Q1

1. There are 8 variables in the file, all of them are nominal variables. The variables that appear to be important are L-BP and CORE-STBL because if we know a patient is either low in L-BP or mod-stable in CORE-STBL, we can tell he/she will remain in hospital straight away.

2. (a) J48

Variation 1: binarySplits = False, minNumObj = 2, unpruned = False

- i) There is only 1 node in the tree, therefore all instance are predicted to be class A (staying in the hospital). By setting unpruned = False, Weka perform an additional step that looks at what nodes can be removed without affecting the performance too much.
- There is no main variable in the decision tree as the classifier predict all instance as class A.

- Ii) The accuracy is 70.93%. From the confusion matrix, there are 61 correct predictions (as remaining in the hospital), 1 that is still in the hospital but classified as discharged and 24 discharged patient but predicted as remaining in the hospital.

```
J48 pruned tree
------------------
: A (86.0/24.0)
```

```
=== Confusion Matrix ===

  a  b   <-- classified as
 61  1 |  a = A
 24  0 |  b = S
```

Variation 2: binarySplits = True, minNumObj = 1, unpruned = True

- i) The tree is much larger in this case since it is not pruned. The variable that provides the largest amount of information gain is L-BP as expected. Since binarySplits is set to True, there are two nodes (True and False) at each level of the tree.
- The main variables are: L-BP, CORE-STBL, BP-STBL, L-SURF, L-CORE,SURF-STBL, L-O2 and decision. All the variables are important.

- Ii) The accuracy is 59.30% in this case. From the confusion matrix, there are 49 correct predictions for class A and 2 correct predictions of class S, 22 patients that are discharged but predicted as remaining in the hospital. 13 patients that are still in the hospital but classified as discharged.

```
J48 unpruned tree
------------------

L-BP = low: A (3.0)
L-BP != low
|   CORE-STBL = mod-stable: A (1.0)
|   CORE-STBL != mod-stable
|   |   CORE-STBL = stable
|   |   |   L-BP = high
|   |   |   |   L-SURF = high
|   |   |   |   |   BP-STBL = stable: S (2.0)
|   |   |   |   |   BP-STBL != stable: A (3.0)
|   |   |   |   L-SURF != high
|   |   |   |   |   L-CORE = high: A (2.0)
|   |   |   |   |   L-CORE != high
|   |   |   |   |   |   BP-STBL = unstable
|   |   |   |   |   |   |   L-CORE = low: S (1.0)
|   |   |   |   |   |   |   L-CORE != low: A (1.0)
|   |   |   |   |   |   BP-STBL != unstable: A (17.0/2.0)
|   |   |   L-BP != high
|   |   |   |   L-SURF = high
|   |   |   |   |   SURF-STBL = stable
|   |   |   |   |   |   BP-STBL = mod-stable: A (2.0)
|   |   |   |   |   |   BP-STBL != mod-stable: S (1.0)
|   |   |   |   |   SURF-STBL != stable: A (6.0)
|   |   |   |   L-SURF != high
|   |   |   |   |   BP-STBL = mod-stable: S (4.0/1.0)
|   |   |   |   |   BP-STBL != mod-stable
|   |   |   |   |   |   L-CORE = mid
|   |   |   |   |   |   |   BP-STBL = stable: A (18.0/4.0)
|   |   |   |   |   |   |   BP-STBL != stable
|   |   |   |   |   |   |   |   L-SURF = low: S (1.0)
|   |   |   |   |   |   |   |   L-SURF != low: A (6.0/1.0)
|   |   |   |   |   |   L-CORE != mid
|   |   |   |   |   |   |   L-O2 = excellent
|   |   |   |   |   |   |   |   L-SURF = low: A (2.0)
|   |   |   |   |   |   |   |   L-SURF != low
|   |   |   |   |   |   |   |   |   L-CORE = high: A (1.0)
|   |   |   |   |   |   |   |   |   L-CORE != high: S (1.0)
|   |   |   |   |   |   |   L-O2 != excellent
|   |   |   |   |   |   |   |   L-CORE = high: S (1.0)
|   |   |   |   |   |   |   |   L-CORE != high
|   |   |   |   |   |   |   |   |   SURF-STBL = stable
|   |   |   |   |   |   |   |   |   |   BP-STBL = stable: S (2.0)
|   |   |   |   |   |   |   |   |   |   BP-STBL != stable: A (1.0)
|   |   |   |   |   |   |   |   |   SURF-STBL != stable: A (5.0/2.0)
|   |   CORE-STBL != stable
|   |   |   BP-STBL = stable: S (3.0)
|   |   |   BP-STBL != stable: A (2.0)

Number of Leaves  :     24

Size of the tree :      47




=== Confusion Matrix ===

  a  b   <-- classified as
 49 13 |  a = A
 22  2 |  b = S
```

(b) Naïve Bayes

```
BP-STBL
  mod-stable        18.0    5.0
  stable            31.0   16.0
  unstable          16.0    6.0
  [total]           65.0   27.0
```

i) A discrete distribution describes the probability of occurrence of each value of a discrete random variable. A discrete random variable is a random variable that has countable values, such as a list of non-negative integers.

BP-STBL probability distribution		
mod-stable	stable	unstable
(17+4)/(62+24) = 0.2442	(30+15)/(62+24) = 0.5233	(15+5)/(62+24) = 0.2326

ii)

P(C = S | L-CORE = mid, L-SURF = low, L-O2 = good, L-BP = high, SURF-STBL = stable, CORE-STBL = stable, BP-STBL = mod-stable)

= α x P(L-CORE = mid | C=S) x P(L-SURF = low | C=S) x P(L-O2 = good | C=S) x P(L-BP = high | C=S) x P(SURF-STBL = stable | C=S) x P(CORE-STBL = stable | C=S) x P(BP-STBL = mod-stable | C=S) x P(C = S)

$= \alpha \times \frac{15}{24} \times \frac{7}{24} \times \frac{13}{24} \times \frac{6}{24} \times \frac{12}{24} \times \frac{21}{24} \times \frac{4}{24} \times \frac{24}{24+62}$

= 0.000502α

P(C = A | L-CORE = mid, L-SURF = low, L-O2 = good, L-BP = high, SURF-STBL = stable, CORE-STBL = stable, BP-STBL = mod-stable)

= α x P(L-CORE = mid | C=A) x P(L-SURF = low | C=A) x P(L-O2 = good | C=A) x P(L-BP = high | C=A) x P(SURF-STBL = stable | C=A) x P(CORE-STBL = stable | C=A) x P(BP-STBL = mod-stable | C=A) x P(C = A)

$= \alpha \times \frac{41}{62} \times \frac{16}{62} \times \frac{32}{62} \times \frac{22}{62} \times \frac{31}{62} \times \frac{59}{62} \times \frac{17}{62} \times \frac{62}{24+62}$

= 0.00294α

The two probabilities calculated above must sum to 1.

0.000502 α +0.00294 α = 1

α = 290.53

P(C = S | L-CORE = mid, L-SURF = low, L-O2 = good, L-BP = high, SURF-STBL = stable, CORE-STBL = stable, BP-STBL = mod-stable)
= 290.53 x 0.000502
= 14.58%

P(C = A | L-CORE = mid, L-SURF = low, L-O2 = good, L-BP = high, SURF-STBL = stable, CORE-STBL = stable, BP-STBL = mod-stable)
= 290.53 x 0.00294
= 85.42%

The probability this patient will be discharged is 14.58%.

iii) Since P(C = A | L-CORE = mid, L-SURF = low, L-O2 = good, L-BP = high, SURF-STBL = stable, CORE-STBL = stable, BP-STBL = mod-stable) = 85.42% is greater than P(C = S | L-CORE = mid, L-SURF = low, L-O2 = good, L-BP = high, SURF-STBL = stable, CORE-STBL = stable, BP-STBL = mod-stable) = 14.58%, therefore, the Naïve Bayes classifier will classified this patient as class A (remain in hospital).

iv) The accuracy of the Naïve Bayes classifier is 69.77%. From the confusion matrix, there are 58 instances correctly predicted as class A, 2 instances correctly predicted as class S. 4 instances that are class A but incorrectly predicted as class S and 22 instances that are class S but incorrectly predicted as class A.

```
Correctly Classified Instances          60               69.7674 %
Incorrectly Classified Instances        26               30.2326 %
Kappa statistic                          0.0244
Mean absolute error                      0.415
Root mean squared error                  0.4732
Relative absolute error                102.4041 %
Root relative squared error            105.3522 %
Total Number of Instances               86

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.935    0.917    0.725      0.935    0.817      0.033    0.437     0.654     A
                 0.083    0.065    0.333      0.083    0.133      0.033    0.437     0.304     S
Weighted Avg.    0.698    0.679    0.616      0.698    0.626      0.033    0.437     0.556

=== Confusion Matrix ===

  a  b   <-- classified as
 58  4 |  a = A
 22  2 |  b = S
```

(c)

i) These 3 instances has the highest Jaccard coefficient with the above patient.

	L-CORE	L-SURF	L-O2	L-BP	SURF-STBL	CORE-STBL	BP-STBL	CLASS
Given	mid	low	good	high	Stable	stable	Mod-stable	-
X1	mid	low	good	high	Stable	Unstable	Mod-stable	A
X2	mid	low	good	high	Stable	stable	Mod-stable	A
X3	mid	low	good	high	unstable	stable	Mod-stable	A

JC(Given,X1) = $\frac{mid,low,good,high,stable,mod-stable}{mid,low,good,high,stable,stable,unstable,mod-stable}$ = 6/8 = 0.75

JC(Given,X2) = $\frac{mid,low,good,high,stable,stable,mod-stable}{mid,low,good,high,stable,stable,mod-stable}$ = 7/7 = 1

JC(Given,X3) =$\frac{mid,low,good,high,stable,mod-stable}{mid,low,good,high,stable,unstable,stable,mod-stable}$ = 6/8 = 0.75

When k=1, this patient will be predicted as class A (remaining in hospital), since the closest neighbor are class A.

When k=3, the patient will be predicted as class A (remaining in hospital), since the 3 closest neighbor are class A.

ii) Variation 1: **<u>K = 1, distanceWeighting = weight by 1 - distance</u>**

The accuracy is 61.63%. From the confusion matrix 53 instances are correctly classified as class A, 9 instances are misclassified as class S and 24 instances are misclassified as class A.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          53               61.6279 %
Incorrectly Classified Instances        33               38.3721 %
Kappa statistic                         -0.1796
Mean absolute error                      0.45
Root mean squared error                  0.596
Relative absolute error                111.0397 %
Root relative squared error            132.6971 %
Total Number of Instances               86

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.855    1.000    0.688      0.855    0.763      -0.213   0.308     0.615     A
                 0.000    0.145    0.000      0.000    0.000      -0.213   0.310     0.211     S
Weighted Avg.    0.616    0.761    0.496      0.616    0.550      -0.213   0.309     0.503

=== Confusion Matrix ===

  a  b   <-- classified as
 53  9 |  a = A
 24  0 |  b = S
```

Variation 2: <u>K = 3, distanceWeighting = No distance weighting</u>

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          60               69.7674 %
Incorrectly Classified Instances        26               30.2326 %
Kappa statistic                         -0.0449
Mean absolute error                      0.4388
Root mean squared error                  0.5133
Relative absolute error                108.2735 %
Root relative squared error            114.2811 %
Total Number of Instances               86

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.968    1.000    0.714      0.968    0.822      -0.096   0.284     0.595     A
                 0.000    0.032    0.000      0.000    0.000      -0.096   0.284     0.211     S
Weighted Avg.    0.698    0.730    0.515      0.698    0.593      -0.096   0.284     0.488

=== Confusion Matrix ===

  a  b   <-- classified as
 60  2 |  a = A
 24  0 |  b = S
```

The accuracy is 69.77%. From the confusion matrix, 60 instances are correctly classified as class A, 2 instances are misclassified as class S and 24 instances are misclassified as class A.

3.

Classifier	**Accuracy**	**Kappa statistic**	**MAE**	**RAE**	**RMSE**	**ROC area**
J48(with pruning)	70.93%	-0.0228	0.4075	100.5513%	0.4554	0.439
Naïve-Bayes	69.77%	0.0244	0.415	102.4041%	0.4732	0.437
k-NN (k = 3)	69.77%	-0.0449	0.4388	108.2735%	0.5133	0.284

The Kappa statistic, MAE and RMSE, ROC all measures how the predictions are close to the actual outcomes, the decision tree classifier performs slightly better in terms of all measures.

J48 performs better than the other 2 classifiers because it simply predict all patient as class A (remaining in the hospital) and the majority of the patients in the training set remain in the hospital.

Q2

1.

(a) The middle-middle square seems to discriminate between classes, therefore it has the most predictive of winning or losing. By placing the first cross at the centre, the player has the highest chance to place three of the marks in a horizontal, vertical, or diagonal row.

(b) By being the first player, one can choose to place his/her mark at the centre.

2.

(i) The main variables are: middle-middle, top-left, bottom-right, top- middle, top-right, bottom-middle, bottom-left, middle-right, middle-left. All variables are important.

(ii) The prediction is negative, which means the classifier predict X player will lose the game.

(iii) The first split is the middle-middle square.

H(S) = $-P\log_2(P) - N\log_2(N)$

$= -\frac{626}{958} \times log_2\left(\frac{626}{958}\right) - \frac{332}{958} \times log_2\left(\frac{332}{958}\right)$

= 0.93095

H(middle-middle = b) = $-\frac{48}{160} \times log_2\left(\frac{48}{160}\right) - \frac{112}{160} \times log_2\left(\frac{112}{160}\right)$ = 0.8813

H(middle-middle = o) = $-\frac{192}{340} \times log_2\left(\frac{192}{340}\right) - \frac{148}{340} \times log_2\left(\frac{148}{340}\right)$ = 0.9879

H(middle-middle = x) = $-\frac{92}{458} \times log_2\left(\frac{92}{458}\right) - \frac{366}{458} \times log_2\left(\frac{366}{458}\right)$ = 0.7236

Information gain

= H(S) – 160/958 * H(middle-middle = b) – 340/958 * H(middle-middle = o) – 458/958 * H(middle-middle = x)

= 0.93095 – 160/958*0.8813 – 340/958*0.9879 – 458/958*0.7236

= 0.0872

(iv)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         810               84.5511 %
Incorrectly Classified Instances       148               15.4489 %
Kappa statistic                          0.6574
Mean absolute error                      0.1695
Root mean squared error                  0.3439
Relative absolute error                 37.4049 %
Root relative squared error             72.262  %
Total Number of Instances              958

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.768    0.113    0.782      0.768    0.775      0.658    0.896     0.835     negative
                 0.887    0.232    0.878      0.887    0.882      0.658    0.896     0.923     positive
Weighted Avg.    0.846    0.191    0.845      0.846    0.845      0.658    0.896     0.892

=== Confusion Matrix ===

   a   b   <-- classified as
 255  77 |   a = negative
  71 555 |   b = positive
```

The accuracy of the decision tree is 84.55%. From the confusion matrix, the model correctly predicted 255 negative cases and 555 positive cases. It misclassified 77 negative cases as positive and misclassified 71 positive cases as negative.

(b)

i) P(C = positive |top-left = x, top-middle = x, top-right = b, middle-left = o, middle-middle = o, middle-right = x, bottom-left = b, bottom-middle = o, bottom-right = b)

= $\alpha\ \times$ P(top-left = x | C = positive) x P(top-middle = x | C = positive) x P(top-right = b | C = positive) x P(middle-left = o | C = positive) x P(middle-middle = o, | C = positive) x P(middle-right = x | C = positive) x P(bottom-left = b | C = positive) x P(bottom-middle = o | C = positive) x P(bottom-right = b | C = positive) x P(C = positive)

$$= \alpha \times \frac{295}{626} \times \frac{225}{626} \times \frac{142}{626} \times \frac{229}{626} \times \frac{148}{626} \times \frac{225}{626} \times \frac{142}{626} \times \frac{229}{626} \times \frac{142}{626} \times \frac{626}{958}$$

= 0.00001469α

P(C = negative |top-left = x, top-middle = x, top-right = b, middle-left = o, middle-middle = o, middle-right = x, bottom-left = b, bottom-middle = o, bottom-right = b)

= $\alpha\ \times$ P(top-left = x | C = negative) x P(top-middle = x | C = negative) x P(top-right = b | C = negative) x P(middle-left = o | C = negative) x P(middle-middle = o, | C = negative) x P(middle-right = x | C = negative) x P(bottom-left = b | C = negative) x

P(bottom-middle = o | C = negative) x P(bottom-right = b | C = negative) x P(C = negative)

$= \alpha \times \frac{123}{332} \times \frac{153}{332} \times \frac{63}{332} \times \frac{101}{332} \times \frac{192}{332} \times \frac{153}{332} \times \frac{63}{332} \times \frac{101}{332} \times \frac{63}{332} \times \frac{332}{958}$

= 0.000009972α

The two probabilities calculated above must sum to 1

0.00001469α + 0.000009972α = 1

α = 40548.22

P(C = positive |top-left = x, top-middle = x, top-right = b, middle-left = o, middle-middle = o, middle-right = x, bottom-left = b, bottom-middle = o, bottom-right = b)
= 40548.22 x 0.00001469
= 59.57%

P(C = negative | top-left = x, top-middle = x, top-right = b, middle-left = o, middle-middle = o, middle-right = x, bottom-left = b, bottom-middle = o, bottom-right = b)
= 40548.22 x 0.000009972
= 40.43%

ii)
The probability is 59.57% as calculated above. The probability for C = positive given the condition is higher, therefore the classifier will predict the X player will win.

iii)

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         667               69.6242 %
Incorrectly Classified Instances       291               30.3758 %
Kappa statistic                          0.2843
Mean absolute error                      0.3702
Root mean squared error                  0.4319
Relative absolute error                 81.7182 %
Root relative squared error             90.7504 %
Total Number of Instances              958

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.428    0.161    0.584      0.428    0.494      0.291    0.744     0.635     negative
                 0.839    0.572    0.734      0.839    0.783      0.291    0.744     0.857     positive
Weighted Avg.    0.696    0.430    0.682      0.696    0.683      0.291    0.744     0.780

=== Confusion Matrix ===

   a   b   <-- classified as
 142 190 |   a = negative
 101 525 |   b = positive
```

The accuracy is 69.62%. From the confusion matrix, the classifier correctly predicted 142 negative case and 525 positive case. However, there are 190 negative cases that are misclassified as positive and 101 positive cases that are misclassified as negative.

(c)

(i)

	T_L	T_M	T_R	M_L	M_M	M_R	B_L	B_M	B_R	CLASS
Given	x	x	b	o	o	x	b	o	b	-
X1	x	x	x	o	o	x	b	o	b	Positive
X2	x	o	b	x	o	x	b	o	b	Negative
X3	x	x	b	o	o	o	b	x	b	Negative

Jc(Given, X1) = $\frac{x,x,o,o,x,b,o,b}{x,x,b,x,o,o,x,b,o,b}$ = 8/10 = 0.8

Jc(Given, X2) = $\frac{x,b,o,x,b,o,b}{x,x,o,b,o,x,o,x,b,o,b}$ = 7/11 = 0.6364

Jc(Given, X3) = $\frac{x,x,b,o,o,b,b}{x,x,b,o,o,x,o,b,o,x,b}$ = 7/11 = 0.6364

When k = 1, the classifier will predict this instance as positive since the X1 has the highest Jaccard Coefficient with it. When k =3, the classifier will predict this instance as negative since two of the three closest neighbor are negative.

(ii)

K=1

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         948               98.9562 %
Incorrectly Classified Instances        10                1.0438 %
Kappa statistic                          0.9768
Mean absolute error                      0.1909
Root mean squared error                  0.2315
Relative absolute error                 42.1333 %
Root relative squared error             48.6378 %
Total Number of Instances              958

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.973    0.002    0.997      0.973    0.985      0.977    1.000     0.999     negative
                 0.998    0.027    0.986      0.998    0.992      0.977    1.000     1.000     positive
Weighted Avg.    0.990    0.018    0.990      0.990    0.990      0.977    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 323   9 |   a = negative
   1 625 |   b = positive
```

The accuracy is 98.96%. From the confusion matrix, 323 instances are correctly classified as negative and 625 instances are correctly classified as positive. 9 instances that are negative are misclassified as positive and 1 instances that is positive but misclassified as negative.

K = 11

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         946               98.7474 %
Incorrectly Classified Instances        12                1.2526 %
Kappa statistic                          0.9721
Mean absolute error                      0.1927
Root mean squared error                  0.2337
Relative absolute error                 42.5461 %
Root relative squared error             49.1143 %
Total Number of Instances              958

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.967    0.002    0.997      0.967    0.982      0.972    1.000     0.999     negative
                 0.998    0.033    0.983      0.998    0.990      0.972    1.000     1.000     positive
Weighted Avg.    0.987    0.022    0.988      0.987    0.987      0.972    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 321  11 |   a = negative
   1 625 |   b = positive
```

The accuracy is 98.75%. The performance deteriorated when n increases. From the confusion matrix, 321 instances are correctly classified as negative and 625 instances are correctly classified as positive. 11 instances that are negative are misclassified as positive and 1 instances that is positive but misclassified as negative.

3.

Classifier	**Accuracy**	**Kappa statistic**	**MAE**	**RAE**	**RMSE**	**ROC area**
J48	84.55%	0.6574	0.1695	37.4049%	0.3439	0.896
Naïve-Bayes	69.62%	0.2843	0.3702	81.7182%	0.4319	0.744
k-NN (k = 3)	98.96%	0.9768	0.1909	42.1333%	0.2315	1

The Kappa statistic, MAE and RMSE, ROC all measures how the predictions are close to the actual outcomes, the k-NN classifier performs significantly better than the other 2 classifiers. The k-NN with a small k did a good job because if we know the outcome of a particular game (with no blank grid), and we observe the same game or similar game in the future, the outcome is very likely to be the same since this game is governed by the same rules.

The performance of Naïve Bayes is the worst because it assumes conditional independence, however in the game of tic tac toe, every move is dependent on previous moves.

Q3

1. Unemployment-Rate = -0.0014 * All-Ords-Index - 0.2452 * Housing-Loan-Interest-Rate + 13.7286

2. True value = 8.4
   Predicted value = -0.0014 * 1779.1 -0.2452 * 15.5 + 13.7286 = 7.43726
   Absolute Error = |True value – Predicted value| = 0.96274

3. MAE = 1.026189
   RMSE = 1.281314

   Both MAE and RMSE express average model prediction error in units of the variable of interest. Both metrics can range from 0 to ∞ and are indifferent to the direction of errors. They are negatively-oriented scores, which means lower values are better.

   Taking the square root of the average squared errors has some interesting implications for RMSE. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE should be more useful when large errors are particularly undesirable.