# Handling Imbalanced Class Problem of Measles InfectionRisk Prediction Model

**Wan MuhamadTaufik Wan Ahmad,NurLailaAb Ghani,Sulfeeza Mohd Drus**

***Abstract:*** ***Measles is an emerging infectious disease with increasing number of reported cases. It is a vaccine-preventable disease;thus, it is common to have imbalanced class problem in the dataset. This study aims to resolve the imbalanced class problem for the prediction of measles infection risk and to compare the predictive results on a balanced dataset based on three machine learningtechniques. The data that was utilized in this study contained 37,884 records of suspected measles casesthat were highly imbalanced towards negative measles cases. The Synthetic Minority Over-Sampling Technique (SMOTE) was performed to balance thedistribution of the target attribute. The balanced dataset was then modelled using logistic regression, decision tree and Naïve Bayes. The predicted results indicated that logistic regression executed on the balanced dataset by SMOTE has the highest and most accurateclassification with 94.5% overall accuracy, 93.9% true positive rate, 5.8% false positive rate and 5.1% false negative rate. Therefore, SMOTE and other over-sampling approaches may be applicable to overcome imbalanced class issues in the medical dataset.***

***Keywords: Data Mining, Classification, Measles, Imbalanced Data***

## I.INTRODUCTION

Measles is an acute respiratory infectious disease commonly diagnosed among young children under five years old.It is transmitted through direct contact with an infected person or the air when the infected person coughs or sneezes. The clinical diagnosis of measles is determined by fever, the appearance of the maculopapular rash, together with any symptoms of cough, runny nose or conjunctivitis. Severe complications induced by measles include pneumonia, seizures, brain damage and death.Statistics from the World Health Organization had shownthat around 2.6 million death occurred in 1980 when the measles vaccine was newly introduced during that period. The introduction of measles vaccine has proved its impact on reducing measles infection as in 2016, the number of reported measles death cases is less than 100,000 cases which is the lowest in history since 1980.

The effectiveness of this vaccine through immunization programs conducted by most of the countrieshas motivated them to expand andincrease their immunization coverage.

However,as stated by the World Health Organization, measles is still ranked first among the diseaseswhich cause death to young children even though the immunization program is stillrunning

The number of measles cases in Malaysia hasbeen reportedly increasing since 2015 (Abdullah et al., 2018).The increase of cases for this past several years can be analyzed to understand its pattern in order to ensure further decision can be made more efficient.In Malaysia, the data on measles cases are stored in a measles surveillance system of Malaysia Health Ministry. This secondary data can be processed and analyzed using machine learningtechniques that can provide deeper insights on measles pattern and individuals who have a higher risk of measles infection.

Machine learning is a part of the data mining process capable of retrieving relationships or patterns from complex datasets. It is often divided into two approaches which are supervised and unsupervised learning. Supervised learning algorithms learn from the input data that already consists oftargeted and correct output. It intends toprovide concise model consisting of all the class labels distributed based on predictor features. For unsupervised learning, the algorithms learn from the data without being explicitly ask onthe desired output and without labeled examples, hence it randomly discover any pattern that groupthe input based on their characteristics or similarities.

The huge amount of complex data related to infectious disease has increased the need of processing and interpreting the data using machine learning (Kaye et al., 2015). A lot of challenges need to be addressed to understand how the disease spread and who will likely get infected. Even though clinical definitionprovidesymptoms of a particular disease, the laboratoryresult can be totally different from earlier clinical diagnosis.This is due to thesimilarity of symptoms between diseases lead to the gap in treatment procedure (Schaepe, 2011).The role of machine learning is to mine and analyzedhuge amount of data, so that a precise pattern can be derived to assist physicians in decisionmaking.

A number of studies related to machine learning modelling of infectious disease have been conducted on unsupervised analysis of several infectious diseases (PeterIdowu *et al.*, 2013),measles outbreak prediction (Liao et al., 2017), dengue outbreak prediction (Rahmawati & Huang, 2016) and dengue infection risk (Fathima & Hundewale, 2012). Our prior work found that decision tree and Naive Bayes are among the techniques commonly used for disease risk prediction (Ahmad et al., 2018).

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

**Wan MuhamadTaufik Wan Ahmad*,** College of Computing and Informatics, Universiti TenagaNasional, Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia

**NurLailaAb Ghani,** College of Computing and Informatics, Universiti TenagaNasional, Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia

**Sulfeeza Mohd Drus,** College of Computing and Informatics, Universiti TenagaNasional, Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia

Imbalanced classification is one of the challenges in machine learning that refers to significantly different number of examples among classes in a training dataset (Buda et al., 2018). It is a common problem on medical diagnosis where instances belonging to the positive or minority class are less in comparison to majority or negative class (Jain et al. 2017).Sampling strategies have been used to overcome the imbalanced class problem by performing under-sampling or over-sampling on the dataset. Under-sampling works by eliminating some data from the majority class while over-sampling adds artificially generated data to the minority class (Rahman & Davis, 2013). The Synthetic Minority Over-Sampling Technique (SMOTE) is one of the most popular advanced oversampling strategies (Chawla et al. 2002).

This study aims to resolve the imbalanced classproblem for the prediction of measles infectionriskand to compare the predictive results of several machine learning techniques.In this paper, Section 2 detailed the data and methodology of the study. Section 3 presented the findings, and Section 4 concluded the paper.

## II.MATERIALS AND METHODS

This study retrieved 37,884 recordsand 49 variables of suspected measles cases in Malaysia between January 2010 and December 2016. This data is secondary data acquired from the database of measles surveillance system, Ministry of Health Malaysia through formal application under a research grant. This data is the most reliable and complete data that records measles incidents in Malaysia. Table 1 tabulated the details of the target and predictor attributes that were employed in this study after going through consultation with the domain expert and data pre-processing stage that excludes some of initial variables. The data processing stage is conducted by removing variables with no records and also having more than fifty percent missing records. This is to ensure the data presented really represent the information of measle incidents. The selected attributes consisted of the age, gender, nationality, vaccination status, clinical symptoms and laboratory result of the individualsuspected of having measles. The attributes also specified the state where the case was reported, the type of case as either local or imported and its outbreak association to check whether the case was reported during an outbreak or otherwise.

**Table. 1 Definition of the variables employed in the study**

| Attribute Name | Data Type | Values | Role |
|---|---|---|---|
| **Age** | Continuous | Positive real number | Predictor |
| **Gender** | Discrete | Female/Male | Predictor |
| **Nationality** | Discrete | Malaysian/Non-Malaysian | Predictor |
| **Vaccination Status** | Discrete | No/Not qualified/Unknown/Yes | Predictor |
| **State** | Discrete | Johor/Kedah/Kelantan/Melaka/Negeri Sembilan/Pahang/Perak/Perlis/Pulau Pinang/Sabah/Sarawak/Selangor/Terengganu/Wilayah Persekutuan Kuala Lumpur/Wilayah Persekutuan Labuan | Predictor |
| **Case Type** | Discrete | Endemic/Import-related/Imported Case/Local Case/Unclassified | Predictor |
| **Outbreak Association** | Discrete | No/Yes | Predictor |
| **Fever** | Discrete | No/Yes | Predictor |
| **Cough** | Discrete | No/Yes | Predictor |
| **Coryza** | Discrete | No/Yes | Predictor |
| **Conjunctivitis** | Discrete | No/Yes | Predictor |
| **Lymphadenopathy** | Discrete | No/Yes | Predictor |
| **lgm_Rubella** | Discrete | Equivocal/Negative/Not Done/Pending/Positive/Rejected | Predictor |
| **1st_lgm_Result** | Discrete | Equivocal/Negative/No Specimen/Not Done/Pending/Positive/Rejected | Predictor |
| **Case Classification** | Discrete | Measles/Not Measles | Target |

### Imbalanced Class Solution using Synthetic Minority Over-Sampling Technique (SMOTE)

Imbalanced class issues must be solved before a dataset can be used for modelling. The particular dataset utilized in this study was highly imbalanced in terms of the number of*'Measles'* and *'Not Measles'* classes.This dataset can be executed on a prediction model; however, the result could lead to bias on majority class. If this happens, the developed model could producea false sense of accuracy.

This problem is common in the healthcare domain, however at the same time is a major challenge for machine learningto producea good model. To overcome this, various techniques have been introduced that include random over-sampling and random under-sampling.

Random over-sampling works by increasing the size of minority class while random under-sampling works by reducing the majority class. Both of this techniques aim to produce a well-balanced class data set in order to achieve accurate result. However, random over-sampling is more frequently being used compared to random under-sampling because random under-sampling cause the lost of important information due to a lot of data being removed from the majority class(Santoso et al., 2017). This will reducethe learning capability of the prediction modeldue to less number of information provided.

In this study, we have applied a random over-sampling technique called SMOTE to achieve a class-balanced or almost class-balanced dataset. SMOTE is different from other random over-sampling as the class of minority is increased with a variety of training examples. It does not only increase the size of the training data but also the training examples variety. It works by using *k*-nearest neighbors methodology, creating examples based on what it learns from the *k*-nearest class (Chawla et al. 2002). Hence, the new data created will be more likely to represent minority class without duplicating them. In one study, SMOTE produces higher classification rate when applied using Support Vector Machine (SVM) algorithm as compared to normal random over-sampling and random under-sampling for liver data set (Lokanayaki& Malathi, 2014).

As a comparison, other normal random over-sampling techniques increase the minority class simply by duplicating instances to achieve balance ratio. The shortcoming of this method is no new information added and it could create overfitting for classifiers (Zheng et al., 2016).This study employed the *DMwR*package of R to implement SMOTE with the number of nearest neighbors taken for consideration in creating a new example is 5.

### Modelling

The prediction model was developed on the balanced dataset created using SMOTE by incorporating logistic regression, decision tree and Naïve Bayestechniques. For logistic regression, it is essential to identify the optimalcut off value that will produce a model with the highest accuracy. The model was tested with a range of cut off values in R.The R package named *rpart* was used to perform decision tree modelling. The accuracy was based on default prune of the model that decides the number of nodes. This number of nodes can be altered to see the changes in the accuracy using a method called pruning. In *rpart*package, the number of nodes is related to the complexity parameter or known as CP. CP is used to control the size of the tree and to determine the most optimal size of the tree.For Naïve Bayes, the conditional probability of having measles was calculated based on Bayesian theorem.

### Performance Evaluation

The split validation technique with 80:20 ratio of train and test datawas employedfor evaluating the models and obtaining unbiased results from the models. This ratio is applied to see the outcome, but another ratio could be applied as there is no rule of thumb for deciding absolute ratio.Several performance measurements which are the accuracy, true positive rate and false positive rate were calculated based on the formula given in Table 2.

**Table. 2 Performance measurement formula.**

| Performance Metrics | Description | Formula |
|---|---|---|
| Accuracy | Proportion of the total number of correct predictions | $\frac{(TP+TN)}{(TP+TN+FN+FP)}$ |
| True Positive Rate | Proportion of actual positive cases that are correctly identified | $\frac{TP}{(TP + FN)}$ |
| False Positive Rate | Proportion of actual negative cases that are incorrectly identified | $\frac{FP}{(TN +FP)}$ |
| False Negative Rate | Proportion of actual positive cases that are incorrectly identified | $\frac{FN}{(TP + FN)}$ |

**TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative*

## III.RESULTS AND DISCUSSIONS

In this study, there were 37,884 casesreported between January 2010 and January 2017. The age range of most positive measles cases were those below ten years old. The issue with the dataset was on the number of *'Not Measles'* class that was significantly higher than *'Measles'* class. There were 5,918 records of positive measles cases and 29,970 records of negative measles cases.The most left side of the plotted visual in Figure 1 showed that the target variable was highly imbalanced, whereby 80 percent of the class is for the*'Not Measles'* class. The ratio of *'Measles'* and *'Not Measles'*classes in this data set was 1:5.

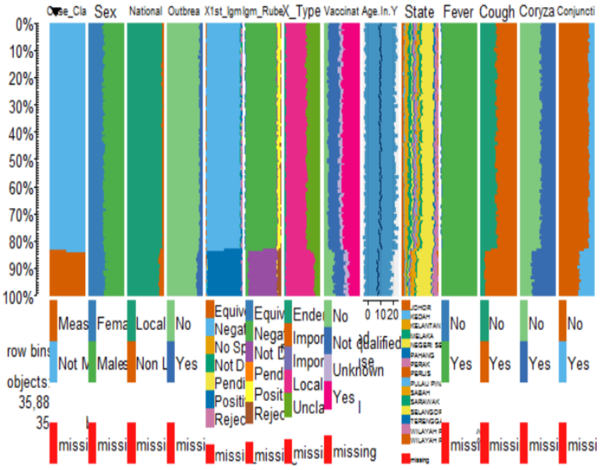

**Fig. 1 Plotting of Attributes and Distribution of Records**

As shown in Figure 2, the application of SMOTE on the dataset has reduced the difference between *'Measles'* and *'Not Measles'* classes from 24,052 to only 8,700.

Retrieval Number: A2649109119/2019©BEIESP
DOI: 10.35940/ijeat.A2649.109119
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

As shown in the most left side of the plotted visual in Figure 2, the target variable has been balanced with the ratio of 1:1 for both classes. This balanced dataset was utilized to predict the risk of measles infection among suspected individuals. The prediction was modelled based on logistic regression, decision tree and Naïve Bayes.

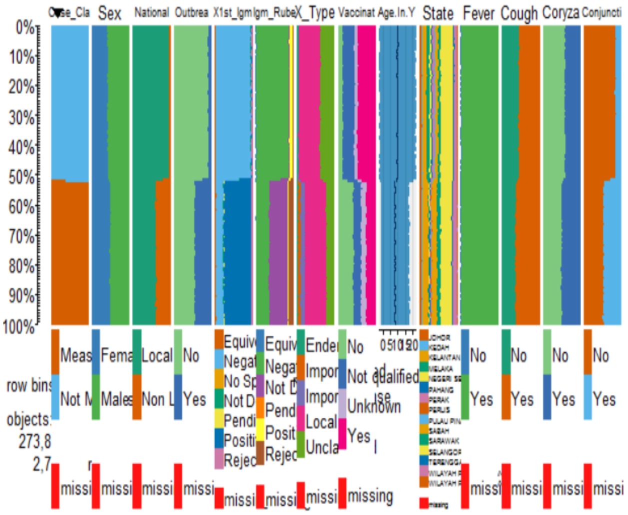

**Fig. 2 Plotting of Attributes and Distribution of Records After SMOTE**

Several experiments were conducted to find a logistic regression model with the optimal cut off value. In real medical cases, a model with good true positive rate and less false negative error is essential as a false positive error is still tolerable compared to false negative error. In this study, a cut off value of 0.7 provided a well-balanced model that can correctly classify positive cases and avoid misclassification of measles cases.

In decision tree modelling, several values of the CPwere tested to see its effect on the accuracy of the model. In this study, theCP of 0.01, tree size of 3 and tree split of 2 produced an optimal decision tree model with the highest accuracy. Other CP values were unable to produce better accuracy. The decrease in the size of tree to two using CP value of 0.26 caused reduce on the overall accuracy.

**Table. 3 Performance comparison of logistic regression, decision tree and Naïve Bayes.**

| Performance Metrics | Logistic Regression | Decision Tree | Naïve Bayes |
|---|---|---|---|
| **Overall Accuracy (%)** | 94.50 | 93.74 | 94.59 |
| **True Positive Rate (%)** | 93.90 | 95.42 | 92.63 |
| **False Positive Rate (%)** | 5.80 | 4.10 | 3.60 |
| **False Negative Rate (%)** | 5.10 | 9.68 | 7.36 |

Based on Table 3, logistic regression produced the well-balanced predictive model as compared to Naïve Bayes and decision tree. Logistic regression has the second highest overall accuracy and the least false negative rate. Even though Naïve Bayes has the highest overall accuracy, its false negative rate is second highest, and its true positive rate is the lowest. Decision tree produced the highest true positive rate. However, its false negative rate is too high which indicates that the model is unstable for this study.

## IV.CONCLUSION

This study investigated and resolved the imbalanced class problem based on SMOTE for the prediction of measles infection risk. As measles is a vaccine-preventable disease, the imbalanced class problem can arise in medical researches. The dataset used in this study was highly imbalanced with *'Measles'* as the minority class. An over-sampling method of SMOTE was performed on the dataset, and three models were built to compare the predictive results. The results from this study indicated that the logistic regression model executed on the balanced dataset by SMOTE produced the highest and most accurate values of performance measurements. Ourresults suggest that SMOTE and other over-samplingapproaches would be beneficial to overcome classimbalanced issues emerging in medical studies.

## ACKNOWLEDGEMENT

This work was supported by the Ministry of Higher Education, Malaysia under Fundamental Research Grant Scheme (FRGS/1/2016/ICT04/UNITEN/03/1) and UNITEN BOLD Grant (10436494/B/2019118).

## REFERENCES

1. Abdullah, A. C., MZ, N. A., & Rosliza, A. M. (2018). Predictors For Inadequate Knowledge And Negative Attitude Towards Childhood Immunization Among Parents In Hulu Langat, Selangor, Malaysia. Malaysian Journal of Public Health Medicine. Vol. 18 (1), 102-112.
2. Buda, M., Maki, A., & Mazurowski, M. A. (2018). A Systematic Study Of The Class Imbalance Problem In Convolutional Neural Networks. Neural Networks, 106, 249-259.
3. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-Sampling Technique. Journal of artificial intelligence research, *16*, 321-357.
4. Chinnayah, T. (2017a). Elımınatıon Of Measles İn Malaysıa By 2018: How Close Are We?. Epidemiology, 7, p. 5. doi: 10.4172/2161-1165-C1-018.
5. Fathima, Shameem A., and Nisar Hundewale. (2012). Comparitive Analysis of Machine Learning Techniques for Classification of Arbovirus. Proceedings - IEEE-EMBS International Conference on Biomedical and Health Informatics: Global Grand Challenge of Health Informatics, BHI 2012 25(Bhi): 376–79.
6. Jain, A., Ratnoo, S., & Kumar, D. (2017). Addressing Class Imbalance Problem In Medical Diagnosis: A Genetic Algorithm Approach. In 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC) (pp. 1-8). IEEE.
7. Kaye, K. S. *et al.* (2015). Guidance For Infection Prevention And Healthcare Epidemiology Programs:Healthcare Epidemiologist Skills And Competencies. Infection Control & Hospital Epidemiology. 36(04),pp. 369–380. doi: 10.1017/ice.2014.79.
8. Liao, Y. *et al.* (2017). A New Method For Assessing The Risk Of Infectious Disease Outbreak. ScientificReports. Nature Publishing Group, 7, p. 40084. doi: 10.1038/srep40084.
9. Lokanayaki K. & Malathi A. (2014). A Prediction for Classification of Highly Imbalanced Medical Dataset using Databoost.IM with SVM. International Journal of Advanced Research in Computer Science and Software Engineering, 4 (4), 276-281.

10. PeterIdowu, A. *et al.* (2013). Data Mining Techniques For Predicting Immunize-Able Diseases: Nigeria AsA Case Study. International Journal of Applied Information Systems. 5(7), pp. 5–15. doi: 10.5120/ijais12-450882.
11. Rahman, M. M., & Davis, D. N. (2013). Addressing The Class Imbalance Problem In Medical Datasets. International Journal of Machine Learning and Computing, 3(2), 224.
12. Rahmawati, D., & Huang, Y. P. (2016). Using C-support vector classification to forecast dengue fever epidemics in Taiwan. In 2016 International Conference on System Science and Engineering (ICSSE) (pp. 1-4). IEEE.
13. Santoso, B., Wijayanto, H., Notodiputro, K. A., & Sartono, B. (2017, March). Synthetic over sampling methods for handling class imbalanced problems: A review. In *IOP Conference Series: Earth and Environmental Science* (Vol. 58, No. 1, p. 012031). IOP Publishing.
14. Schaepe, K. S. (2011). Bad News And First Impressions: Patient And Family Caregiver Accounts Of LearningThe Cancer Diagnosis. Social Science & Medicine. 73(6), pp. 912–921. doi: 10.1016/j.socscimed.2011.06.038.
15. Ahmad, W. M. T. W., Ab Ghani, N. L., & Drus, S. M. (2018). Data Mining Techniques for Disease Risk Prediction Model: A Systematic Literature Review. In International Conference of Reliable Information and Communication Technology (pp. 40-46). Springer, Cham.
16. Zheng, Z., Cai, Y., & Li, Y. (2016). Oversampling method for imbalanced classification. *Computing and Informatics*, *34*(5), 1017-1037.