1. **a)**

$E[Z] = E[\,|X - Y|^2\,] = E[(X - Y)^2]$

$= E[X^2 - 2XY - Y^2]$

$= E[X^2] - 2E[XY] + E[Y^2]$ #By the independence of variables

$= E[X^2] - 2E[X]E[Y] + E[Y^2]$ #By the independence of variables

$E[X] = E[Y] = \frac{1}{2} * (a + b) = \frac{1}{2} * (0 + 1) = \frac{1}{2}$ # X, Y ~ Unif(0, 1)

$Var[X] = Var[Y] = 1/12 * (b - a)^2 = 1/12 * (1 - 0)^2 = 1/12$ # Formula of σ^2 for Unif(0, 1)

$E[X^2] = E[Y^2] = Var[X] + E[X]^2 = Var[Y] + E[Y]^2$ # $Var[X] = E[X^2] - E[X]^2$

$= 1/12 + (\frac{1}{2})^2 = \frac{1}{3}$

$E[Z] = E[X^2] - 2E[X]E[Y] + E[Y^2]$

$= \frac{1}{3} - 2 * \frac{1}{2} * \frac{1}{2} + \frac{1}{3}$ # Substitute values

$= \frac{1}{6}$

$E[X^4] = \frac{1}{4+1} \sum_{i=0}^{4} 0^i 1^{4-i} = \frac{1}{5}$ # Fourth moment for Unif(0, 1)

$Var[X^2] = Var[Y^2] = E[X^4] - E[X^2]^2 = \frac{1}{5} - (\frac{1}{3})^2 = 4/45$

$Var[XY] = E[(XY)^2] - (E[XY])^2$ # Formula for variances

$= E[X^2Y^2] - E[X]^2E[Y]^2$ #By the independence of variables

$= E[X^2]E[Y^2] - E[X]^2E[Y]^2$

$=\frac{1}{3} * \frac{1}{3} - \frac{1}{4} * \frac{1}{4}$

$= 1/9 - 1/16$

$= 7/144$

$Var[Z] = Var[\,|X - Y|^2\,] = Var[X^2 -2XY + Y^2]$

$= Var[X^2 - 2XY] + Var[Y^2] + 2Cov[X^2 - 2XY, Y^2]$

$= Var[X^2] + Var[-2XY] - 2Cov[X^2, 2XY] + Var[Y^2] + 2Cov[X^2 - 2XY, Y^2]$

$= Var[X^2] + 4Var[XY] - 2Cov[X^2, 2XY] + Var[Y^2] + 2Cov[X^2 - 2XY, Y^2]$

$= Var[X^2] + 4Var[XY] - 2Cov[X^2, 2XY] + Var[Y^2] + 2Cov[X^2 - 2XY, Y^2]$

$= 4/45 + 4 * (7/144) - \frac{1}{6} + 4/45 - \frac{1}{6}$

$= 7/180$

Therefore, the expected value of Z is ⅙ , and the variance of Z is 7/180.

b) $E[R] = E[\,\sum_{i=1}^{d} Z_i\,] = \sum_{i=1}^{d} E[Z_i] = \sum_{i=1}^{d} \frac{1}{6} = \frac{d}{6}$

$$\text{Var[R]} = \text{Var}[\sum_{i=1}^{d} Z_i] = \sum_{i=1}^{d} \text{Var}[\text{Z}_\text{i}] = \sum_{i=1}^{d} \frac{7}{180} = \frac{7d}{180}$$

c) The maximum possible squared Euclidean distance within the d-dimensional unit cube:

$$\sqrt{\sum_{i=1}^{d} (1-0)^2}^{\,2} = \sum_{i=1}^{d} 1 = \text{d}$$

The standard deviation of R: $\sigma = \sqrt{Var[R]} = \frac{7}{180}\sqrt{d}$

The expected value of R is the probability-weighted mean of R, so $\mu = \frac{d}{6}$.

The mean of R supports “most points are far away” because of the mean itself is almost as great as the maximum possible squared Euclidean distance within the d-dimensional unit cube.

The standard deviation is quite small compared to the mean of R, so these points are “approximately the same distance”.

2. a) $\text{H(X)} = \sum_{x \in X} p(x) log_2 \frac{1}{p(x)}$

$$= -\sum_{x \in X} p(x) log_2 p(x)$$

Since p(x) ∈ [0, 1], then $\text{log}_2\text{p(x)}$ will always be non-positive.

⇒ $\text{p(x)log}_2\text{p(x)}$ will always be non-positive

⇒ $\text{-p(x)log}_2\text{p(x)}$ will always be non-negative

⇒ $-\sum_{x \in X} p(x) log_2 p(x)$ will always be non-negative.

b) $\text{H(X, Y)} = -\sum_{x \in X}\sum_{y \in Y} \text{p(x, y)log}_2\text{p(x, y)}$

$$= -\sum_{x \in X}\sum_{y \in Y} \text{p(x, y)log}_2\text{p(x)p(y)} \qquad \text{\#By the independence of variables}$$

$$= -\sum_{x \in X}\sum_{y \in Y} (\text{p(x, y)log}_2\text{p(x)}) - \sum_{x \in X}\sum_{y \in Y} (\text{p(x, y)log}_2\text{p(y)})$$

$$= -\sum_{x \in X} (\text{p(x)log}_2\text{p(x)}) - \sum_{y \in Y} (\text{p(y)log}_2\text{p(y)}) \qquad \text{\#By the law of total probability}$$

$$= \text{H(X)} + \text{H(Y)}$$

c) $\mathrm{H(X, Y)} = -\sum_{x \in X}\sum_{y \in Y} \mathrm{p(x, y)log_2p(x, y)}$

$= -\sum_{x \in X}\sum_{y \in Y} \mathrm{p(x, y)log_2p(x)p(y|x)}$

$= -\sum_{x \in X}\sum_{y \in Y} \mathrm{(p(x, y)log_2p(x))} - \sum_{x \in X}\sum_{y \in Y} \mathrm{(p(x, y)log_2p(y|x))}$

$= -\sum_{x \in X}\sum_{y \in Y} \mathrm{(p(x, y)log_2p(x))} + \mathrm{H(Y|X)}$

$= -\sum_{x \in X} \mathrm{(p(x)log_2p(x))} + \mathrm{H(Y|X)}$ #By the law of total probability

$= \mathrm{H(X) + H(Y|X)}$

d) $\mathrm{KL(p||q)} = \sum_{x \in X} p(x) log_2 \frac{p(x)}{q(x)}$

$= -\sum_{x \in X} p(x) log_2 \frac{q(x)}{p(x)}$

Since log(x) is a concave function, -log(x) will be a convex function.

Therefore, by Jensen's Inequality:

$$\mathrm{E[-log_2(q(x)/p(x))] \geq -log_2(E[q(x)/p(x)])}$$

This implies that:

$\mathrm{KL(p||q)} = -\sum_{x \in X} p(x) log_2 \frac{q(x)}{p(x)}$

$\geq -log_2 \sum_{x \in X} p(x) \frac{q(x)}{p(x)}$

$= \mathrm{-\ log_2} \sum_{x \in X} q(x)$

$= \mathrm{-log_2 1}$

$= 0$

Thus, KL(p||q) is always non-negative.

e) $\mathrm{KL(p(x, y)||p(x)p(y))} = \sum_{x \in X}\sum_{y \in Y} \mathrm{p(x, y)log_2} \frac{p(x,y)}{p(x)p(y)}$

$= \sum_{x \in X}\sum_{y \in Y} \mathrm{p(x, y)log_2} \frac{p(x)p(y|x)}{p(x)p(y)}$

$= \sum_{x \in X}\sum_{y \in Y} \mathrm{p(x, y)log_2} \frac{p(y|x)}{p(y)}$

$$= \sum_{x\in X}\sum_{y\in Y}(p(x, y)\log_2 p(y|x)) - \sum_{x\in X}\sum_{y\in Y}(p(x, y)\log_2 p(y))$$

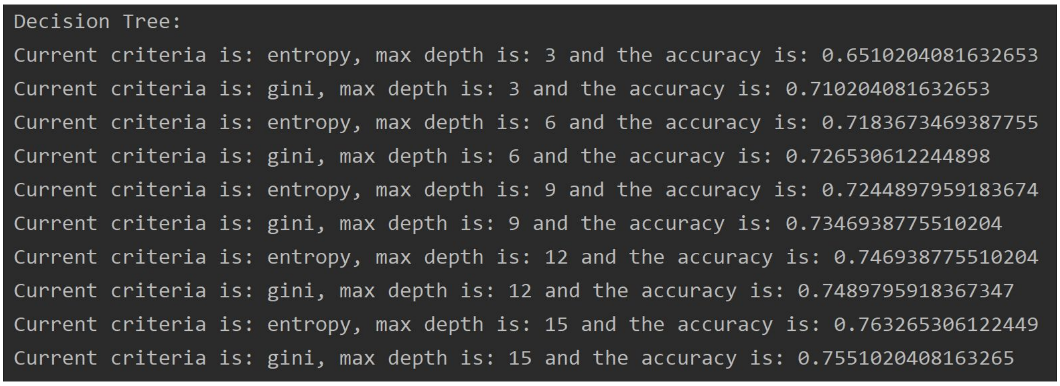
$$= \sum_{x\in X}\sum_{y\in Y}(p(x, y)\log_2 p(y|x)) - \sum_{x\in X}(p(y)\log_2 p(y))$$

$$= - H(Y|X) + H(X)$$

$$= H(X) - H(Y|X)$$

$$= I(Y|X)$$

3. a) See hw1_code.py

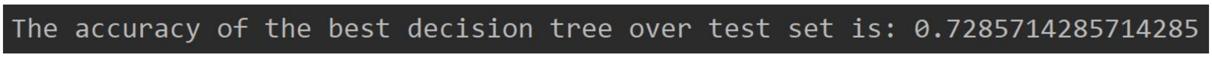


```
Decision Tree:
Current criteria is: entropy, max depth is: 3 and the accuracy is: 0.6510204081632653
Current criteria is: gini, max depth is: 3 and the accuracy is: 0.710204081632653
Current criteria is: entropy, max depth is: 6 and the accuracy is: 0.7183673469387755
Current criteria is: gini, max depth is: 6 and the accuracy is: 0.726530612244898
Current criteria is: entropy, max depth is: 9 and the accuracy is: 0.7244897959183674
Current criteria is: gini, max depth is: 9 and the accuracy is: 0.7346938775510204
Current criteria is: entropy, max depth is: 12 and the accuracy is: 0.746938775510204
Current criteria is: gini, max depth is: 12 and the accuracy is: 0.7489795918367347
Current criteria is: entropy, max depth is: 15 and the accuracy is: 0.763265306122449
Current criteria is: gini, max depth is: 15 and the accuracy is: 0.7551020408163265
```

b)

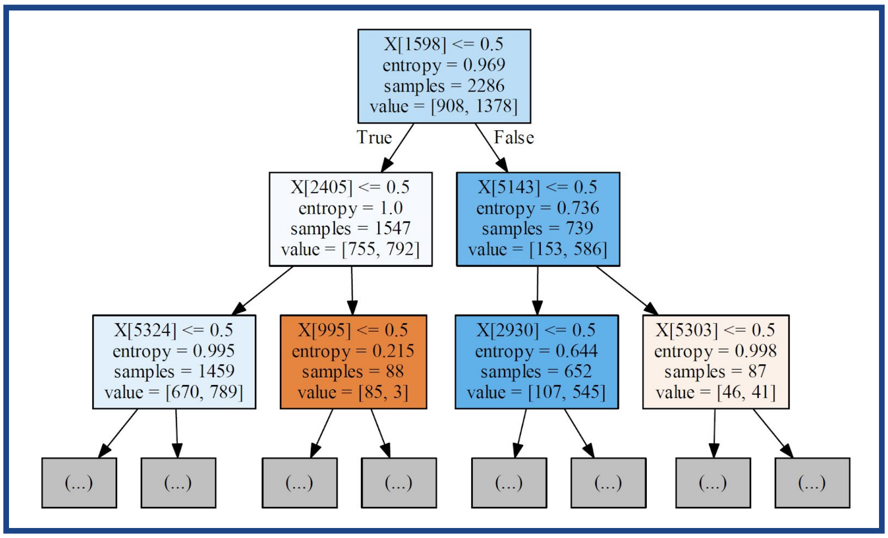
I used [3, 5, 7, 9, 11] as the input of different max depths for the decision tree, and the model with the highest accuracy on the validation set is the one with a max depth of 15 and having entropy as its criteria.

c) The accuracy of the model achieved the highest validation accuracy:

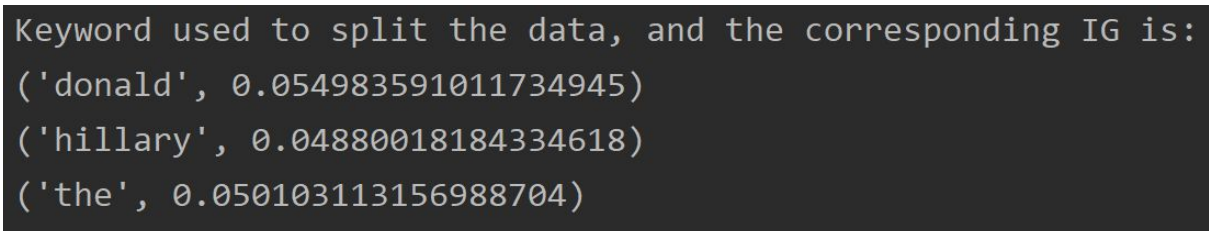


```
The accuracy of the best decision tree over test set is: 0.7285714285714285
```

The visualization of the first two layers of the decision tree:



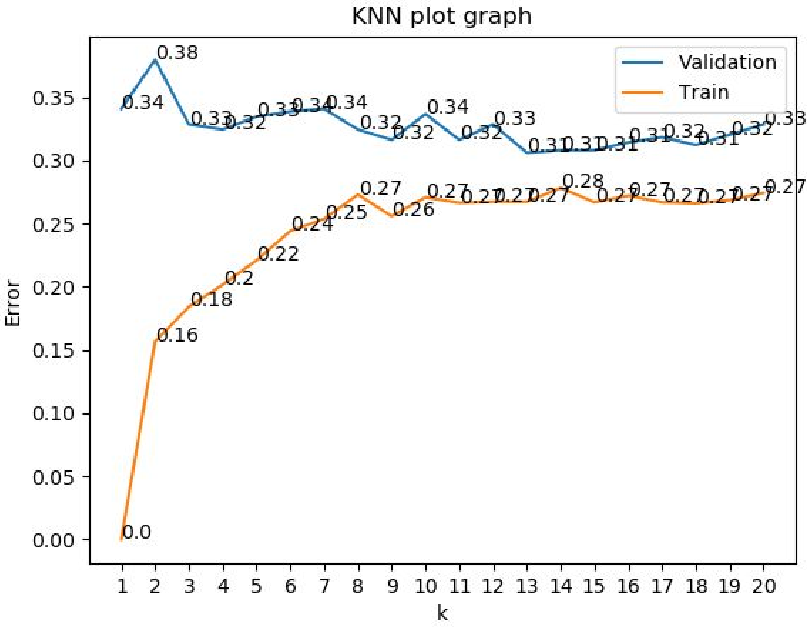
d) The information gain of a few keywords:



```
Keyword used to split the data, and the corresponding IG is:
('donald', 0.054983591011734945)
('hillary', 0.04880018184334618)
('the', 0.05010311315698870 4)
```

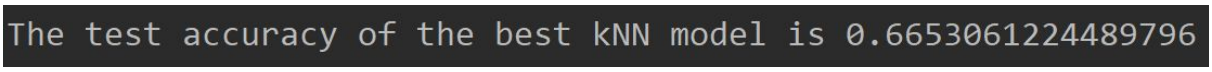
The keyword ‘donald’ is for the topmost split, and ‘hillary’ and ‘the’ corresponds to the split of the left child and right child of the root respectively. The float in each tuple indicates the information gain from the split with respect to the corresponding keyword.

e) The graph of errors:



From the graph, it is clear to see that the lowest validation error was achieved when k = 13.

The test accuracy of kNN when k = 13 is:



```
The test accuracy of the best kNN model is 0.6653061224489796
```