Question 1

1.

a) $z = (x-y)^2 = x^2 - 2xy + y^2 \quad x, y \in [0, 1]$

Expectation:

$E(x) = \int_a^b x f(x) dx$

uniform distribution

$f(x) = \frac{1}{b-a}$

$$E(z) = E(x-y)^2 = E(x^2) - E(2xy) + E(y^2)$$
$$= E(x^2) - 2E(x) \cdot E(y) + E(y^2)$$
$$= \int_0^1 \frac{x^2}{1-0} dx - 2\int_0^1 x \cdot \frac{1}{1-0} dx \int_0^1 y \cdot \frac{1}{1-0} dy + \int_0^1 \frac{y^2}{1-0} dy$$
$$= \frac{x^3}{3}\Big|_0^1 - 2 \frac{x^2}{2}\Big|_0^1 \cdot \frac{y^2}{2}\Big|_0^1 + \frac{y^3}{3}\Big|_0^1$$
$$= \frac{1}{3} + \frac{1}{3} - \cancel{2} \cdot \frac{1}{2} \times \cancel{\frac{1}{2}}$$
$$= \frac{2}{3} - \frac{1}{2}$$
$$= \frac{1}{6}$$

Variance $\quad Var(x) = E(x^2) - E(x)^2$

$$Var(z) = E(z^2) - E(z)^2$$
$$E(z^2) = E((x-y)^4) = E(x^4 - 4x^3y + 6x^2y^2 - 4xy^3 + y^4)$$
$$= E(x^4) - 4E(x^3y) + 6E(x^2y^2) - 4E(xy^3) + E(y^4)$$
$$= \int_0^1 x^4 \frac{1}{1-0} dx - 4\int_0^1 x^3 \frac{1}{1-0} dx \int_0^1 y \frac{1}{1-0} dy + 6\int_0^1 x^2 \frac{1}{1-0} dx \int_0^1 y^2 \frac{1}{1-0} dy$$
$$- 4\int_0^1 y^3 \frac{1}{1-0} dy \cdot \int_0^1 x \frac{1}{1-0} dx + \int_0^1 y^4 \frac{1}{1-0} dy$$
$$= \frac{1}{5} - \cancel{4} \cdot \frac{1}{\cancel{4}} \cdot \frac{1}{2} + \cancel{6}^{2} \frac{1}{\cancel{3}} \cdot \frac{1}{3} - \cancel{4} \cdot \frac{1}{\cancel{4}} \cdot \frac{1}{2} + \frac{1}{5}$$
$$= \frac{2}{5} + \frac{2}{3} - 1 = \frac{2}{5} - \frac{1}{3} = \frac{1}{15}$$

$$\therefore Var(z) = E(z)^2 - E(z)^2 = \frac{1}{15} - \frac{1}{6} \cdot \frac{1}{6}$$
$$= \frac{1}{15} - \frac{1}{36}$$
$$= \frac{7}{180}$$

b>

$$R = Z_1 + \cdots Z_d = \sum_{i=1}^{d} Z_i$$

$$Z_i = (X_i - Y_i)^2$$

from a) $E(Z) = \frac{1}{6}$

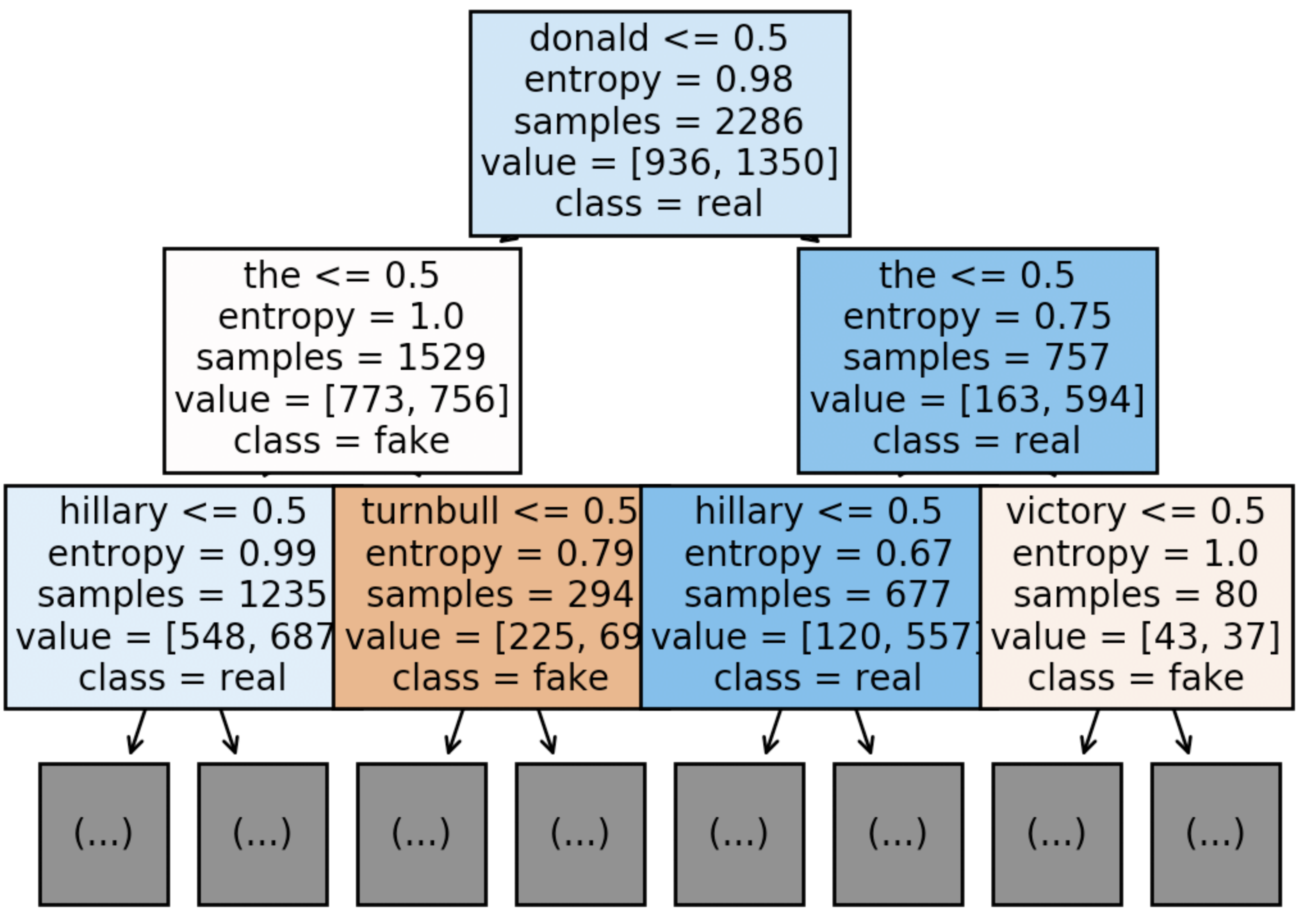
$\mathrm{Var}(Z) = \frac{7}{180}$

$$E(R) = E\left(\sum_{i=1}^{d} Z_i\right) = \sum_{i=1}^{d} E(Z_i) = d \cdot \frac{1}{6} = \frac{d}{6}$$

$$\mathrm{Var}(R) = \mathrm{Var}\left(\sum_{i=1}^{d} Z_i\right) = \sum_{i=1}^{d} V(Z_i) = \frac{7d}{180}$$

Question 2:

```
criteria: gini max_depth: 24 accuracy: 0.789795918367347
criteria: gini max_depth: 13 accuracy: 0.7816326530612245
criteria: gini max_depth: 9 accuracy: 0.7591836734693878
criteria: gini max_depth: 7 accuracy: 0.7346938775510204
criteria: gini max_depth: 2 accuracy: 0.6816326530612244
criteria: entropy max_depth: 24 accuracy: 0.7938775510204081
criteria: entropy max_depth: 13 accuracy: 0.7510204081632653
criteria: entropy max_depth: 9 accuracy: 0.7428571428571429
criteria: entropy max_depth: 7 accuracy: 0.7204081632653061
criteria: entropy max_depth: 2 accuracy: 0.6816326530612244
the best situation is: criteria: entropy max_depth: 24 accuracy: 0.7938775510204081
```



3. a) $H(x) = \sum_x p(x) \log_2 \left(\frac{1}{p(x)}\right)$ $\quad x \in \{1, 2, \ldots N\}$

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

since $p(x) \in [0, 1]$

$\therefore \log_2 p(x) \leq 0$

$\therefore p(x) \cdot \log_2 p(x) \leq 0$

$\therefore -p(x) \cdot \log_2 p(x) \geq 0$

$\therefore \sum_{x \in X} (-) p(x) \cdot \log_2 p(x) \geq 0$

$\therefore H(x) \geq 0$ is non negative

b) $KL(p \| q) = \sum_x p(x) \cdot \log_2 \frac{p(x)}{q(x)}$

assume $p(x) > 0$ and $q(x) > 0$ for all x

according to Jensen's inequality

$$E(f(x)) \geq f(E(x))$$

logarithm $(\log(x))$ is concave on the set of positive real numbers

$\therefore -\log(x)$ is convex

$\therefore KL(p \| q) = \sum_x p(x) \left(-\log_2 \frac{p(x)}{q(x)}\right)$ is convex

$\therefore -KL(p \| q) = \sum_x p(x) \log_2 \frac{q(x)}{p(x)}$

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

$$= \log_2 \sum_x p(x) \cdot \frac{q(x)}{p(x)}$$

$$= \log_2 \sum_x q(x) \geq 0$$

$\therefore KL(p \| q) \geq 0$

c) $I(Y;X) = H(Y) - H(Y|X)$

$$H(Y|X) = \sum_{x \in X} P(x) \cdot H(Y|X=x)$$

$$= \sum_{x \in X} P(x)(-)\sum_{y \in Y} P(Y|X) \cdot \log P(Y|X)$$

$$= -\sum_{x \in X} P(x) \cdot \sum_{y \in Y} P(Y|X) \cdot \log P(Y|X)$$

$$= -\sum_{x,y \in X,Y} P(x,y) \cdot \log P(Y|X)$$

$\Rightarrow \therefore I(Y;X) = H(Y) - H(Y|X)$

$$= -\sum_y P(y) \cdot \log_2 P(y) + \sum_{x,y \in X,Y} P(x,y) \cdot \log P(Y|X)$$

$$= -\sum_{x,y} P(x,y) \cdot \log_2 P(y) + \sum P(x,y) \log P(Y|X)$$

$$= \sum_{x,y} P(x,y) \cdot \log_2 \frac{P(Y|X)}{P(y)}$$

$$= \sum_{x,y} P(x,y) \cdot \log_2 \frac{P(Y|X) \cdot P(x)}{P(x) \cdot P(y)} = \sum_{x,y} P(x,y) \cdot \log_2 \frac{P(x,y)}{P(x) \cdot P(y)}$$

$$= KL(P(x,y) \| P(x)P(y))$$