Evaluation of Malaysia's 15th General Election Prediction Models by using Cross Validation Approaches

Article history

Received:
xx xxx 2019

Received in revised form:
xx xxx 2019

Accepted:
xx xxx 2019

Published online:
xx xxx 2019

*Corresponding author
izardy@.graduate.utm.my

Izardy Amiruddin
Syahid Anuar

Razak Faculty, Universiti Teknology Malaysia
izardy@graduate.utm.my
syahid.anuar@utm.my

Abstract

The 14th General Election (GE-14) outcome in Malaysia reflected how the public reacted towards numerous issues within the previous Barisan Nasional (BN) government. Somehow the voting pattern indicates BN still managed to secure a huge number of votes from parliaments within the rural area which consist majority of Malay and Bumiputera voters. However, some of the areas with the majority of Malay voters won by Pakatan Harapan (PH) due to the split of Malay popular votes between BN and another Malay based party PAS. This scenario resulted in the statistical gain for PH. BN could possibly win the previous GE-14 and the incoming GE-15 if they collaborated with PAS to avoid 3 corner fights. This paper evaluates multiple approaches of prediction models for GE-15, based on data models developed from GE-14 results.

Keywords: *Election, Politics, Malaysia, Prediction, Malay, Bumiputera, Parliament*

1. Introduction

The GE-14 results indicated BN lost in 54 numbers of parliament seats from the previous elections. Out of these 54 seats, 31 were marginally won by PH due to split of Malay support between BN and PAS. By considering 79 numbers of seats secured by BN during GE-14 and the 31 numbers of PH marginally won seats, BN have the potential to at least secure 110 parliament seats if the party managed to collaborate with PAS for 2 corner fights with PH. BN 110 seats combined with PAS 18 seats would be able to form a stable government post GE-14. Considering the current event whereby PAS became part of Perikatan Nasional (PN) instead of

* Corresponding author. masaini@abc.xxx.xxx

collaborating with BN through Muafakat Nasional (MN), the tendency to create another 3-corner fight with PH is high and both parties will finally end-up be losing to PH. Considering voting patterns will remain the same, 3 classification machine learning models implemented on the developed data model based on GE-14 result consisting of 60 numbers of attributes. The importance level of these attributes to be measured using Chi-Square method and only relevance attributes to be selected for model training purpose. This study implemented Naive Bayes, Logistic Regression and Random Forest classification machine learning model for comparison. Best classification model which is evaluated using the K-Fold Cross Validation technique to be proposed for the simulation of GE-15 results.

2. Related Work

A research paper published by Asian Journal of Communication proposed the methodology to predict elections from the sentiment of social media. Malaysia, India and Pakistan Twitter platform users had been selected by the authors as their research scope. This research performed sentiment analysis using machine learning models. This study found that the machine learning model performed quite well for India and Pakistan, however it was an opposite outcome for Malaysia. The authors also suggested that combinations of sentiment and volume information are effective at predicting smaller vote shares (Ahmed, Skoric & Hilber, 2018).

Another research paper published by Journal of Ambient Intelligence and Humanized Computing suggested that the election result can be forecasted by leveraging data mining approach from social media platforms. This paper evaluated association on numbers of volumetric parameters, sentiment and social network method to project political favoritism from the social media (Chauhan & Sharma, 2020). The author believed by doing so, a population's sentiments or opinion with regards to their political tendencies can be measured.

While many of the recent election prediction related papers associated with social media sentiment, a study which was conducted in 2006 published by International Journal of Forecasting suggested that voters' choice could be modelled to predict the final outcome of a two-stage election. The author addressed that the election forecast from the single stage process would not be effective as compared to the two-stage process which had been utilized by more than 40 countries. The Nested Logit Factor Model of electors' preference was developed based on political science theories extracted from literature reviews. Politicians would be able to understand their competitive level at the second stage election and would be able to observe their rival performance based on the model outcome using the first stage of election information as training data (Kamakura, Mazzon & Bruyn, 2006).

3. Methodology

The methodology of this research utilized data preprocessing from multiple resources. These different sets of data which consist of information associated with results from the previous GE-14 was then cleaned and merged into a single table. The data preprocessing, cleaning and modelling performed within Jupyter Notebook environment utilizing Python's library Pandas and NumPy. The next process was then performed using Weka data mining software. Since it is a small dataset with only 60 features, the feature selection method was implemented just to confirm the relevancy rank of continuous features which require normalization. After confirmation, those continuous features dropped from the dataset prior to the machine learning implementation process. The feature normalization was not implemented for the rest of attributes as most of the continuous data represented in terms of percentage. The selected classification models performed within the Weka platform implemented both 10 fold cross validation and Leave-one-out cross-validation (LOOCV) by using 222 folds which are equivalent to number of instances. Each of the classification models performance evaluated using both method. Best method which produces the highest accuracy selected for the consideration on simulation of GE-15 outcome.

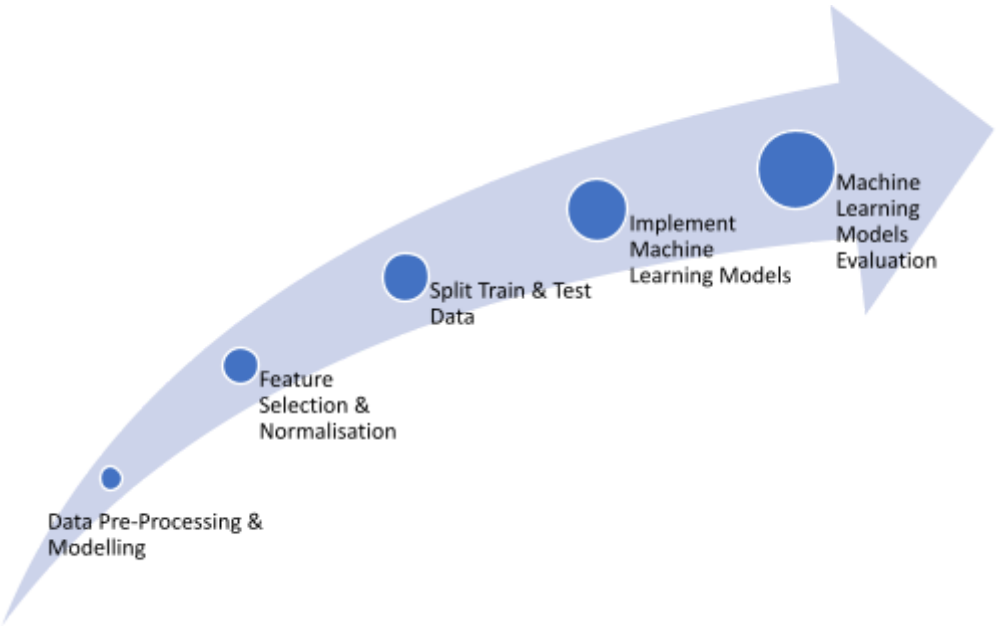

Figure 1. Process Flow

3.1 Data Sourcing, Transformation and Modelling

The raw data for GE-14 result obtained from Kaggle (https://www.kaggle.com/terenctb/malaysia-ge14-election-results-parliament). The raw dataset consists of the parliament id, seat name, candidates, candidates party representation, votes obtained, total registered voters, total votes, turn out and spoilt votes. The raw data for voters' race composition obtained from The Star (https://election.thestar.com.my/) . Though it is not detailed, the figures reflected in the table can be a good indicator for analysis. The data scraped from the website using Google Chrome tool 'Scraper'.

The status of parliament on urbanization categories obtained from the International Journal of Law, Government and Communication (IJLGC) from the paper 'Trend of Voting in Malaysia General Election 2018 in Urban and Rural Area'. The paper highlighted the list of parliaments and the urbanization categories of which these parliament fall on. The status of parliament on FELDA categories obtained from the Akademika journal published by UKM from the paper 'FELDA's Voter Behavior in GE-14: Party Identification, Sociological or Rational Choice)'. The paper highlighted the list of parliaments which fell under FELDA categories.

The data on voters age categories composition obtained from Github repository (https://github.com/khoo-j/MsiaGE14/blob/master/GE14_Age-Ethnicity-bySeats.xlsx) . However, the source accuracy cannot be confirmed. The owner repository claimed that the uploaded data as per supply from DAP Malaysia, Tindak Malaysia and SPR.

There are three types of raw data obtained for this project. The first type is in the csv and spreadsheet format which is a direct transformation process executed using Jupyter Notebook and Python's library (Pandas and NumPy). The second type is the data obtained directly from websites and journals. Data from websites scraped using online tools (Google Chrome Scraper), while data obtained from journals manually extracted to csv. The third type of data is the shapefile data, in order to use this data in Power BI environment, the data must be transformed into topo json format. The transformation process was executed within mapshaper.org website.

All the raw files were then loaded into Jupyter Notebook for pre-processing which include standardization of column name and cell values. All the process files were then saved into separate folders for the purpose of data modelling. The data model was established by joining all the necessary parameters using a common key which either the parliament id or parliament name. This process was also executed in Jupyter Notebook. The process is reflected in Figure 2 whereby the data model represented in Microsoft Power BI environment.

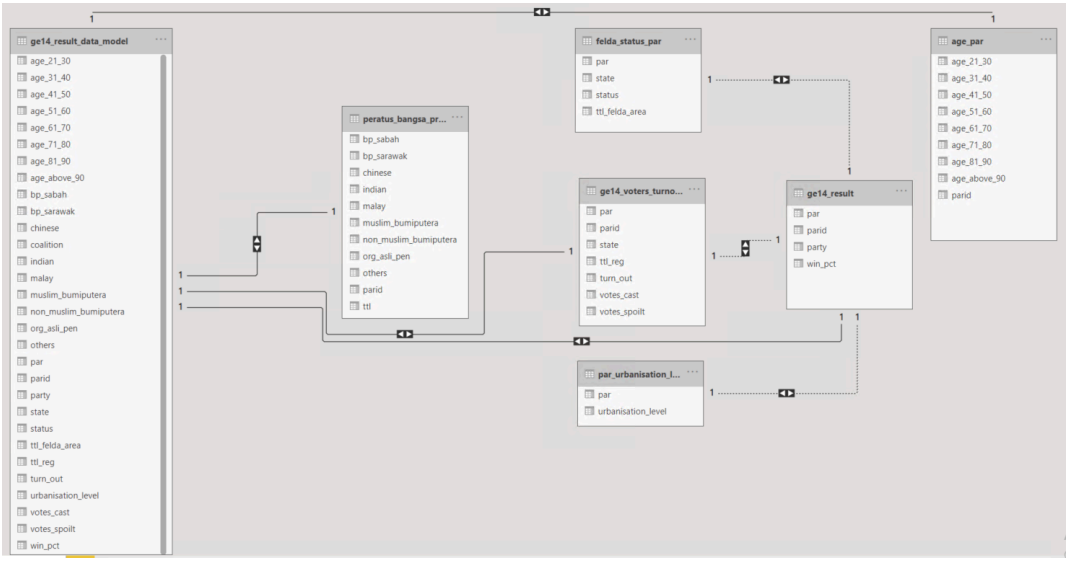

Figure 2. Data Model

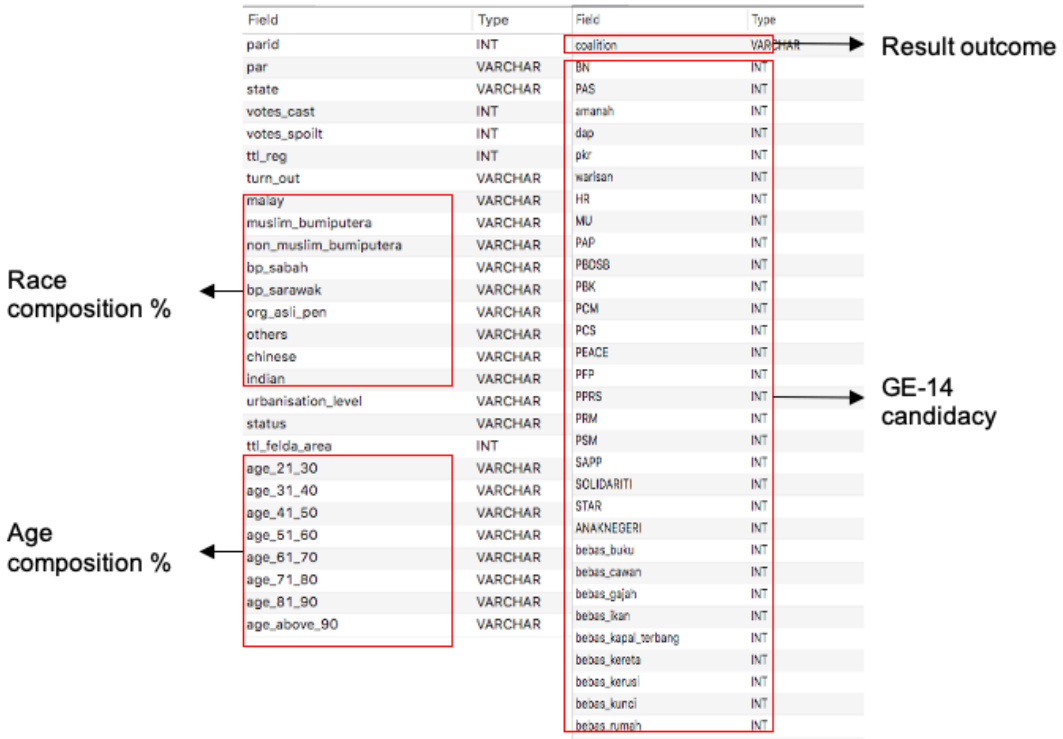

Figure 3. Independent Fields

4. Algorithm

The following section summarized on the algorithms implemented for this research which are pre-developed as part of the tools in Weka Data Mining software

4.1 Cross Validation

The approach of data re-sampling for machine learning model evaluation known as cross validation method. This method applied to test different sets of sample data on the performance of a machine learning model. Cross validation is useful when it comes to minimizing the over-fitting issue. The k-Fold term refers to the number of chunks from a dataset which to be splitted. In general, a small size of dataset is recommended to be applied with higher k , however another approach which is suitable to use with a small dataset is Leave One Out Cross Validation (LOOCV). This approach produces higher accuracy but slower processing whereby only one instance is used for every iteration while the rest for training purpose.

4.2 Naïve Bayes

Naive Bayes classification was based on Bayes Theorem which addressed the assumption of independency among predictors. The theorem assumed the existence of a certain feature in a class is not linked to the presence of any other feature. This method is best utilised for 2 types of class prediction. Despite their seemingly over simplified assumptions, in many real world scenarios, especially document classification and spam filtering, Naive Bayes classifiers have performed very well. In order to estimate the required parameters, this algorithm only needs a small amount of training data. In terms of speed, Naive Bayes classifiers produced faster results compared to other advanced methods. While it performed well in classifying related problems, the estimator performance is bad (H. Zhang, 2004).

4.3 Logistic Regression

Logistic regression is a linear regression-like statistical method because the derived equation for this algorithm predicts outcome from one or more predictor variables, X for binary variables, Y. The conditional probability follows a logistic distribution given by P(Y=1|X=xi) (Ekinci, Omurca & Acun, 2018) The predictor variables may however be categorical or continuous, unlike linear regression which solely require variables in continuous form. This model best applied to determine outcome for 2 types of categorical dependent variables. This algorithm derived from numbers of probabilistic mathematical equations such as Odds Ratio, Logit, Sigmoid and Cross Entropy functions.

4.4 Random Forest

Random Forest works as a large collection of decorrelated decision trees whereby the features selection within the train dataset is selected randomly. It is a modification method of bagged decision trees and within the ensemble type machine learning algorithm which relates to Bootstrap Aggregation or bagging. As per statistical term, bootstrap is a method to estimate quantity from a data

sample. Bootstrap method generates new sub samples to improve estimation accuracy. The bootstrap concept implemented in the form of aggregation to produce more accurate predictions based on combinations of multiple machine learning algorithms compared to single algorithms. The advantage or Random Forest includes non-parametric nature, high performance accuracy the the capacity to identify ranked of variables (Galiano, Ghimire, Rogan, Olma & Sanchez, 2012)

Table 1. Algorithms Performance Measured Using 10-Fold Cross Validation

Classification Algorithms	Time, seconds	TP Rate	Precision	ROC Area
Naive Bayes	0.00	79.73	79.70	89.5
Logistic Regression	0.09	71.17	71.10	72.0
Random Forest	0.06	81.53	81.50	92.3

Table 2. Algorithms Performance Measured Using LOOCV

Classification Algorithms	Time, seconds	TP Rate	Precision	ROC Area
Naive Bayes	0.00	79.28	79.30	89.1
Logistic Regression	0.09	78.83	78.90	81.5
Random Forest	0.07	84.23	84.20	91.7

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         177               79.7297 %
Incorrectly Classified Instances        45               20.2703 %
Kappa statistic                          0.5903
Mean absolute error                      0.1995
Root mean squared error                  0.4284
Relative absolute error                 40.2295 %
Root relative squared error             86.0184 %
Total Number of Instances              222

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.762    0.174    0.786      0.762    0.774      0.590    0.895     0.861     OPPOSITION
                 0.826    0.238    0.806      0.826    0.816      0.590    0.895     0.926     GOVERNMENT
Weighted Avg.    0.797    0.208    0.797      0.797    0.797      0.590    0.895     0.896

=== Confusion Matrix ===

   a   b   <-- classified as
  77  24 |   a = OPPOSITION
  21 100 |   b = GOVERNMENT
```

Figure 4. Naïve Bayes Result using 10-Folds Cross Validation

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         176               79.2793 %
Incorrectly Classified Instances        46               20.7207 %
Kappa statistic                          0.5815
Mean absolute error                      0.202
Root mean squared error                  0.4301
Relative absolute error                 40.5403 %
Root relative squared error             85.9758 %
Total Number of Instances              222

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.762    0.182    0.778      0.762    0.770      0.582    0.891     0.839     OPPOSITION
                 0.818    0.238    0.805      0.818    0.811      0.582    0.891     0.924     GOVERNMENT
Weighted Avg.    0.793    0.212    0.793      0.793    0.793      0.582    0.891     0.886

=== Confusion Matrix ===

  a  b   <-- classified as
 77 24 |  a = OPPOSITION
 22 99 |  b = GOVERNMENT
```

Figure 5. Naïve Bayes Result using LOOCV

```
Time taken to build model: 0.09 seconds

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

Correctly Classified Instances         158               71.1712 %
Incorrectly Classified Instances        64               28.8288 %
Kappa statistic                          0.4129
Mean absolute error                      0.2867
Root mean squared error                  0.5257
Relative absolute error                 57.7949 %
Root relative squared error            105.5734 %
Total Number of Instances              222

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.624    0.215    0.708      0.624    0.663      0.415    0.723     0.639     OPPOSITION
                 0.785    0.376    0.714      0.785    0.748      0.415    0.717     0.720     GOVERNMENT
Weighted Avg.    0.712    0.303    0.711      0.712    0.709      0.415    0.720     0.683

=== Confusion Matrix ===

  a  b   <-- classified as
 63 38 |  a = OPPOSITION
 26 95 |  b = GOVERNMENT
```

Figure 6. Logistic Regression Result using 10-Folds Cross Validation

```
Time taken to build model: 0.09 seconds

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

Correctly Classified Instances         175               78.8288 %
Incorrectly Classified Instances        47               21.1712 %
Kappa statistic                          0.5699
Mean absolute error                      0.22
Root mean squared error                  0.4455
Relative absolute error                 44.159  %
Root relative squared error             89.0681 %
Total Number of Instances              222

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.723    0.157    0.793      0.723    0.756      0.572    0.814     0.729     OPPOSITION
                 0.843    0.277    0.785      0.843    0.813      0.572    0.816     0.816     GOVERNMENT
Weighted Avg.    0.788    0.223    0.789      0.788    0.787      0.572    0.815     0.777

=== Confusion Matrix ===

   a   b   <-- classified as
  73  28 |   a = OPPOSITION
  19 102 |   b = GOVERNMENT
```

Figure 7. Logistic Regression Result using LOOCV

```
Time taken to build model: 0.06 seconds

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

Correctly Classified Instances         181               81.5315 %
Incorrectly Classified Instances        41               18.4685 %
Kappa statistic                          0.6261
Mean absolute error                      0.3119
Root mean squared error                  0.3568
Relative absolute error                 62.8767 %
Root relative squared error             71.6556 %
Total Number of Instances              222

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.772    0.149    0.813      0.772    0.792      0.627    0.923     0.911     OPPOSITION
                 0.851    0.228    0.817      0.851    0.834      0.627    0.923     0.940     GOVERNMENT
Weighted Avg.    0.815    0.192    0.815      0.815    0.815      0.627    0.923     0.927

=== Confusion Matrix ===

   a   b   <-- classified as
  78  23 |   a = OPPOSITION
  18 103 |   b = GOVERNMENT
```

Figure 8. Random Forest Classification Result using 10-Folds Cross Validation

```
Time taken to build model: 0.07 seconds

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

Correctly Classified Instances         187               84.2342 %
Incorrectly Classified Instances        35               15.7658 %
Kappa statistic                          0.6818
Mean absolute error                      0.3107
Root mean squared error                  0.3581
Relative absolute error                 62.3697 %
Root relative squared error             71.5903 %
Total Number of Instances              222

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.822    0.140    0.830      0.822    0.826      0.682    0.917     0.906     OPPOSITION
                 0.860    0.178    0.852      0.860    0.856      0.682    0.917     0.934     GOVERNMENT
Weighted Avg.    0.842    0.161    0.842      0.842    0.842      0.682    0.917     0.922

=== Confusion Matrix ===

   a   b   <-- classified as
  83  18 |   a = OPPOSITION
  17 104 |   b = GOVERNMENT
```

Figure 9. Random Forest Classification Result using LOOCV

5. Results and Discussion

The modelled dataset which is loaded and processed within the Weka platform indicates Random Forest Classification algorithm together with LOOCV produced the best result. This is indicated in Figure 9, whereby correctly classified instances is 84.23% compared to Logistic Regression at 78.83% (Figure 7) and Naïve Bayes at 79.28% (Figure 5). Similar pattern indicated for average True Positive rate (TP rate) and Precision for all machine learning model.

The Receiver Operating Characteristic (ROC) area result also indicate best performance for Random Forest Classification compared to the other two. Random Forest and Logistic Regression also indicate significant increase in performance for average TP rate and Precision using LOOCV compared to 10-fold cross validation. However, using 10-fold cross validation, the ROC area results for Random Forest Classification and Naïve Bayes indicate better performance. This approach could be utilized whenever time of processing came into consideration whereby number on instances increased. In example to predict support tendency at Dewan Undangan Negeri (DUN), Daerah Mengundi (DM) or locality level.

For parliamentary level, the best machine learning model to be implemented as per result comparison in Table 1 and Table 2 is Random Forest Classification. The processing time however is slower compared to Naive Bayes and faster compared to Logistic Regression for both 10-fold cross validation and LOOCV. The least performed model is Naive Bayes; however, the processing time is fastest compared to others and on top of that, this model produces quite impressive average ROC area result with just 0.028 difference with the best performing model. The Naïve Bayes model with 10-fold cross validation could be utilized if there is requirement to evaluate tendency of support at micro-level whereby number of instances and features increased and processing time will be tremendously increase if LOOCV approach being implement.

The Logistic Regression algorithm shows that it comparatively performed against Naive Bayes but has disadvantage on the processing time. All the algorithms show better performance when evaluated using the LOOCV approach. Since the size of data is small, the processing time for both 10-fold cross validation and LOOCV is the same for Naive Bayes and Logistic Regression algorithms. However, the processing time for Random Forest Classification slightly increased on the implementation of LOOCV, though this can be neglected as the increased is not significant.

4. Conclusion

The research outcome indicates each of the classification models have their own strength, however in terms of accuracy, Random Forest Classifier reflects the best performance for both k fold cross validation and LOOCV. The LOOCV evaluator increases the performance by 2.7%, therefore this model is proposed to be used together with LOOCV evaluator for the purpose of GE-15 prediction. Though, if there is a requirement to deep analyze further to the locality level which tremendously increases size of data, the k-fold approach will be utilized for the sake of processing speed. The research outcome also indicates, the performance of Naive Bayes algorithms reduced by 0.45% when LOOCV evaluator implemented. Therefore, improvement of machine learning models does not necessarily apply to all algorithms when LOOCV evaluator is used. This study can be extend in the future by integrating other independent data such as estimated percentage of B40, public utilities availability, frequency of political activities from both government or opposition within the constituency, frequency of sentiment related news and frequency of issues raised to improve model accuracy. The same structure can also be implemented at DUN level for state election, DM and locality.

5. Acknowledgement

Data used for this research obtained from various online sources. Some of these data had already been taken out from its original link. Though, this information is important in the development of data model. We acknowledge efforts from organization like Tindak Malaysia in sharing elections related data to the public via their GitHub repository.

6. References

6.1. Journal Article

[1] Jaidka, Kokil, et al. "Predicting elections from social media: a three-country, three-method comparative study." Asian Journal of Communication 29.3 (2019): 252-273.

[2] Chauhan, Priyavrat, Nonita Sharma, and Geeta Sikka. "The emergence of social media data and sentiment analysis in election prediction." Journal of Ambient Intelligence and Humanized Computing (2020): 1-27.

[3] Kamakura, Wagner A., José Afonso Mazzon, and Arnaud De Bruyn. "Modeling voter choice to predict the final outcome of two-stage elections." International Journal of Forecasting 22.4 (2006): 689-706.

[4] Zhang, Harry. "The Optimality of Naive Bayes, 2004." American Association for Artificial Intelligence (www. aaai. org) (2004).

[5] Ekinci, E. Omurca, and N. Acun. "A comparative study on machine learning techniques using Titanic dataset." 7th international conference on advanced technologies. 2018.

[6] Rodriguez-Galiano, Victor Francisco, et al. "An assessment of the effectiveness of a random forest classifier for land-cover classification." ISPRS Journal of Photogrammetry and Remote Sensing 67 (2012): 93-104.

6.2. Website

[1] https://www.kaggle.com/terenctb/malaysia-ge14-election-results-parliament
[2] https://daneshtindak.carto.com/tables/malaysia_parliamentary_carto_2018/
[3] https://election.thestar.com.my/
[4] https://github.com/khoo-j/MsiaGE14/blob/master/GE14_Age-Ethnicity-bySeats.xlsx