**Initial results (Using default configuration/without optimization)**

**Methodology**

1. Preprocessing

* Remove row data if na value exists on column manual\_assessment
* Remove column username,word\_origin,word\_translation
* Partition dataset using 80/20 proportion, random sampling with consideration of distribution on each target class

1. Model Creation by assign algorithms to learn training dataset
2. Performance assessment by assign generated model to predict test dataset
3. Compare prediction with real target value in test dataset, calculate accuracy

**In summary**

Naïve Bayes : 67.58 % - 70.21%

SVM

- Multiclass SVM : 80.1 % -

Decision Tree

- CHAID : 85.23 % - 86.24%

- CART : 82.03 % - 82.03%

- C50 : 90.32 % - 90.72 %

Random Forest : 90.75 % - 93.3%

**96%quantile – 92.69%**

**> str(train\_dataset1\_normMDA)**

**'data.frame': 18668 obs. of 4 variables:**

**$ origin\_mean\_vote\_up : num 1.67 1.5 1.95 1.8 1.77 ...**

**$ translation\_mean\_duration: num 6.93 17.38 11.92 26.77 91.59 ...**

**$ translation\_jaro\_distance: num 0.557 0.404 0 0.59 0.527 ...**

**$ manual\_origin\_source : num 3 1 3 3 1 1 1 3 3 3 ...**

97%quantile – 82%

'data.frame': 18668 obs. of 4 variables:

$ translation\_mean\_duration: num 6.93 17.38 11.92 26.77 91.59 ...

$ translation\_jaro\_distance: num 0.557 0.404 0 0.59 0.527 ...

$ manual\_origin\_source : num 3 1 3 3 1 1 1 3 3 3 ..

**Details**

**Random Forest**

pred 1 2 3 4 5

1 115 17 15 5 20

2 1 22 0 1 0

3 24 20 366 16 17

4 1 3 1 109 2

5 58 54 89 89 3634

Confusion Matrix and Statistics

Reference

Prediction 1 2 3 4 5

1 115 17 15 5 20

2 1 22 0 1 0

3 24 20 366 16 17

4 1 3 1 109 2

5 58 54 89 89 3634

Overall Statistics

Accuracy : 0.9075

95% CI : (0.8988, 0.9156)

No Information Rate : 0.785

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.719

Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

Class: 1 Class: 2 Class: 3 Class: 4 Class: 5

Sensitivity 0.57789 0.189655 0.77707 0.49545 0.9894

Specificity 0.98728 0.999562 0.98170 0.99843 0.7117

Pos Pred Value 0.66860 0.916667 0.82619 0.93966 0.9261

Neg Pred Value 0.98136 0.979807 0.97521 0.97567 0.9483

Prevalence 0.04253 0.024792 0.10066 0.04702 0.7850

Detection Rate 0.02458 0.004702 0.07822 0.02330 0.7767

Detection Prevalence 0.03676 0.005129 0.09468 0.02479 0.8386

Balanced Accuracy 0.78258 0.594608 0.87939 0.74694 0.8506

**Naïve Bayes**

pred 1 2 3 4 5

1 23 6 19 1 26

2 80 58 108 68 342

3 21 15 131 18 85

4 25 13 68 39 309

5 50 24 145 94 2911

Confusion Matrix and Statistics

Reference

Prediction 1 2 3 4 5

1 23 6 19 1 26

2 80 58 108 68 342

3 21 15 131 18 85

4 25 13 68 39 309

5 50 24 145 94 2911

Overall Statistics

Accuracy : 0.6758

95% CI : (0.6622, 0.6892)

No Information Rate : 0.785

P-Value [Acc > NIR] : 1

Kappa : 0.2707

Mcnemar's Test P-Value : <2e-16

Statistics by Class:

Class: 1 Class: 2 Class: 3 Class: 4 Class: 5

Sensitivity 0.115578 0.50000 0.2781 0.177273 0.7925

Specificity 0.988393 0.86895 0.9670 0.906930 0.6889

Pos Pred Value 0.306667 0.08841 0.4852 0.085903 0.9029

Neg Pred Value 0.961772 0.98558 0.9229 0.957160 0.4763

Prevalence 0.042530 0.02479 0.1007 0.047019 0.7850

Detection Rate 0.004916 0.01240 0.0280 0.008335 0.6221

Detection Prevalence 0.016029 0.14020 0.0577 0.097029 0.6890

Balanced Accuracy 0.551985 0.68447 0.6225 0.542101 0.7407

“Multiclass SVM”

Confusion Matrix and Statistics

Reference

Prediction 1 2 3 4 5

1 35 9 22 11 58

2 0 0 0 0 0

3 20 18 41 10 39

4 0 0 0 0 1

5 144 89 408 199 3575

Overall Statistics

Accuracy : 0.7803

95% CI : (0.7682, 0.7921)

No Information Rate : 0.785

P-Value [Acc > NIR] : 0.7887

Kappa : 0.1394

Mcnemar's Test P-Value : NA

Statistics by Class:

Class: 1 Class: 2 Class: 3 Class: 4 Class: 5

Sensitivity 0.17588 0.00000 0.087049 0.0000000 0.9733

Specificity 0.97768 1.00000 0.979325 0.9997757 0.1650

Pos Pred Value 0.25926 NaN 0.320313 0.0000000 0.8097

Neg Pred Value 0.96391 0.97521 0.905515 0.9529714 0.6288

Prevalence 0.04253 0.02479 0.100663 0.0470186 0.7850

Detection Rate 0.00748 0.00000 0.008763 0.0000000 0.7641

Detection Prevalence 0.02885 0.00000 0.027356 0.0002137 0.9436

Balanced Accuracy 0.57678 0.50000 0.533187 0.4998879 0.5692

**"CHAID"**

Confusion Matrix and Statistics

Reference

Prediction 1 2 3 4 5

1 99 24 23 11 41

2 2 9 2 1 5

3 28 22 265 26 44

4 5 10 12 44 12

5 65 51 169 138 3571

Overall Statistics

Accuracy : 0.8523

95% CI : (0.8418, 0.8624)

No Information Rate : 0.785

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.5369

Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

Class: 1 Class: 2 Class: 3 Class: 4 Class: 5

Sensitivity 0.49749 0.077586 0.56263 0.200000 0.9722

Specificity 0.97790 0.997808 0.97148 0.991254 0.5795

Pos Pred Value 0.50000 0.473684 0.68831 0.530120 0.8941

Neg Pred Value 0.97768 0.977039 0.95203 0.961706 0.8511

Prevalence 0.04253 0.024792 0.10066 0.047019 0.7850

Detection Rate 0.02116 0.001923 0.05664 0.009404 0.7632

Detection Prevalence 0.04232 0.004061 0.08228 0.017739 0.8536

Balanced Accuracy 0.73769 0.537697 0.76706 0.595627 0.7759

**"CART"**

Confusion Matrix and Statistics

Reference

Prediction 1 2 3 4 5

1 16 0 5 3 0

2 0 0 0 0 0

3 79 46 262 42 113

4 0 0 0 0 0

5 104 70 204 175 3560

Overall Statistics

Accuracy : 0.8203

95% CI : (0.809, 0.8312)

No Information Rate : 0.785

P-Value [Acc > NIR] : 1.142e-09

Kappa : 0.397

Mcnemar's Test P-Value : NA

Statistics by Class:

Class: 1 Class: 2 Class: 3 Class: 4 Class: 5

Sensitivity 0.080402 0.00000 0.55626 0.00000 0.9692

Specificity 0.998214 1.00000 0.93346 1.00000 0.4503

Pos Pred Value 0.666667 NaN 0.48339 NaN 0.8655

Neg Pred Value 0.960687 0.97521 0.94948 0.95298 0.8004

Prevalence 0.042530 0.02479 0.10066 0.04702 0.7850

Detection Rate 0.003420 0.00000 0.05599 0.00000 0.7608

Detection Prevalence 0.005129 0.00000 0.11584 0.00000 0.8790

Balanced Accuracy 0.539308 0.50000 0.74486 0.50000 0.7098

**"C50"**

Confusion Matrix and Statistics

Reference

Prediction 1 2 3 4 5

1 134 30 23 18 35

2 8 22 7 4 13

3 16 17 360 12 27

4 6 10 10 129 17

5 35 37 71 57 3581

Overall Statistics

Accuracy : 0.9032

95% CI : (0.8944, 0.9115)

No Information Rate : 0.785

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.7251

Mcnemar's Test P-Value : 5.47e-13

Statistics by Class:

Class: 1 Class: 2 Class: 3 Class: 4 Class: 5

Sensitivity 0.67337 0.189655 0.76433 0.58636 0.9750

Specificity 0.97634 0.992987 0.98289 0.99036 0.8012

Pos Pred Value 0.55833 0.407407 0.83333 0.75000 0.9471

Neg Pred Value 0.98536 0.979676 0.97386 0.97981 0.8976

Prevalence 0.04253 0.024792 0.10066 0.04702 0.7850

Detection Rate 0.02864 0.004702 0.07694 0.02757 0.7653

Detection Prevalence 0.05129 0.011541 0.09233 0.03676 0.8081

Balanced Accuracy 0.82485 0.591321 0.87361 0.78836 0.8881

Feature Importance

> importance(modelRandomForest);

1 2 3 4 5

user\_point 2.4760477 -3.3784361 1.419018559 -2.00064354 11.2890773

user\_is\_male 0.1311174 -0.9155260 -2.003424807 -0.43368246 2.0647504

user\_is\_female 0.3854888 -1.1165018 -1.550633653 0.15563636 0.5455110

user\_is\_age\_0 -1.0010015 0.0000000 0.000000000 0.00000000 0.0000000

user\_is\_age\_15 -0.4665611 -0.2539880 0.231623701 -0.04467362 3.2869497

user\_is\_age\_20 0.1196253 -2.1635150 -3.049539109 -2.10656694 2.7540629

user\_is\_age\_25 -0.3803639 -2.3692463 0.840503567 -1.49294067 -0.6544116

user\_is\_age\_30 2.4768461 0.4664017 0.474540025 -0.27167909 0.9110807

user\_is\_age\_35 0.3527174 0.7008253 -0.273839024 -0.44330566 0.1482448

user\_is\_age\_40 1.7370550 -1.0010015 2.509759676 -2.46296159 0.4099715

user\_languages 0.8496861 0.5008030 -1.314517819 -0.37624686 2.5757967

user\_number\_duration\_played 1.3019054 -1.1974180 0.568488927 -2.42922034 6.7251241

user\_mean\_duration\_played 0.2978190 -1.4262738 -1.093118576 -0.10085139 5.7107729

user\_median\_duration\_played 1.6025478 -0.5477582 -1.905081706 -2.32010546 7.2045675

user\_total\_duration\_played 1.6470532 -0.7774005 2.270072582 -3.39386905 12.6561895

user\_number\_time\_to\_translate 3.0015178 -1.0371768 1.157141451 -3.31453348 11.1767297

user\_mean\_time\_to\_translate 1.2140959 -1.7872201 -1.919980885 -1.44388424 7.8878632

user\_median\_time\_to\_translate 0.4937931 -0.3019443 -2.800458522 -1.55607558 9.7519950

user\_total\_time\_to\_translate 1.4934943 -1.7038039 0.316187937 -2.97491708 8.8617775

user\_number\_time\_to\_alternate -0.1815548 -2.7317580 0.492984968 -3.93521776 12.3995854

user\_mean\_time\_to\_alternate 0.6402928 -0.3896531 1.127950681 -0.99162587 8.9308801

user\_median\_time\_to\_alternate 1.0804180 0.1638903 1.209372586 -1.06883666 10.9565061

user\_total\_time\_to\_alternate 0.5354554 -3.6598592 1.200144108 -4.37702708 11.2516528

user\_number\_time\_to\_vote\_up 3.6961469 -0.1722077 0.331827857 -2.11668710 13.8451231

user\_mean\_time\_to\_vote\_up 2.8725346 -1.0525006 -1.079464128 -2.81013910 10.3517964

user\_median\_time\_to\_vote\_up 3.2169681 -2.0195404 -1.121304972 -2.68564667 11.4876075

user\_total\_time\_to\_vote\_up 1.5098318 -0.8534593 0.835154710 -1.24506061 12.4720087

user\_number\_time\_to\_vote\_down 6.8334223 -4.6911687 0.659427715 -3.79775978 11.5359738

user\_mean\_time\_to\_vote\_down -1.2346252 -0.9917419 0.427033859 -2.71284440 8.6217214

user\_median\_time\_to\_vote\_down 1.2215868 -1.1801584 0.479124591 -2.13313897 9.1239133

user\_total\_time\_to\_vote\_down 3.8389239 -2.2585817 -1.140040456 -3.85369214 9.2187427

user\_number\_time\_to\_categorize 3.7309971 -3.4680957 1.792028941 -3.10721277 10.8015443

user\_mean\_time\_to\_categorize 0.4151415 -2.4789993 3.040013902 -1.82137743 12.1028941

user\_median\_time\_to\_categorize 1.9646895 -2.0536447 -0.505819911 -4.82921362 9.2842842

user\_total\_time\_to\_categorize 3.6170024 -3.1484459 2.014607463 -2.14789317 12.6078863

user\_number\_skip\_translate 1.7419037 -1.1257717 -0.818237529 -3.20792423 11.7821219

user\_mean\_skip\_translate -0.8386201 -3.3684312 -2.132045979 -2.18061246 6.8462826

user\_median\_skip\_translate 0.2381228 -2.3367218 -1.001813252 -4.15400488 8.1030220

user\_total\_skip\_translate 0.5509849 -1.3653656 0.413563244 -3.09274848 12.9708565

user\_number\_skip\_alternate 1.9810248 -2.7799547 0.028565508 -0.58768297 10.7543359

user\_mean\_skip\_alternate 2.3013477 -0.6985294 0.609792344 -0.49666135 10.9695479

user\_median\_skip\_alternate 1.8267978 -2.5784725 -1.177074076 -4.64324455 9.9944606

user\_total\_skip\_alternate 1.4346551 -0.1293248 1.791468680 -3.32464826 11.9090975

user\_number\_skip\_vote 1.4873896 -3.7345571 0.004532935 -3.46414979 9.7257258

user\_mean\_skip\_vote -0.5473838 -0.7233228 -0.457678108 -1.99518122 7.7481916

user\_median\_skip\_vote -1.2810103 -1.8821771 -0.881411139 -2.37458994 9.1765476

user\_total\_skip\_vote 5.3759284 -1.4936003 0.081176298 -3.86058777 9.6819194

user\_number\_skip\_categorize 1.6789344 0.4287061 1.040699675 -2.74496154 9.8867367

user\_mean\_skip\_categorize 1.8743161 -2.2344380 0.108482362 -1.04365680 8.1099461

user\_median\_skip\_categorize 3.9512836 -1.3697534 -2.616203978 -2.51999212 9.0999696

user\_total\_skip\_categorize 4.1833091 0.1058044 0.191942436 -2.80157583 10.8360622

origin\_n\_show 18.7343059 17.4557235 25.352703030 25.84335452 24.4678094

origin\_total\_duration 29.0338660 22.1143561 32.623501145 33.84924292 32.4695938

origin\_mean\_duration 29.9948051 23.5234069 29.561081281 30.21851594 28.3639225

origin\_median\_duration 32.3529833 24.3290917 33.904365957 31.50156981 29.4707152

origin\_total\_vote\_up 20.8835171 21.6247323 27.083415780 24.55994733 23.8577017

origin\_mean\_vote\_up 27.3772905 25.5471854 36.010603459 33.00037529 34.3114394

origin\_median\_vote\_up 12.2237680 7.8821218 14.639400389 13.24271085 12.4124851

origin\_total\_vote\_down 19.7498135 18.5547442 25.355140969 25.57164143 18.3569446

origin\_mean\_vote\_down 28.7095438 22.4739702 33.279422605 31.22013262 31.5367540

origin\_median\_vote\_down 8.2714081 1.3656607 10.272013702 7.94644787 5.1742295

origin\_n\_gram 12.4987395 14.1210091 16.666647621 14.56291905 9.4913067

origin\_n\_char 24.3742004 21.1469105 31.487835809 32.35423049 28.0073230

origin\_n\_user 19.2000399 18.4384052 27.343638353 28.12020846 23.3941136

origin\_n\_diff\_translation 21.8295334 18.0438745 23.541065086 21.98283839 23.4690266

origin\_readibility 15.1822862 12.3047896 23.049588531 21.06051777 11.5085941

origin\_mean\_n\_gram\_translation 24.9138773 23.4534175 32.137562121 29.44542854 24.4847704

origin\_mean\_n\_char\_translation 26.4633803 22.9073244 33.908952710 35.83593088 27.1976694

origin\_num\_most\_common\_translation 22.3386733 22.9028717 31.671462504 28.14806699 29.7451876

origin\_most\_common\_n\_gram 11.5652156 12.8216073 14.047835479 13.13764217 11.4274247

origin\_most\_common\_n\_char 24.3617208 18.8640917 31.560446676 30.69219426 24.6844092

origin\_word\_entropy 25.9920744 20.2454857 35.000097411 32.88377857 29.6041249

translation\_n\_show 23.8770827 21.5689507 26.831864548 27.02796763 27.7218753

translation\_total\_duration 23.7989245 21.1591592 33.246696681 34.55287301 31.7232741

translation\_mean\_duration 22.3943222 20.1899687 32.534968044 35.95177307 37.7289515

translation\_median\_duration 25.1860211 23.7765213 35.945891087 34.97134140 32.2862499

translation\_total\_vote\_up 26.7706626 25.9365997 31.383052991 32.88673791 29.7183366

translation\_mean\_vote\_up 26.8051724 22.0537065 35.421419246 32.69640606 31.1925442

translation\_median\_vote\_up 14.7864414 10.8422151 13.699388578 13.96393477 12.1372811

translation\_total\_vote\_down 19.6046767 17.2111180 20.673888570 20.74927368 21.8346936

translation\_mean\_vote\_down 26.8162778 26.4405398 27.924762337 27.55773160 30.9262122

translation\_median\_vote\_down 17.3018735 8.0348291 5.615981271 11.23851780 10.6499666

translation\_n\_gram 9.2754796 11.4375429 14.595140120 13.85959760 13.2118043

translation\_n\_char 19.5388962 18.3071977 31.379609507 28.81540809 28.2817306

translation\_n\_user 25.3351505 23.2055011 28.285518920 28.45692925 25.4445218

action\_duration 3.7343487 0.5268645 3.736951824 -1.82811335 20.0867278

action\_vote\_up 7.4331056 2.8710209 0.172911271 0.35968986 7.2289113

action\_vote\_down 9.9595775 0.8397590 -0.789446899 2.44969629 8.4685717

is\_top\_translation 12.6524989 12.4096013 13.125259472 14.85956580 18.5333783

translation\_lv\_distance 20.2176516 19.3222563 31.968349101 29.16455930 25.7978796

translation\_jaro\_distance 28.4757782 23.9600626 39.670023993 36.98991543 43.1786624

manual\_origin\_source 57.9155145 47.0369660 71.522501616 54.60374084 64.0584008

MeanDecreaseAccuracy MeanDecreaseGini

user\_point 11.32448936 27.2806777

user\_is\_male 0.54732443 6.2308702

user\_is\_female -0.20659002 6.1564108

user\_is\_age\_0 -1.00100150 0.1518895

user\_is\_age\_15 2.90156111 5.2696333

user\_is\_age\_20 -0.04162483 7.5153853

user\_is\_age\_25 -1.32174480 5.9750831

user\_is\_age\_30 1.63145319 4.4725497

user\_is\_age\_35 0.19372081 1.4984112

user\_is\_age\_40 0.88007773 1.4603721

user\_languages 1.93864885 8.8452502

user\_number\_duration\_played 5.93612934 22.6997455

user\_mean\_duration\_played 4.18692248 40.3480007

user\_median\_duration\_played 5.24369039 43.0674773

user\_total\_duration\_played 11.65189552 29.1987599

user\_number\_time\_to\_translate 11.17479279 26.9709039

user\_mean\_time\_to\_translate 6.04529198 35.2702518

user\_median\_time\_to\_translate 7.91351410 34.5197641

user\_total\_time\_to\_translate 8.30073610 29.0800067

user\_number\_time\_to\_alternate 10.33424672 27.0975406

user\_mean\_time\_to\_alternate 8.92619385 33.1314000

user\_median\_time\_to\_alternate 10.21389410 34.2489971

user\_total\_time\_to\_alternate 10.04370227 29.5697281

user\_number\_time\_to\_vote\_up 12.99225536 26.8349694

user\_mean\_time\_to\_vote\_up 9.32244122 35.3085722

user\_median\_time\_to\_vote\_up 10.00692104 34.8150636

user\_total\_time\_to\_vote\_up 12.04926450 27.7766848

user\_number\_time\_to\_vote\_down 11.23717595 28.4558812

user\_mean\_time\_to\_vote\_down 6.64463275 36.1042610

user\_median\_time\_to\_vote\_down 7.65134624 36.9714299

user\_total\_time\_to\_vote\_down 8.17922970 31.4540313

user\_number\_time\_to\_categorize 10.50033982 25.1158028

user\_mean\_time\_to\_categorize 11.88544008 34.8631955

user\_median\_time\_to\_categorize 7.93478962 34.3348416

user\_total\_time\_to\_categorize 12.01307974 26.9709454

user\_number\_skip\_translate 10.95432182 24.2240282

user\_mean\_skip\_translate 4.29201759 32.8868347

user\_median\_skip\_translate 5.91835597 32.0555734

user\_total\_skip\_translate 11.51526134 27.2814766

user\_number\_skip\_alternate 9.59385292 27.3491311

user\_mean\_skip\_alternate 10.25489862 34.2280076

user\_median\_skip\_alternate 7.05686299 33.7518849

user\_total\_skip\_alternate 11.48507483 28.3707480

user\_number\_skip\_vote 8.56467826 22.4669305

user\_mean\_skip\_vote 6.06486092 33.1473253

user\_median\_skip\_vote 6.14547885 32.6941194

user\_total\_skip\_vote 9.34243521 25.3123836

user\_number\_skip\_categorize 9.34656819 24.0584817

user\_mean\_skip\_categorize 7.19686852 33.6997988

user\_median\_skip\_categorize 7.64514202 35.0226440

user\_total\_skip\_categorize 11.09195820 27.5998390

origin\_n\_show 28.00198230 108.4323302

origin\_total\_duration 36.55509276 185.5823873

origin\_mean\_duration 31.10752079 214.1755309

origin\_median\_duration 33.25012813 203.8064595

origin\_total\_vote\_up 26.80439862 157.5484696

origin\_mean\_vote\_up 38.69663053 214.1294740

origin\_median\_vote\_up 14.74244743 37.5996394

origin\_total\_vote\_down 22.79405229 107.1404427

origin\_mean\_vote\_down 35.27877103 205.1437484

origin\_median\_vote\_down 8.84330570 10.5737489

origin\_n\_gram 11.68930370 44.0145714

origin\_n\_char 34.00069517 138.9759492

origin\_n\_user 27.98224375 112.3402606

origin\_n\_diff\_translation 24.46383195 124.3713888

origin\_readibility 15.88254006 50.2723518

origin\_mean\_n\_gram\_translation 28.28444282 171.7571909

origin\_mean\_n\_char\_translation 31.96858902 177.3539841

origin\_num\_most\_common\_translation 31.95236868 165.7565916

origin\_most\_common\_n\_gram 12.65899376 44.2818886

origin\_most\_common\_n\_char 29.41761374 130.8925343

origin\_word\_entropy 34.25656789 191.3959356

translation\_n\_show 30.08336588 177.2010587

translation\_total\_duration 35.92904780 219.5611139

translation\_mean\_duration 41.98332054 171.8039683

translation\_median\_duration 36.09838827 176.7086036

translation\_total\_vote\_up 33.18386386 235.8515487

translation\_mean\_vote\_up 36.95166597 147.8252951

translation\_median\_vote\_up 15.06858774 64.8201625

translation\_total\_vote\_down 23.05544033 61.5901049

translation\_mean\_vote\_down 33.16000658 127.9521245

translation\_median\_vote\_down 12.92307853 47.6364129

translation\_n\_gram 14.33651051 47.6145059

translation\_n\_char 31.78927513 135.3955711

translation\_n\_user 29.28792144 172.6665524

action\_duration 19.73206624 71.3528942

action\_vote\_up 8.09900130 16.2225226

action\_vote\_down 9.68671232 21.7556801

is\_top\_translation 17.95665258 39.2085615

translation\_lv\_distance 29.99301876 135.5719979

translation\_jaro\_distance 46.42554435 186.6519246

manual\_origin\_source 70.65807078 708.2648855

FEATURE SELECTION

95%quantile – 93.08%

> str(train\_dataset1\_normMDA)

'data.frame': 18668 obs. of 6 variables:

$ origin\_total\_duration : num 206 303 217 628 483 ...

$ origin\_mean\_vote\_up : num 1.67 1.5 1.95 1.8 1.77 ...

$ translation\_mean\_duration: num 6.93 17.38 11.92 26.77 91.59 ...

$ translation\_jaro\_distance: num 0.557 0.404 0 0.59 0.527 ...

$ manual\_origin\_source : num 3 1 3 3 1 1 1 3 3 3 ...

**96%quantile – 92.69%**

**> str(train\_dataset1\_normMDA)**

**'data.frame': 18668 obs. of 5 variables:**

**$ origin\_mean\_vote\_up : num 1.67 1.5 1.95 1.8 1.77 ...**

**$ translation\_mean\_duration: num 6.93 17.38 11.92 26.77 91.59 ...**

**$ translation\_jaro\_distance: num 0.557 0.404 0 0.59 0.527 ...**

**$ manual\_origin\_source : num 3 1 3 3 1 1 1 3 3 3 ...**

97%quantile – 82%

'data.frame': 18668 obs. of 4 variables:

$ translation\_mean\_duration: num 6.93 17.38 11.92 26.77 91.59 ...

$ translation\_jaro\_distance: num 0.557 0.404 0 0.59 0.527 ...

$ manual\_origin\_source : num 3 1 3 3 1 1 1 3 3 3 ...

98%quantile

> str(train\_dataset1\_normMDA)

'data.frame': 18668 obs. of 3 variables:

$ translation\_jaro\_distance: num 0.557 0.404 0 0.59 0.527 ...

$ manual\_origin\_source : num 3 1 3 3 1 1 1 3 3 3 ...

> set.seed(12345);

+

+ modelRandomForest = hitungRandomForest(train\_dataset1\_normMDA, test\_dataset1\_normMDA);

+ hitungNaiveBayes(train\_dataset1\_normMDA, test\_dataset1\_normMDA);

+ hitungDecisionTree(train\_dataset1\_normMDA, test\_dataset1\_normMDA);

+

pred 1 2 3 4 5

1 111 6 8 5 8

2 2 38 0 3 2

3 24 26 400 11 17

4 2 0 3 155 7

5 60 46 60 46 3639

Confusion Matrix and Statistics

Reference

Prediction 1 2 3 4 5

1 111 6 8 5 8

2 2 38 0 3 2

3 24 26 400 11 17

4 2 0 3 155 7

5 60 46 60 46 3639

Overall Statistics

Accuracy : 0.9282

95% CI : (0.9204, 0.9354)

No Information Rate : 0.785

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.7891

Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

Class: 1 Class: 2 Class: 3 Class: 4 Class: 5

Sensitivity 0.55779 0.327586 0.84926 0.70455 0.9907

Specificity 0.99397 0.998466 0.98146 0.99731 0.7893

Pos Pred Value 0.80435 0.844444 0.83682 0.92814 0.9449

Neg Pred Value 0.98062 0.983168 0.98310 0.98559 0.9589

Prevalence 0.04253 0.024792 0.10066 0.04702 0.7850

Detection Rate 0.02372 0.008121 0.08549 0.03313 0.7777

Detection Prevalence 0.02949 0.009617 0.10216 0.03569 0.8230

Balanced Accuracy 0.77588 0.663026 0.91536 0.85093 0.8900

pred 1 2 3 4 5

1 19 11 17 8 22

2 4 2 4 6 13

3 89 53 277 50 145

4 1 2 4 3 20

5 86 48 169 153 3473

Confusion Matrix and Statistics

Reference

Prediction 1 2 3 4 5

1 19 11 17 8 22

2 4 2 4 6 13

3 89 53 277 50 145

4 1 2 4 3 20

5 86 48 169 153 3473

Overall Statistics

Accuracy : 0.8066

95% CI : (0.795, 0.8178)

No Information Rate : 0.785

P-Value [Acc > NIR] : 0.0001483

Kappa : 0.4075

Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

Class: 1 Class: 2 Class: 3 Class: 4 Class: 5

Sensitivity 0.095477 0.0172414 0.5881 0.0136364 0.9455

Specificity 0.987054 0.9940828 0.9199 0.9939448 0.5467

Pos Pred Value 0.246753 0.0689655 0.4511 0.1000000 0.8839

Neg Pred Value 0.960887 0.9754839 0.9523 0.9533233 0.7333

Prevalence 0.042530 0.0247916 0.1007 0.0470186 0.7850

Detection Rate 0.004061 0.0004274 0.0592 0.0006412 0.7423

Detection Prevalence 0.016457 0.0061979 0.1312 0.0064116 0.8397

Balanced Accuracy 0.541265 0.5056621 0.7540 0.5037906 0.7461

[1] "CHAID"

Confusion Matrix and Statistics

Reference

Prediction 1 2 3 4 5

1 20 5 7 1 3

2 0 8 0 0 2

3 76 37 265 49 89

4 0 0 0 6 2

5 103 66 199 164 3577

Overall Statistics

Accuracy : 0.8284

95% CI : (0.8173, 0.8391)

No Information Rate : 0.785

P-Value [Acc > NIR] : 6.944e-14

Kappa : 0.4262

Mcnemar's Test P-Value : NA

Statistics by Class:

Class: 1 Class: 2 Class: 3 Class: 4 Class: 5

Sensitivity 0.100503 0.068966 0.56263 0.027273 0.9739

Specificity 0.996429 0.999562 0.94035 0.999551 0.4712

Pos Pred Value 0.555556 0.800000 0.51357 0.750000 0.8705

Neg Pred Value 0.961447 0.976869 0.95052 0.954185 0.8316

Prevalence 0.042530 0.024792 0.10066 0.047019 0.7850

Detection Rate 0.004274 0.001710 0.05664 0.001282 0.7645

Detection Prevalence 0.007694 0.002137 0.11028 0.001710 0.8782

Balanced Accuracy 0.548466 0.534264 0.75149 0.513412 0.7225

[1] "CART"

Confusion Matrix and Statistics

Reference

Prediction 1 2 3 4 5

1 0 0 0 0 0

2 0 0 0 0 0

3 95 46 267 45 113

4 0 0 0 0 0

5 104 70 204 175 3560

Overall Statistics

Accuracy : 0.8179

95% CI : (0.8065, 0.8289)

No Information Rate : 0.785

P-Value [Acc > NIR] : 1.289e-08

Kappa : 0.3885

Mcnemar's Test P-Value : NA

Statistics by Class:

Class: 1 Class: 2 Class: 3 Class: 4 Class: 5

Sensitivity 0.00000 0.00000 0.56688 0.00000 0.9692

Specificity 1.00000 1.00000 0.92894 1.00000 0.4503

Pos Pred Value NaN NaN 0.47173 NaN 0.8655

Neg Pred Value 0.95747 0.97521 0.95040 0.95298 0.8004

Prevalence 0.04253 0.02479 0.10066 0.04702 0.7850

Detection Rate 0.00000 0.00000 0.05706 0.00000 0.7608

Detection Prevalence 0.00000 0.00000 0.12097 0.00000 0.8790

Balanced Accuracy 0.50000 0.50000 0.74791 0.50000 0.7098

[1] "C50"

Confusion Matrix and Statistics

Reference

Prediction 1 2 3 4 5

1 101 14 10 6 23

2 4 23 3 8 7

3 14 16 348 10 18

4 8 9 12 125 25

5 72 54 98 71 3600

Overall Statistics

Accuracy : 0.897

95% CI : (0.8879, 0.9056)

No Information Rate : 0.785

P-Value [Acc > NIR] : < 2.2e-16

Kappa : 0.6919

Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

Class: 1 Class: 2 Class: 3 Class: 4 Class: 5

Sensitivity 0.50754 0.198276 0.73885 0.56818 0.9801

Specificity 0.98817 0.995179 0.98622 0.98789 0.7068

Pos Pred Value 0.65584 0.511111 0.85714 0.69832 0.9243

Neg Pred Value 0.97834 0.979931 0.97121 0.97889 0.9069

Prevalence 0.04253 0.024792 0.10066 0.04702 0.7850

Detection Rate 0.02159 0.004916 0.07437 0.02672 0.7694

Detection Prevalence 0.03291 0.009617 0.08677 0.03826 0.8324

Balanced Accuracy 0.74785 0.596727 0.86254 0.77804 0.8434