Question 1

(a) $E(X^2) = E(Y^2) = Var(X) + E(X)^2$

$$= \frac{(b-a)^2}{12} + \frac{b+a}{2}$$

$$= \frac{1}{12} + \frac{1}{2} = \frac{13}{12}$$

$$E(Z) = E(X^2 - 2XY + Y^2)$$

$$= E(X^2) - 2E(X)E(Y) + E(Y)^2$$

$$= \frac{13}{2} - 2 + \frac{13}{2} = \frac{1}{6}$$

$$Var(X^2) = Var(Y^2)$$

$$= E(X^4) - E(X^2)^2$$

By fourth moment of mgf

$$E(X^4) = \frac{9}{5}$$

$$= \frac{9}{5} - \frac{13}{12}^2 = \frac{451}{720}$$

$$Var(X) = \frac{(b-a)^2}{12} = \frac{1}{12}$$

$$Var(Z) = Var(|X-Y|^2)$$

$$= Var(X^2 - 2XY + Y^2)$$

$$= Var(X^2) + 4Var(XY) + Var(Y^2)$$

$\because$ X and Y are independent

$\therefore$ $Var(XY) = Var(X)Var(Y)$

$$= Var(X^2) + 4Var(X)Var(Y) + Var(Y^2)$$

$$= \frac{451}{720} + 4 \cdot \frac{1}{12} \cdot \frac{1}{12} + \frac{451}{720}$$

$$= \frac{461}{360}$$

b) $E[R] = E[Z_1 + \dots + Z_d]$

$\because$ since $X_1, Y_1 \dots X_d, Y_d \sim \text{Uniform}(0, 1$

$\therefore E(|X_1 - Y_1|^2) = \dots = E(|X_d - Y_d|^2)$

$\therefore E(Z_1) = \dots = E(Z_d)$

$= d\,E(Z)$

$= d \cdot \frac{1}{6} = \frac{d}{6}$

$Var[R] = Var[Z_1 + \dots + Z_d]$

$\because Var[Z_1] = \dots = Var[Z_d]$

$\therefore Var[Z_1 + \dots + Z_d] = Var(dZ_1)$

$= d^2\, Var[Z_1]$)

$= d^2 \frac{461}{720}$

d)

$$\sigma = \sqrt{d^2 \frac{461}{720}} = d\sqrt{\frac{461}{720}}$$

$$\mu = d \cdot \frac{1}{6}$$

$$\because \frac{1}{6} < \sqrt{\frac{461}{720}} \approx 0.8$$

$$\therefore \mu < \sigma$$

As dimension d increase, σ increase faster than μ.

maxium distance $= d\,E(z)$

Therefore as d increase,

average distance distance increase,

so every points get further away.

But they get relative same distance,

which seems to be closer.

Question 2

a) $H(X) = \sum_{x} p(x) \log_2\left(\frac{1}{p(x)}\right)$

$\because \ 0 \leq P(X) \leq 1$,

$\therefore \ \frac{1}{p(X)} \geq 1$

$\therefore \ \log_2\left(\frac{1}{p(x)}\right) \geq \log_2(1)$

$\Rightarrow \log_2\left(\frac{1}{p(x)}\right) \geq 0$

b) $H(X,Y) = H(X) + H(Y|X)$

by properties $H(Y|X) = H(Y)$

$H(X,Y) = H(X) + H(Y)$

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

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

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

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

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

d)

$$KL(p\|q) = \int p(x)\log_2\left(\frac{p(x)}{q(x)}\right)dx$$

$$= -\int p(x)\log_2\left(\frac{q(x)}{p(x)}\right)dx$$

$$= -E\left(\log_2\frac{q}{p}\right)$$

by Jensen's inequality

$$E\left(\log_2\frac{q}{p}\right) \geq \log_2\left(E\left(\frac{q}{p}\right)\right)$$

$$-E\left(\log_2\frac{q}{p}\right) < -\log_2\left(E\left(\frac{q}{p}\right)\right)$$

$\because$ Since

$$\therefore \quad -\log_2\left(\Sigma\, p(x)\,\frac{q(x)}{p(x)}\right) = 0$$

$$\therefore -E\left(\log_2\frac{q}{p}\right) < 0$$

$$\therefore E\left(\log_2\frac{q}{p}\right) \geq 0$$

e)

$$KL(p(x,y)\|p(x)p(y)) = \Sigma\, p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

$$= \Sigma\, p(x,y)\log_2\left(\frac{p(y)\,p(x|y)}{p(x)p(y)}\right)$$

$$= \Sigma\, p(x,y)\log_2\left(\frac{p(x|y)}{p(x)}\right)$$

$$= \Sigma\, p(x,y)\log_2\left(\frac{1}{p(x)}\right) - \Sigma\, p(x,y)\log_2\left(\frac{1}{p(x|y)}\right)$$

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

$$= I(X;Y)$$

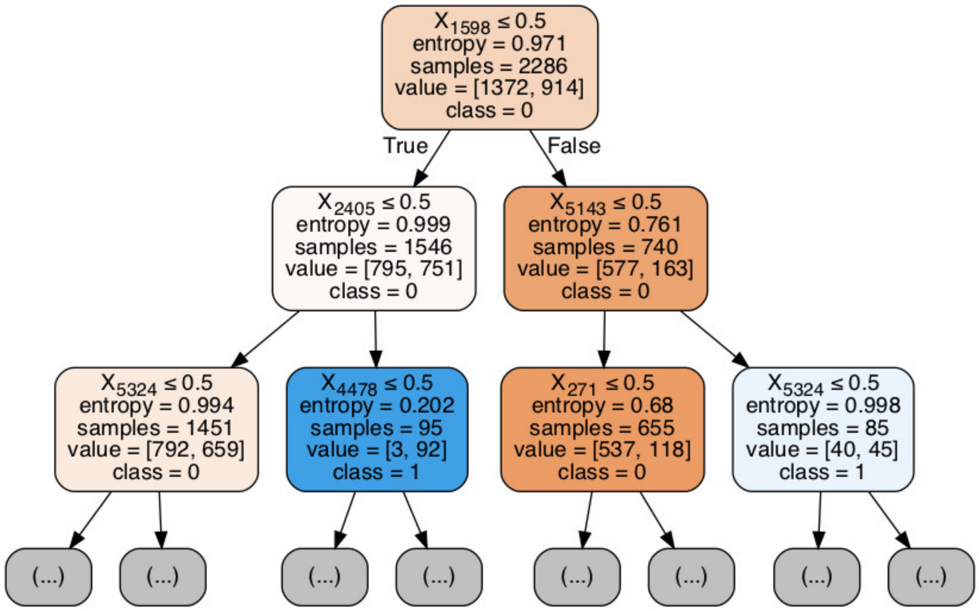
Question 3

b)

```
In [87]: models = [gini10, gini30, gini50, gini100, gini150, entro10, entro30, entro50, entro100, entro150]
         scores = []
         for model in models:
             score = model.score(x_validation_array, y_validation_array)
             scores.append(score)
             print(model.criterion, model.max_depth, 'Model Accuarcy:', score)
         best = max(scores)
         best_model = models[scores.index(best)]
         print('Best Model :', best_model)
         print("Best Model Accuracy ", test_accuracy(best_knn_model, x_test_array, y_test_array))
```

```
gini 10 Model Accuarcy: 0.7081632653061225
gini 30 Model Accuarcy: 0.7653061224489796
gini 50 Model Accuarcy: 0.7714285714285715
gini 100 Model Accuarcy: 0.753061224489796
gini 150 Model Accuarcy: 0.7714285714285715
entropy 10 Model Accuarcy: 0.7204081632653061
entropy 30 Model Accuarcy: 0.7612244897959184
entropy 50 Model Accuarcy: 0.7673469387755102
entropy 100 Model Accuarcy: 0.7775510204081633
entropy 150 Model Accuarcy: 0.7673469387755102
Best Model : DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=100,
                       max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort=False,
                       random_state=None, splitter='best')
Best Model Accuracy  0.5346938775510204
```

c)



In [83]:

```
models = [gini10, gini30, gini50, gini100, gini150, entro10, entro30, entro50, entro100, entro150]
scores = []
for model in models:
    score = model.score(x_validation_array, y_validation_array)
    scores.append(score)
best = max(scores)
best_model = models[scores.index(best)]
print(best_model)
print("Best Model Accuracy ", test_accuracy(best_model, x_test_array, y_test_array))
```

```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=100,
                       max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort=False,
                       random_state=None, splitter='best')
Best Model Accuracy  0.5102040816326531
```

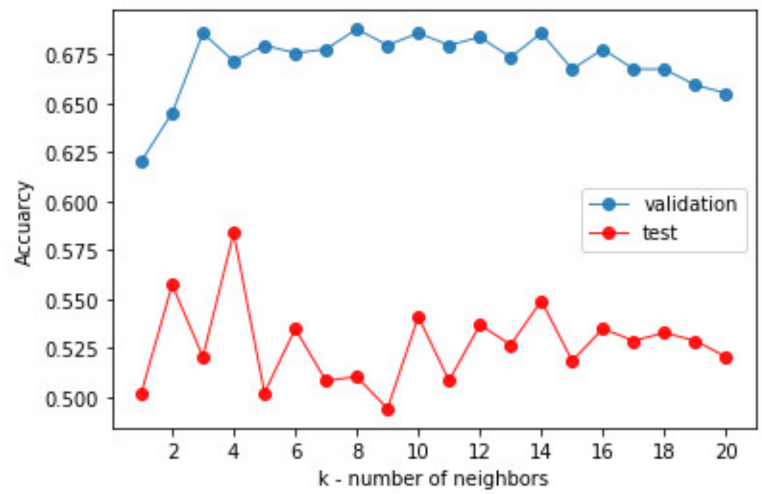
d)

```
In [213]: print(compute_info_gain(vectorizer.inverse_transform(x_train),y_train_array, 'the')) #Topmost
          print(compute_info_gain(vectorizer.inverse_transform(x_train),y_train_array, 'hillary'))
          print(compute_info_gain(vectorizer.inverse_transform(x_train),y_train_array, 'donald'))

0.05330262393898977
0.040765700073505995
0.0501023886302121

In [ ]:
```

e)




```
In [74]: print(best_knn_model)
         print("Best Model Accuracy ", test_accuracy(best_knn_model, x_test_array, y_test_array))

KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
                     metric_params=None, n_jobs=None, n_neighbors=17, p=2,
                     weights='uniform')
Best Model Accuracy  0.5285714285714286
```