Evaluation of Applications for Nursery School using Machine Learning Approach

Mohd Hashim Ashaari
Jabatan Pendidikan Politeknik dan Kolej Komuniti
hashim.ashaari@mohe.gov.my

Mohd Gadaffi Osman
Politeknik Sultan Mizan Zainal Abidin
gadaffi@psmza.edu.my

Abdul Hadi Mustaffa
Jabatan Pendidikan Politeknik dan Kolej Komuniti
hadi.mustaffa@mohe.gov.my

Abstract: This Paper demonstrates how several classification techniques can be used to perform classify nursery dataset. There are various classification techniques that can be used to classify dataset. Therefore, the need to choose the right classification technique can affect the result of classification accuracy of the model. Naïve Bayes, Support Vector Machine and k-Nearest Neighbours are three machine learning models that used in this study to evaluate the application of nursery schools. This paper is aimed to find the best machine learning models that can be used to classify the nursery dataset. The pre-processing tasks that have been conducted in this study include data cleaning, attribute or feature reduction, feature engineering technique and dimensional reduction. The data that has been undergoing pre-processed was then used in the third stage to identify the best machine learning technique by develop a comparative analysis between the three chosen techniques. Outcome of this study demonstrates the accuracy of each classification techniques and the accuracy result of three classification models were compared before and after tuning parameter was conducted to determine which classification technique is the best to classify the nursery dataset. The impact of the study can be used as a reference to develop classification model for this kind dataset.

Keywords: classification technique, machine learning models, tuning parameter

1.0 Introduction

Machine learning is the progressions of artificial intelligent programs that enable the programs learn without human interference or clear direction using specific algorithms to analyse, observe and recognize the patterns of certain dataset. By observing the data that have supplied to the program, their learning over time will be improved. The prediction target is to accurately classify the result based on sensible relationships from any prepared data (Sani et al, 2018).

Supervised and unsupervised learning are two learning algorithm method that used in machine learning. In supervised learning, the machine was train using data that have been labels or known data. The machine will learn from feed training data and discover the pattern of the data (Osisanwoi et al, 2017). Then new data or unforeseen data was feed to the built model to predict outcomes for the new data. Classification problem normally uses supervised learning normally method to solve the problem. Different with unsupervised learning, the machine will work by itself to discover the

important information from the data. The unsupervised learning was usually implemented for unlabelled data or unknown data.

The analysis for complex and large datasets can be performed using machine learning. Many of different tasks can be automate using machine learning such as handwritten and image recognition, classification problems, fraud detection, cancer detection and weather forecasting.

In this paper the nursery dataset was process using different classification method to develop and select the most suitable algorithm that can be used to processing the nursery dataset. The objective of this study nursery dataset was to evaluate the applications of nursery schools whether it is recommended, not recommend, very recommend priority or specific priority using machine learning approach from the dataset that had been obtained from UCI Machine Learning Repository. The dataset has 12960 instances and eight attributes with five possible classes. The eight class are parent's occupation which are; usual, pretentious or great pretentious, form of the family which are; complete, completed, incomplete or foster, child nursery condition which are; proper, less proper, improper, critical, very critical, number of children which are; 1, 2, 3 or more than 3, housing condition which are; convenient, less convenience or critical, financial standing of the family which are; convenient or inconvenient, health conditions which are; recommended, priority or not recommended and social conditions which are; non problematic, slightly problematic or problematic.

2.0 Related Work

Recently series studies have been introduced for classify of nursery schools' application. For example, Asmita Singh, Malka N. Halgamuge and Rajasekaran Lakshmiganthan has used a different kind of classification techniques to evaluate the nursery school's application. In their study, the researcher used Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms to solve the classification problem. The researchers found out that K-Nearest Neighbors perform very well compared to Naïve Bayes and Random Forest by giving result mean accuracy of 94% compared to the mean accuracy of Naïve Bayes which around 31% (Singh et al, 2018).

Another study has been conducted by Pardeep Kumar, Nitin, Vivek Kumar Sehgal and Durg Singh Chauhan used different algorithm which are CHAID, QUEST, C4.5, Neural Network, Logistic Regression, k-means, genetic algorithm and SVM algorithms. The result from their study shows k-means algorithm can perform very well by giving 100% accuracy to solve the classification problems and SVM perform low by giving 55% accuracy result compared to other techniques (Kumar et al, 2018).

3.0 Methodology

The research methodology of this study is divided into three stages. The first stages start by understanding the data and analysing the quantity, details, class and attributes. The second stage was followed by pre-processing task to prepare the data for processing method. The pre-processing tasks that have been conducted in this study include data cleaning, attribute or feature reduction, feature engineering technique and dimensional reduction. The

data that has been undergoing pre-processed was then used in the third stage to identify the best machine learning technique by develop a comparative analysis between the three chosen techniques. Lastly the accuracy result of three classification models were compared before and after tuning parameter was conducted.

3.1 Pre-Processing

Data pre-processing is an important technique to transform a raw dataset into data form or format that can be understand by machine or tools for further processing. Real-world data is usually always inconsistent, lacking, incomplete and likely to consist of many errors, noise and outlier. Therefore, to ensure the data in understandable format and suitable form to further processing, data pre-processing must be done to overcome the problems (Sani et al, 2018). The pre-processing tasks that have been conducted in this study include data cleaning, attribute or feature reduction, feature engineering technique and dimensional reduction. Now days, there are a lot of data mining tools that can be used for data pre-processing task. In this study, the 'Jupyter Notebook' and 'Weka' version 3.8 was used as a tool to perform the pre-processing task and to perform classification task. Jupyter Notebook was machine learning software that runs using Python Language which developed by non-profit organization. It contains numerous types of machine learning algorithms and visualization tools to perform data pre-processing, data analysis and develop predictive modelling for the study.

3.1.1 Data Cleaning

Before conducting data cleaning process and to better understand of the data, data exploration and visualization was plot and explore in Weka version 3.8. Figure 1 shows data visualization for all the attributes in the Nursery dataset in a graphical view.

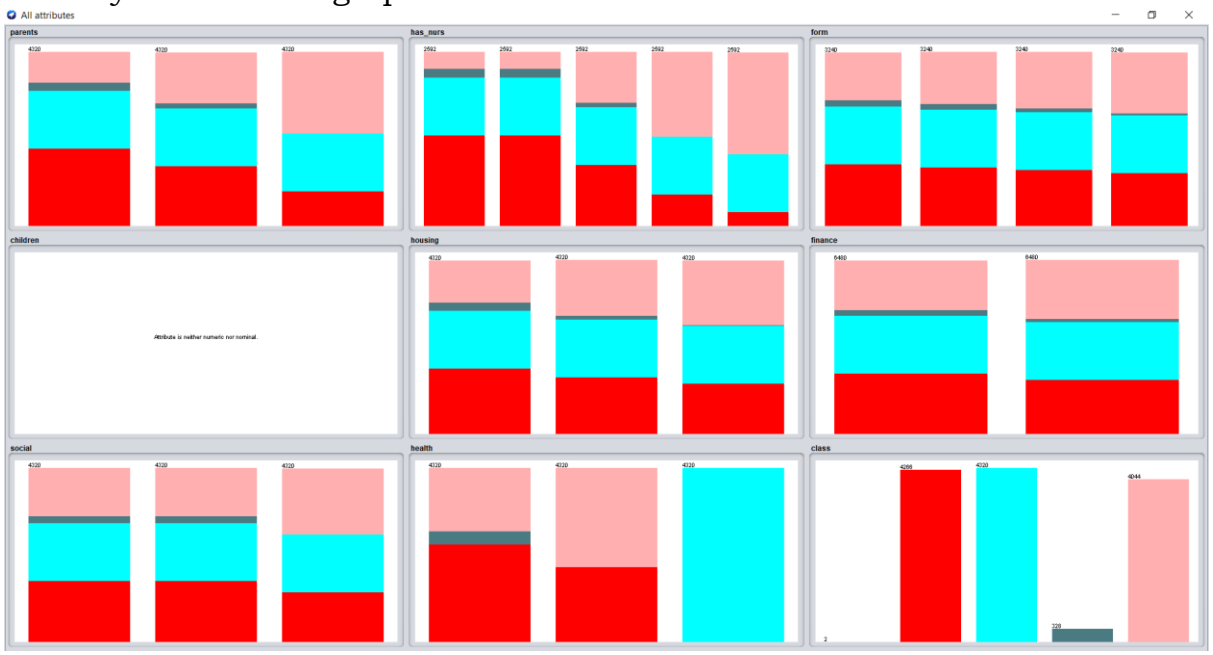

Figure 1: Data Visualization for All Attributes (Nursery Dataset) in Weka

From the data exploration and checking in Weka, there no missing value for all attributes in the nursery dataset. The nursery dataset also did not contain any outlier after conducted filter technique using Interquatile Range in Weka. An outlier refers to instances of datasets that diverge from other inspections, probably cause improper data collection (Sani et al, 2018).

3.1.2 Attribute or Feature Reduction

Zipping method and correlation test was conducted to observe the relationship between each attribute with class. From the task we can see that most of the attributes are dispersed and contributing to the dataset except "finance" attribute which has almost all equal sectors in the chart. Since finance attributes not dispersed, it is not contributing much to the class and finance attribute was drop.

3.1.3 Feature Engineering

The dataset was converted from the form of strings to numbers using categorical codes and dummy values. Then, the attributes of the dataset was converted into categorical values and some attributes into binary values. Those attribute whose categories have somewhat equal priority are divided using get_dummies and the rest are divided using categorical values. This process was conducted in Jupyter Notebook.

3.1.4 Dimensional Reduction

Since the dataset have 14 dimensions after conducted feature engineering, Principal Component Analysis (PCA) was used to do dimensionality reduction. PCA uses refined underlying mathematical principles reduced the possibly correlated attributes into a lesser number of attributes called principal components by finding the correlation between all attributes. After applying PCA, the dataset has been reduced to 8 dimensions. The dataset was converted from the form of strings to numbers using categorical codes and dummy values. Then, the attributes of the dataset were converted into categorical values and some attributes into binary values. Those attribute whose categories have somewhat equal priority are divided using get_dummies and the rest are divided using categorical values.
The attribute reduction, feature engineering and dimensional reduction task was conducted in Jupyter Notebook.

3.2 Classification Algorithms

3.2.1 Naïve Bayes

These are very simple probabilistic algorithm which applied Bayes' theorem that composed of directed acyclic graphs with a strong assumption of independence (Osisanwoi et al, 2017). The algorithms was name independence model based on estimating and naïve state (Sani et al, 2018). This algorithm frequently used for supervised learning method by assume the existence attribute of class does not relate to the existence of any other attributes. It can be train very can be trained very accurately depending on the correct probability model (Berend et al, 2018). Figure 2 shows the Naïve Bayes Model.

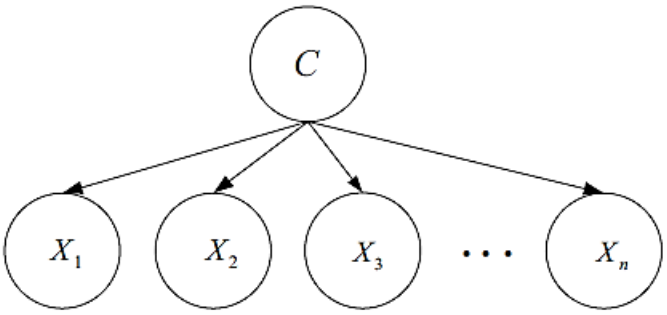

Figure 2: Naïve Bayes Model (Taheri et al, 2013)

3.2.2 Support Vector Machine

Support Vector Machine is the latest supervised machine learning technique mostly used in classification and regression problems analysis (Osisanwoi et al, 2017). These models are closely like neural networks specifically classical multilayer perceptron (MLP). SVM create one or more than one hyperplane that separates the class of the data in classification analysis. To reduce the error of classification analysis using SVM, the distance and the margin between the separating hyperplane and the instances on either side of it must be maximize (Durgesh et al, 2018). The concept of SVM can be shown in Figure 3.

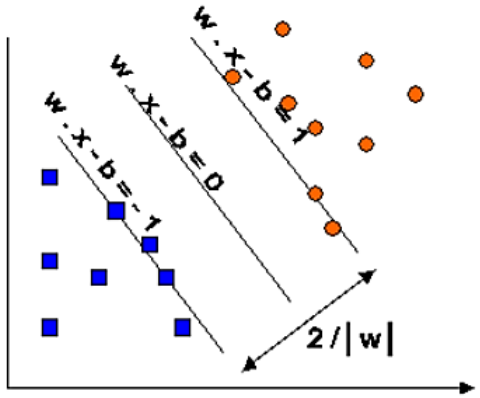

Figure 3: SVM Model (Durgesh et al, 2018)

3.3.3 k-Nearest Neighbors

The k-Nearest Neighbors (KNN) is another simple supervised learning algorithms for classification analysis. This classification algorithm operate by predicts the nearest neighbours of test samples based on to the K training samples by determine which has the largest category probability (Suguna et al, 2018). KNN has proved with very good performance in the classification analysis of different datasets. The concept of SVM can be shown in Figure 4.

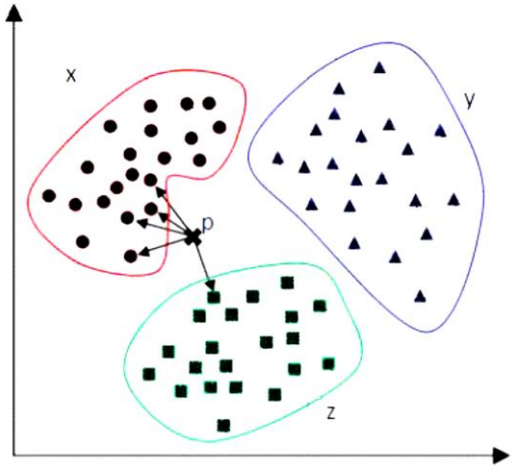

Figure 4: An example of K-NN classifier (Suguna et al, 2018)

3.4 Machine Learning Task

In this study, Naïve Bayes, Support Vector Machine and k-Nearest Neighbors algorithms was compared to analyse the nursery dataset. The dataset was divided using percentage split which is 80% for training and 20% for test data. Table 1 shows the classification accuracy of three classification method that has been selected to solve the classification problem of nursery dataset before conducted parameter tuning.

Table 1: Classification accuracy before conduct tuning parameter

No.	Classifier	Accuracy (%)
1	Naïve Bayes	68
2	SVM	92
3	k-Nearest Neighbors	92

To produce better accuracy result, tuning parameters is needed for each classifier. To get the optimum result of accuracy for each classifier, set of experiment was conducted with different tuning parameters. Then, the performance of classifiers was evaluated and compare to the previous result. Naïve Bayes classifier has been tuned using cross-validation method. Cross-validation split the original dataset into two sets which are training set to train the model and the test to evaluate the model. SVM and KNN was tuned by tuning the hyper-parameters of on estimator to find the optimal hyper-parameters using grid search. Regularization technique or set penalty parameter was implemented during tuning parameters to overcome overfitting issues. Overfitting or modelling error is situation where the predictive model learns all the noise and target function during train the model, therefore this will be reducing the performance of that model on an undiscovered data (Sani et al, 2018).

4.0 Results And Discussion

This study was conducted by divided the dataset using percentage split which is 80% for training and 20% for test data. Table 2 shows the classification accuracy for Naïve Bayes classifier before and after conducted tuning parameter using 5 fold cross validation. The result shows that there is

no improvement in classification accuracy before and after conducting tuning parameter compared to others classifier (KNN and SVM).

Table 2: Naïve Bayes classification accuracy before and after tuning parameter

No.	Naïve Bayes Tuning Parameter	Accuracy (%)
1	Before Tuning	74
2	After Tuning – Cross-Validation	74

For k-Nearest Neighbors classifier, we can see that there is improvement in performance of the model after conducted tuning parameter using Grid Search Cross Validation Method by tuning the hyper-parameters. The accuracy has improved from 92% to 94% after conducted the tuning parameter as shown in Figure 3. From the KNN tuning process we best number of neighbours was 9 and best leaf size was 5.

Table 3: KNN classification accuracy before and after tuning parameter

No.	KNN Tuning Parameter	Accuracy (%)
1	Before Tuning	92
2	After Tuning Hyper-parameters	94

As shown in Table 4, after conducting tuning parameter the performance of the model was improved from 92% to 94% using Grid Search Cross Validation method.

Table 4: SVM classification accuracy before and after tuning parameter

No.	SVM Tuning Parameter	Accuracy (%)
1	Before Tuning	92
2	After Tuning Hyper-parameters	93

Table 5: Comparative Classification accuracy of three classifiers

No.	Classifier	Accuracy (%)
1	Naïve Bayes	74
2	k-Nearest Neighbors (KNN)	94
3	Support Vector Machine (SVM)	93

From Table 5, the highest accuracy to evaluate the applications of nursery schools can obtain by conducted experiment using k-Nearest Neighbors (KNN) classifier. The performance of Naïve Bayes algorithms for this

dataset is low compared to KNN and SVM classifier and the result also shows the best performer was KNN classifier with 94% accuracy.

The overall results from this study also show the performance of the models or algorithms can be improve by conducted tuning parameter for the models. Selected the suitable classifier and tuning parameter can affect the accuracy result of the experiment.

5.0 Conclusion

This study identifies performance comparison between three classifiers which is Naïve Bayes, Support Vector Machine and K-Nearest Neighbors in classifying evaluation of application to nursery schools using nursery dataset. The accuracy result of these three classification techniques was compared to each other to identify the best performer of classifier. Several pre-processing tasks has been conducted before performing the performance comparison such as data cleaning, attribute reduction, feature engineering and dimensional reduction using Principal Component Analysis (PCA).

The study shows the best classifier to classify nursery dataset was k-Nearest Neighbors (KNN) with the 94% accuracy The result also shows that, to get the optimum result of accuracy for each classifier tuning parameters can be implemented to increase the performance of classifiers compare to the classifier without parameter tuning. The result also indicate that Naïve Bayes technique perform not so good to classify the nursery dataset. However, this studies only clearly comparing the classifiers in term of accuracy. To re-analyse and extract useful knowledge from the nursery dataset other classification techniques such as Decision Tree, Artificial Neural Network and Random Forest classifiers which not named in this study could be used to in the analysis.

References

Sani, N. S., Rahman, M. A., Bakar, A. A., Sahran, S., & Sarim, H. M. (2018). Machine learning approach for bottom 40 percent households (B40) poverty classification. *Int. J. Adv. Sci. Eng. Inf. Technol*, 8(4-2), 1698.

Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3), 128-138.

Singh, A., & Lakshmiganthan, R. (2018). Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms.

Kumar, P., Sehgal, V. K., & Chauhan, D. S. (2012). A benchmark to select data mining based classification algorithms for business intelligence and decision support systems. arXiv preprint arXiv:1210.3139.

Berend, D., & Kontorovich, A. (2015). A finite sample analysis of the Naive Bayes classifier. J. Mach. Learn. Res., 16(1), 1519-1545.

Taheri, S., & Mammadov, M. (2013). Learning the naive Bayes classifier with optimization models. International Journal of Applied Mathematics and Computer Science, 23(4).

Durgesh, K. S., & Lekha, B. (2010). Data classification using support vector machine. Journal of theoretical and applied information technology, 12(1), 1-7.

N. Suguna & K. Thanushkodi (2018), *An Improved k-Nearest Neighbor Classification Using Genetic Algorithm, International Journal of Computer Science*, 7(4), 18-21, 2010.

Suguna, N., & Thanushkodi, K. (2010). An improved k-nearest neighbor classification using genetic algorithm. International Journal of Computer Science Issues, 7(2), 18-21.