Privacy for IoT TD/TP

Submitted By:

1. Sheikh Shah RAHMAN
2. Bogdan GORELKIN

TD_1:

1.1 Pseudonymization

1. Where was **Alice** likely born and what is her most likely marital status?
    a. Female, 19-25, **Saarbrücken**, Game of Thrones, **single**
2. Can you also get personal information about **Charlie**?
    a. **Male**, 19-25, Saarbrücken, Game of Thrones, **single**
    b. **Male**, 16-18, Trier, Game of Thrones, **single**
    c. **Male**, 19-25, Berlin, Big Bang Theory, **single**
3. Can you also learn some personal information about Bob?
    a. Male, 12-15, München, Friends!, in relationship

1.2 k-anonymity

1. Does the dataset 1 in Figure 1.1 satisfy the k-anonymity? If so, what is the maximum value of k?
    a. yes, we always have 2 similar records. Max k =3 (male, GOT, age 19-25)
2. Same question for datasets 2
    a. k-anonimity is not satisfied because all its records are unique (female, Grey'A,19-25)
3. Same question for datasets 3
    b. not satisfied because it contains 1 unique record

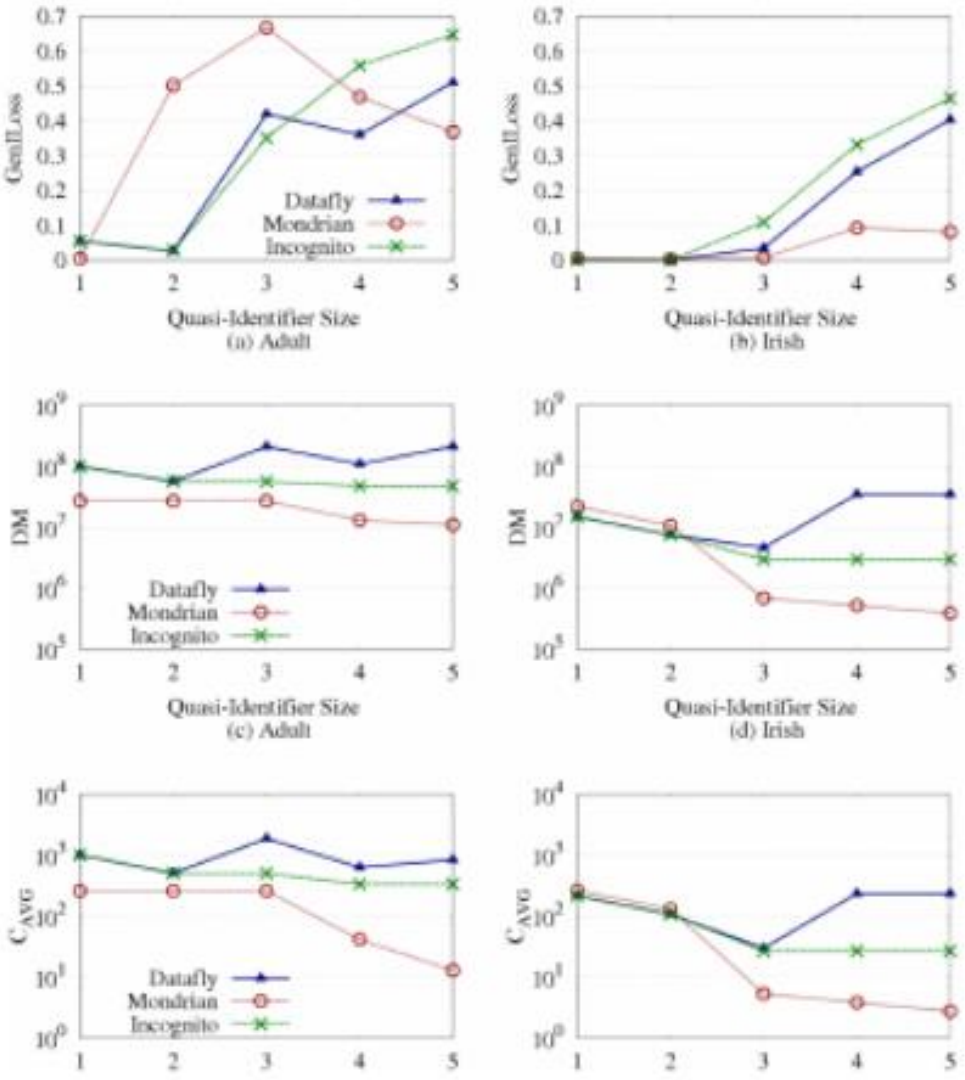
Exercise 1.3:

1. Propose a 3-anonymous version considering the following possible generalizations:

Id	MARITAL STATUS	AGE	ZIP
1	*	[25,29]	3204*
2	*	[20,24]	3202*
3	*	[20,24]	3202*
4	*	[25,29]	3204*
5	*	[25,29]	3204*
6	*	[20,24]	3202*

for any combination of these three attributes there are always 3 identical entries.

2. Propose a 3-anonymous version considering Mondrian's algorithm.
    a.
3. Calculate the values of C_{AVG}
    b. C_{AVG} = (6/2)*(1/3)=1; discernibility = 3^2 + 3^2 = 18
4. Figure 1.2 compares the usefulness of different k -anonymization methods on two different datasets. What can we conclude from these experiments?



    a.

TP_1:

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure
                 0.526    0.240    0.540      0.526    0.533
                 0.760    0.474    0.750      0.760    0.755
Weighted Avg.    0.678    0.392    0.676      0.678    0.677

=== Confusion Matrix ===

   a   b   <-- classified as
 141 127 |   a = 1
 120 380 |   b = 0
```

For 02:

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure
                 0.739    0.362    0.522      0.739    0.612
                 0.638    0.261    0.820      0.638    0.718
Weighted Avg.    0.673    0.296    0.716      0.673    0.681

=== Confusion Matrix ===

   a   b   <-- classified as
 198  70 |   a = 1
 181 319 |   b = 0
```

For 03:

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure
                 0.571    0.198    0.607      0.571    0.588
                 0.802    0.429    0.777      0.802    0.789
Weighted Avg.    0.721    0.348    0.718      0.721    0.719

=== Confusion Matrix ===

   a   b   <-- classified as
 153 115 |   a = 1
  99 401 |   b = 0
```

For 11 and 15 F-measure is 0.751 and 0.752

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure
                 0.616    0.174    0.655      0.616    0.635
                 0.826    0.384    0.800      0.826    0.813
Weighted Avg.    0.753    0.311    0.750      0.753    0.751

=== Confusion Matrix ===

   a   b   <-- classified as
 165 103 |   a = 1
  87 413 |   b = 0
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure
                 0.601    0.162    0.665      0.601    0.631
                 0.838    0.399    0.797      0.838    0.817
Weighted Avg.    0.755    0.316    0.751      0.755    0.752

=== Confusion Matrix ===

   a   b   <-- classified as
 161 107 |   a = 1
  81 419 |   b = 0
```

Bayesian Network

```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall   F-Measure
                 0.608     0.184     0.639       0.608    0.623
                 0.816     0.392     0.795       0.816    0.806
Weighted Avg.    0.743     0.319     0.741       0.743    0.742

=== Confusion Matrix ===

   a   b   <-- classified as
 163 105 |   a = 1
  92 408 |   b = 0
```

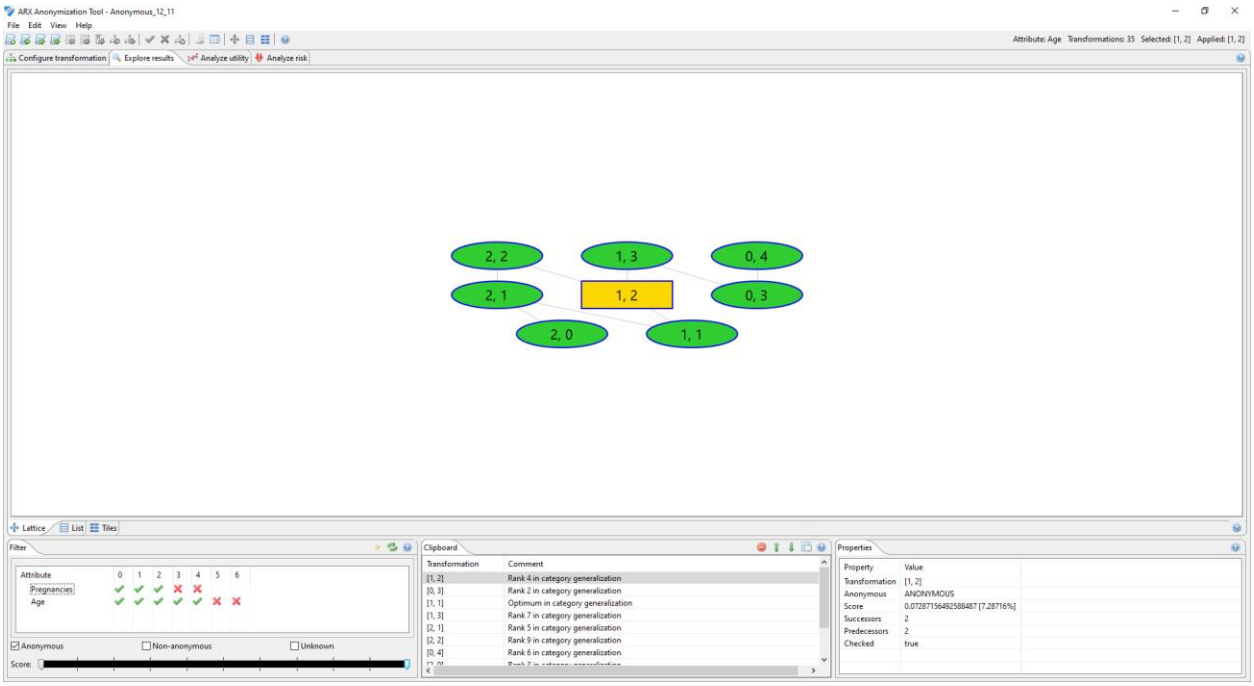


Figure 1 – Generalization lattice

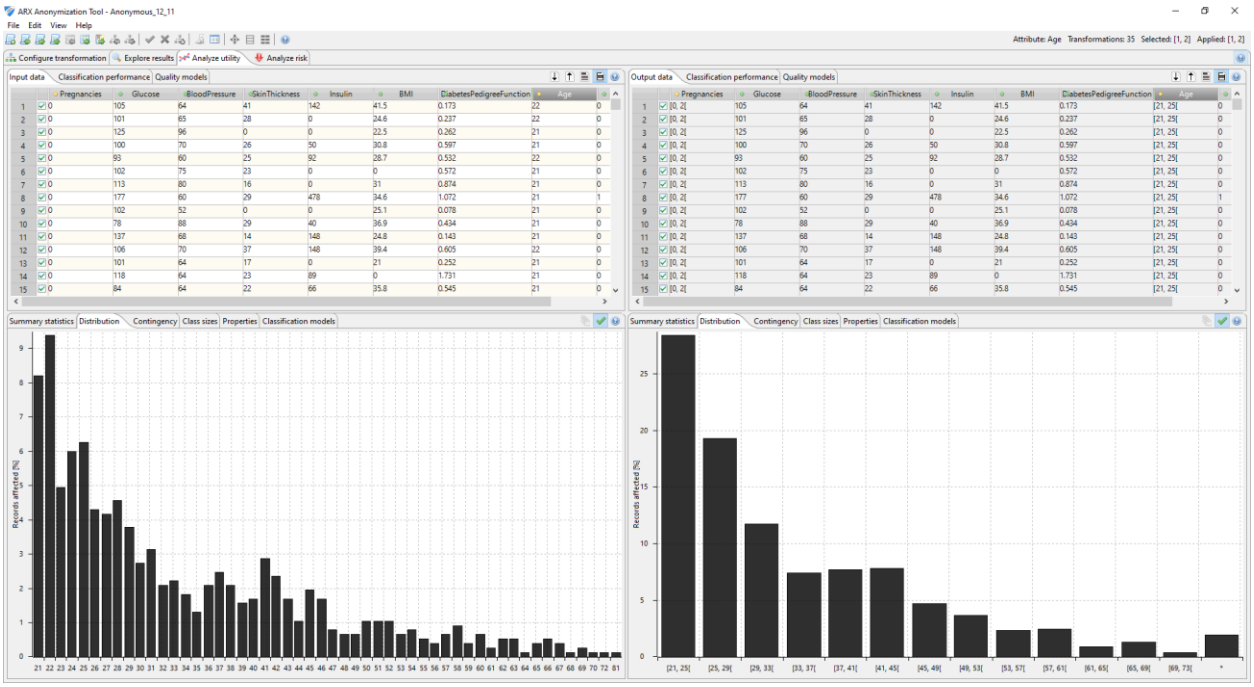


Figure 2 – Age attribute distributions

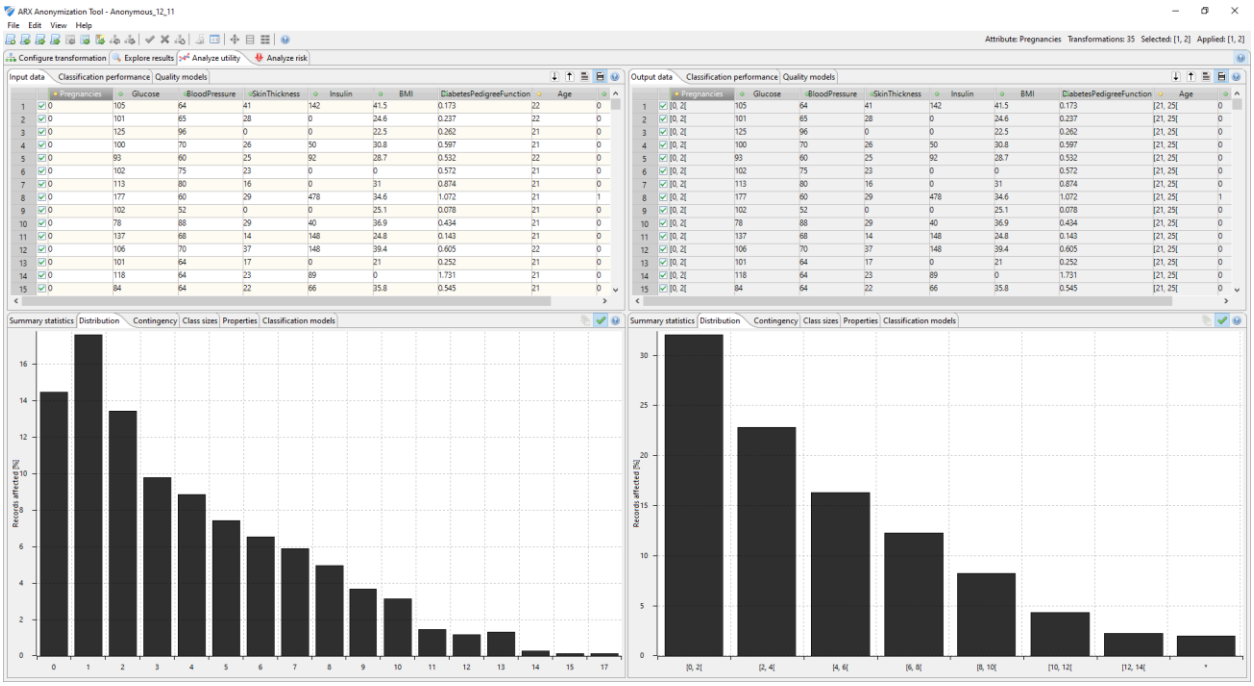


Figure 3 – Pregnancies attribute distributions

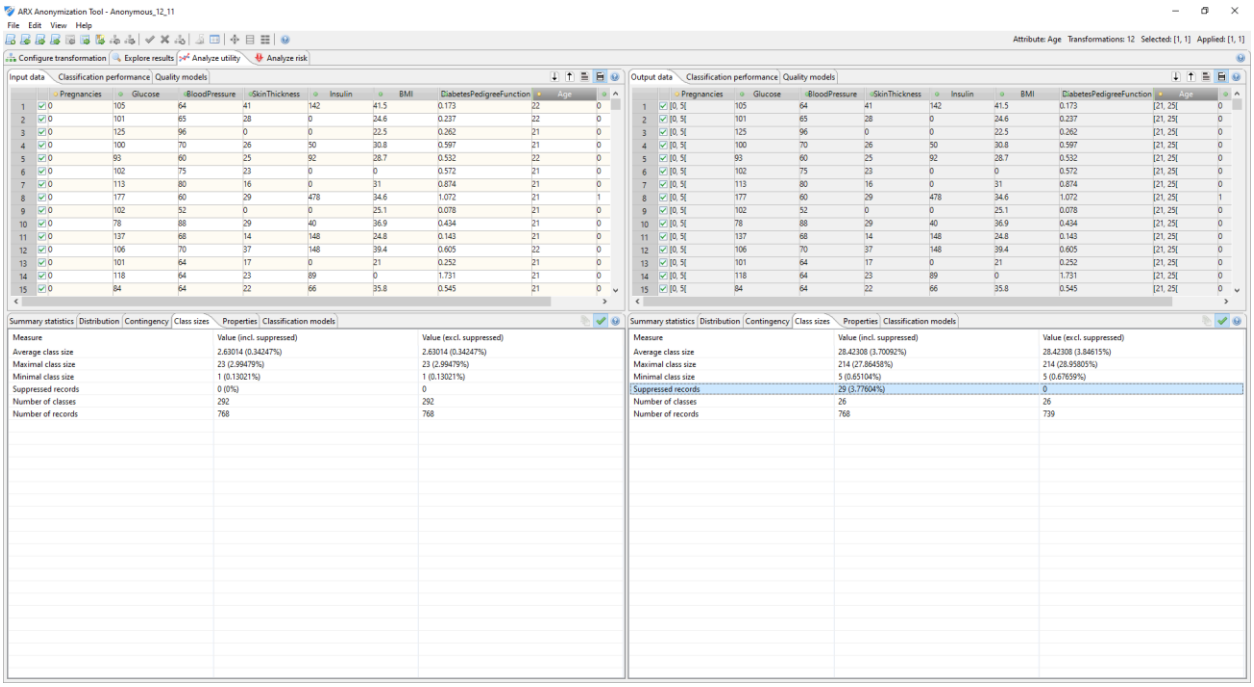


Figure 4 – How much data has been deleted

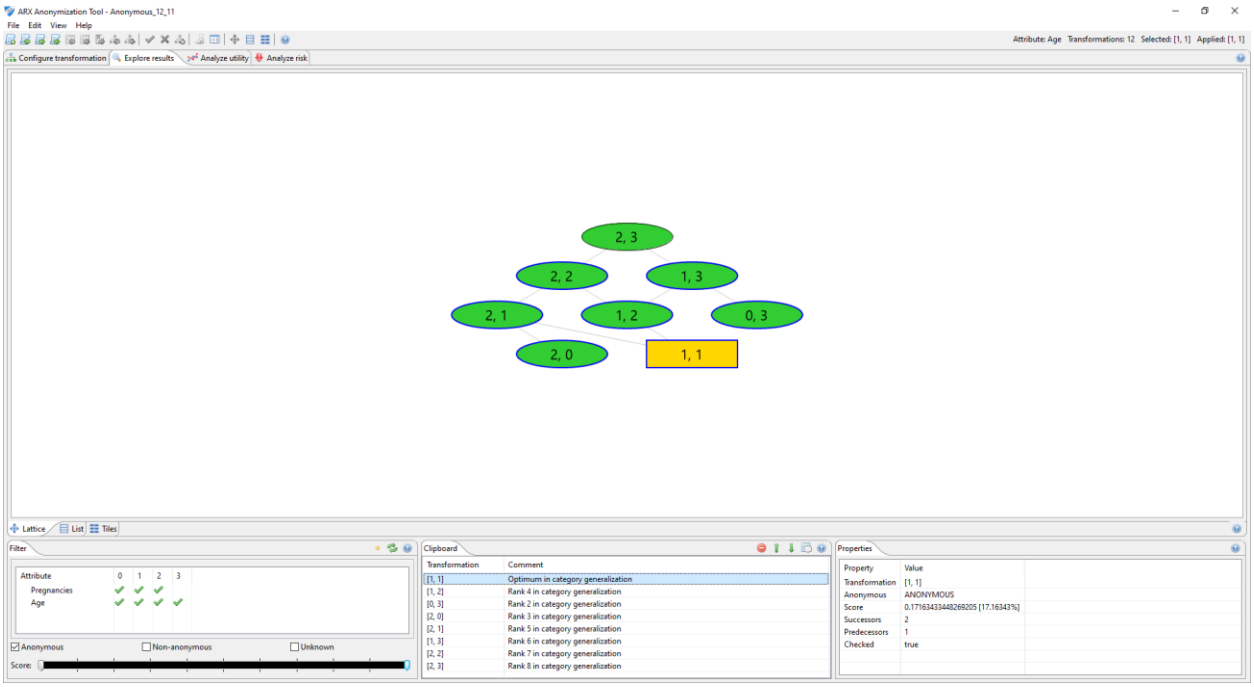


Figure 5 – optimum

Exercise 1.2. Using the weka tool, try to learn marital status, using the three classification approaches with the default parameters for J48, Random Forests, and Naive Bayes. What is the maximum score in terms of learning (UMax)?

```
=== Summary ===

Correctly Classified Instances       30565               67.5888 %
Incorrectly Classified Instances     14657               32.4112 %
Kappa statistic                          0.4833
Mean absolute error                      0.1344
Root mean squared error                  0.2635
Relative absolute error                 71.5163 %
Root relative squared error             85.9672 %
Total Number of Instances            45222

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.679    0.123    0.724      0.679    0.701      0.566    0.812     0.699     Never-married
                 0.844    0.301    0.710      0.844    0.771      0.545    0.795     0.703     Married-civ-spouse
                 0.379    0.076    0.446      0.379    0.410      0.325    0.723     0.321     Divorced
                 0.000    0.000    0.000      0.000    0.000      -0.002   0.524     0.016     Married-spouse-absent
                 0.020    0.003    0.183      0.020    0.036      0.051    0.628     0.072     Separated
                 0.000    0.000    0.000      0.000    0.000      0.000    0.513     0.001     Married-AF-spouse
                 0.360    0.011    0.477      0.360    0.410      0.400    0.806     0.281     Widowed
Weighted Avg.    0.676    0.191    0.646      0.676    0.655      0.495    0.782     0.608

=== Confusion Matrix ===

    a     b     c     d     e     f     g   <-- classified as
 9917  3539   996     7    42     0    97 |     a =  Never-married
 2151 17773   998     5    22     0   106 |     b =  Married-civ-spouse
  992  2619  2387     6    43     0   250 |     c =  Divorced
  164   263   106     0     4     0    15 |     d =  Married-spouse-absent
  386   543   413     4    28     0    37 |     e =  Separated
   15    12     5     0     0     0     0 |     f =  Married-AF-spouse
   66   294   443     0    14     0   460 |     g =  Widowed
```

Figure 6 – For J48: Accuracy - 67.5888%

```
Correctly Classified Instances       28394               62.788  %
Incorrectly Classified Instances     16828               37.212  %
Kappa statistic                          0.4145
Mean absolute error                      0.129
Root mean squared error                  0.278
Relative absolute error                 68.6466 %
Root relative squared error             90.6916 %
Total Number of Instances            45222

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.662    0.152    0.675      0.662    0.668      0.513    0.828     0.741     Never-married
                 0.780    0.305    0.690      0.780    0.732      0.474    0.795     0.730     Married-civ-spouse
                 0.295    0.085    0.361      0.295    0.325      0.229    0.705     0.284     Divorced
                 0.009    0.005    0.023      0.009    0.013      0.007    0.614     0.019     Married-spouse-absent
                 0.050    0.015    0.098      0.050    0.067      0.049    0.630     0.055     Separated
                 0.000    0.000    0.000      0.000    0.000      -0.000   0.499     0.001     Married-AF-spouse
                 0.286    0.014    0.371      0.286    0.323      0.309    0.822     0.253     Widowed
Weighted Avg.    0.628    0.204    0.603      0.628    0.613      0.429    0.786     0.628

=== Confusion Matrix ===

    a     b     c     d     e     f     g   <-- classified as
 9668  3543  1052    44   186     3   102 |     a =  Never-married
 2760 16425  1408    89   214     2   157 |     b =  Married-civ-spouse
 1184  2729  1860    52   188     2   282 |     c =  Divorced
  163   245   102     5    17     0    20 |     d =  Married-spouse-absent
  423   524   324    10    71     2    57 |     e =  Separated
   14     9     6     0     3     0     0 |     f =  Married-AF-spouse
  119   328   406    14    44     1   365 |     g =  Widowed
```

Figure 7 – For Random Forests: Accuracy - 62.778%

```
Correctly Classified Instances       30194               66.7684 %
Incorrectly Classified Instances     15028               33.2316 %
Kappa statistic                          0.4725
Mean absolute error                      0.1264
Root mean squared error                  0.2604
Relative absolute error                 67.2558 %
Root relative squared error             84.9482 %
Total Number of Instances            45222

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.667    0.131    0.709      0.667    0.687      0.545    0.852     0.779     Never-married
                 0.835    0.295    0.712      0.835    0.768      0.541    0.825     0.763     Married-civ-spouse
                 0.378    0.080    0.435      0.378    0.405      0.317    0.766     0.355     Divorced
                 0.031    0.004    0.094      0.031    0.046      0.047    0.716     0.041     Married-spouse-absent
                 0.016    0.002    0.176      0.016    0.029      0.044    0.741     0.094     Separated
                 0.000    0.000    0.000      0.000    0.000      0.000    0.633     0.001     Married-AF-spouse
                 0.359    0.012    0.460      0.359    0.403      0.391    0.918     0.365     Widowed
Weighted Avg.    0.668    0.191    0.640      0.668    0.649      0.485    0.824     0.670

=== Confusion Matrix ===

     a     b     c     d     e     f     g   <-- classified as
  9742  3625  1057    43    34     0    97 |     a =  Never-married
  2217 17572  1063    60    24     0   119 |     b =  Married-civ-spouse
  1119  2482  2383    24    30     0   259 |     c =  Divorced
   160   239   108    17    10     0    18 |     d =  Married-spouse-absent
   440   468   415    21    22     0    45 |     e =  Separated
    17    10     5     0     0     0     0 |     f =  Married-AF-spouse
    52   294   452    16     5     0   458 |     g =  Widowed
```

Figure 8 – For Naive Bayes: Accuracy - 66.7684%