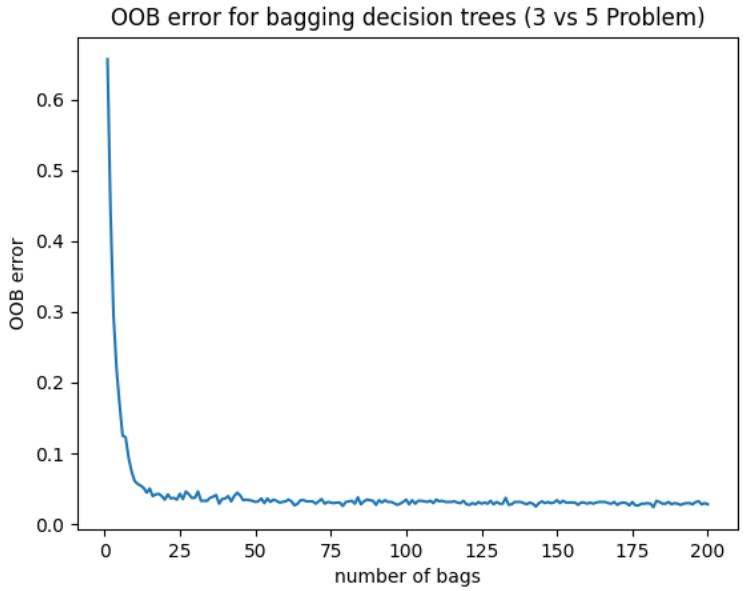
1)

a.



- OOB Error of bagging decision trees : 0.0024053
- Test Error of
  - single decision tree : 0.01395
  - bagging decision trees : 0.01163



- OOB Error of bagging decision trees : 0.0313
- Test Error of
  - single decision tree : 0.1288
  - bagging decision trees : 0.079754

b.

- 3 vs 5 Problem has higher OOB & Test Errors of both single decision & bagging decision trees. This can mean that 1 and 3 are easier to differentiate which lead to smaller errors.
- Increasing the number of bags will lower the OOB Error in both Problems, especially significantly at around 10 bags for both problem. After 10 bags, OOB error seems to be stable
- OOB error seems to be smaller than the test error in both problems.

2,

a, $P_{yes} = \frac{2}{5}$, $P_{no} = \frac{3}{5}$

↳ $H(D) = \sum_{i=1}^{k} P_i \log_2 \frac{1}{P_i} = \frac{2}{5}\log_2\left(\frac{5}{2}\right) + \frac{3}{5}\log_2\left(\frac{5}{3}\right) = 0.971$

- Gain (D, color):

$H(D_{purple}) = \frac{2}{4}\log_2\left(\frac{4}{2}\right) + \frac{2}{4}\log_2\left(\frac{4}{2}\right) = 1$

$H(D_{red}) = 1\log_2(1) = 0$

→ Gain (D, color) $= 0.971 - \left(\frac{4}{5} \times 1 + \frac{1}{5} \times 0\right) = 0.171$ (1)

- Gain (D, stripes)

$H(D_{no}) = \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0$

$H(D_{yes}) = \frac{1}{3}\log_2\left(\frac{3}{1}\right) + \frac{2}{3}\log_2\left(\frac{3}{2}\right) = 0.918$

→ Gain (D, stripes) $= 0.971 - \left(\frac{2}{5} \times 0 - \frac{3}{5} \times 0.918\right) = 0.9202$ (2)

- Gain (D, Texture)

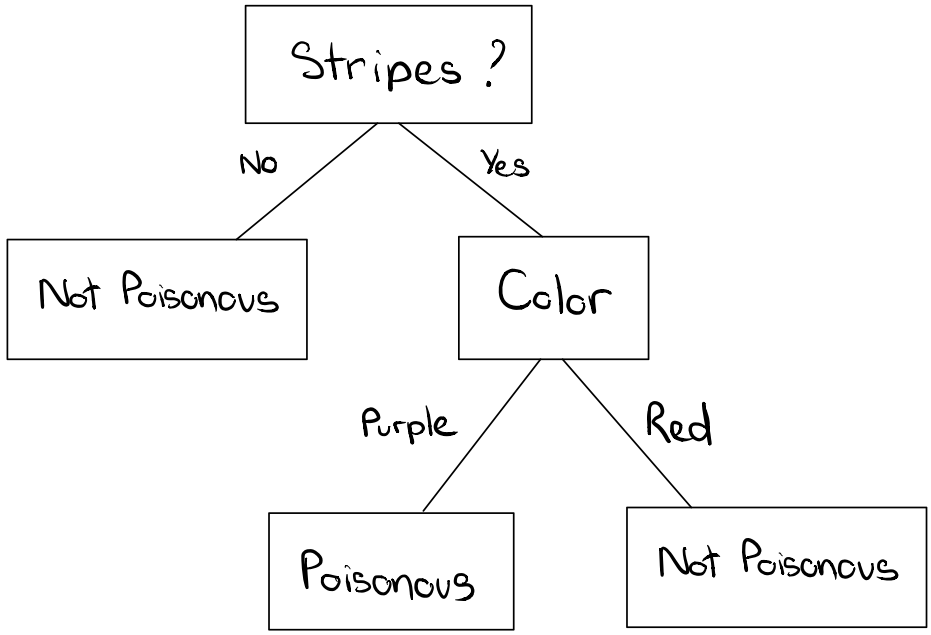
$H(D_{smooth}) = \frac{2}{3}\log_2\left(\frac{3}{2}\right) + \frac{1}{3}\log_2\left(\frac{3}{1}\right) = 0.918$

$H(D_{rough}) = \frac{1}{2}\log_2\left(\frac{2}{1}\right) + \frac{1}{2}\log_2\left(\frac{2}{1}\right) = 1$

→ Gain (D, texture) = $0.971 - \left(\frac{3}{5} \times 0.918 + \frac{2}{5} \times 1\right) = 0.0202$ ③

①②③ → The root attribute of the tree is "Stripes"

b,



3)

- At the first iteration:

$E_{in}(g_1) = 0.2$; $g_1(\vec{x}_n) = 1$ (since majority will be positive (+1) - 80%)

↳ $\varepsilon_1 = 0.2 \rightarrow \gamma = \sqrt{\frac{1-\varepsilon_1}{\varepsilon_1}} = \sqrt{\frac{1-0.2}{0.2}} = 2$

↳ $\alpha_1 = \frac{1}{2}\ln\left(\frac{1-\varepsilon_1}{\varepsilon_1}\right) = \frac{1}{2}\ln\left(\frac{1-0.2}{0.2}\right) = \ln 2$

↳ $z_1 = \gamma\varepsilon_1 + \frac{1}{\gamma}(1-\varepsilon_1) = 2(0.2) + \frac{1}{2}(1-0.2) = \frac{4}{5}$

$D_1(n) = \frac{1}{N}$

↳ $D_2(n) = \frac{1}{z_1}\left(\frac{1}{N}\right)e^{-\alpha_1 y_n g_1(\vec{x}_n)}$

$= \frac{5}{4}\left(\frac{1}{N}\right)e^{-(\ln 2)y_n} = \frac{5}{4}\left(\frac{1}{N}\right)\left(\frac{1}{2}\right)^{y_n}$

- $y_n = 1 \rightarrow D_2(n) = \frac{5}{4}\left(\frac{1}{N}\right)\left(\frac{1}{2}\right) = \frac{5}{4}\left(\frac{1}{2N}\right) = \frac{5}{8N}$
- $y_n = -1 \rightarrow D_2(n) = \frac{5}{4}\left(\frac{1}{N}\right)\left(\frac{1}{2}\right)^{-1} = \frac{5}{4}\left(\frac{2}{N}\right) = \frac{5}{2N}$

↳ Cumulative weights of { (+) points = $\frac{5}{8N}(80\%\,N) = 50\%$; (-) points = $\frac{5}{2N}(20\%\,N) = 50\%$

∴ Using depth 0 decision tree is not a good idea since at the second iteration, it will be a random guess (50% (+) vs 50% (-))

4)

Euclidean distance : $d(x, x') = \sqrt{(x - x')^2} = |x - x'|$

→ $d(x_{target}, x) = [0.2, 1.8, 3.8, 1.2, 0.2]$

↳ (x_1, y_1), (x_4, y_4) & (x_5, y_5) are 3 nearest neighbor

↳ $y_{prediction} = \frac{y_1 + y_4 + y_5}{3} = \frac{5 + 11 + 8}{3} = 8$

5,

a. I chose the article *“Study funds gender and skin-type bias in commercial artificial-intelligence systems.”* The article talks about error rates in facial detection between skin color and gender. Many US major technology companies claimed that their facial-recognition system had an accuracy rate of over 97% but they never mentioned that their training data used for this system is highly biases towards lighter skinned and male individuals. In short, for women, the darker the skin color, the higher the error rates yielded by the system.

b. There was a lot of bias in the training data when developing this facial-recognition system. Big companies cannot jump into conclusion with biased data. The system yields nearly 50% error rate for darkest-skinned women on a simple binary classification task.

c. One of the solutions could involve developing and training a more balanced data set with more pictures and images of female and darker-skinned individuals. Companies and scientists must be fair with no bias and/or discrimination and cannot jump into conclusion with biased data. Researchers can definitely gather more training images of female and darker-skinned individuals to train the model and system to yield higher accuracy rate and to more importantly, avoid a biased training data set.