IDENTIFYING IMPORTANT CHARACTERISTICS IN AVERAGE GRADE OF THE SCHOOL IN SELANGOR: A MACHINE LEARNING APPROACH

Faiza Rusrianti Tajul Arus[1] and Mohamad Nor Hakkim Nor Azlan[2]

ABSTRAK

Pembelajaran mesin memainkan peranan penting kepada Sains Data dengan menyediakan kemahiran dalam mengumpulkan data daripada pelbagai sumber dan menggunakan analisis yang sesuai untuk mengekstrak maklumat kritikal. Oleh itu, ia membantu Saintis Data atau Penganalisis Data untuk mengumpulkan maklumat masa nyata dan membuat keputusan berasaskan bukti melalui pembelajaran mesin, analitik ramalan dan analisis sentimen. Pada masa hadapan, pembelajaran mesin diunjurkan menjadi teknologi arus perdana dan mempunyai impak yang ketara merentas banyak industri di dunia. Selaras dengan kemunculan pembelajaran mesin, kertas kerja ini menyerlahkan bagaimana pembelajaran mesin berinovasi untuk analisis dan membuat keputusan yang lebih baik dalam pelaksanaan penarafan sekolah di Selangor dengan menggunakan skor Purata Gred Sekolah (Gred Purata Sekolah atau GPS) yang dibangunkan oleh Kementerian Pendidikan Malaysia. Selain itu, kertas kerja ini juga menunjukkan bagaimana algoritma pembelajaran mesin yang dilatih pada hanya 917 responden mencukupi untuk menganalisis GPS dengan ketepatan antara 71.5% hingga 87.3% di Negeri Selangor. Keputusan menunjukkan model Regresi Logistik membuat ramalan yang paling tepat selain model regresi. Justeru, kajian ini mencadangkan untuk menggunakan pendekatan pembelajaran mesin bagi mengenal pasti pembolehubah penting dalam pembentukan GPS. Sumbangan utama kajian ini adalah untuk mengintegrasikan ciri-ciri sosioekonomi ke dalam tafsiran yang lebih baik terhadap keputusan ranking sekolah di Selangor.

Kata kunci: Pembelajaran Mesin, Regresi Logistik, Gred Purata Sekolah

ABSTRACT

Machine learning plays a vital role to a Data Science by providing skills in gathering data from multiple sources and applying appropriate analysis to extract critical information. Thus, its helps the Data Scientist or Data Analyst to gather real time information and to make evidence-based decision making through machine learning, predictive analytics and sentiment analysis. In the future, machine learning is projected to become mainstream technologies and will have a significant impact across many industries in the world. In line with the emergence of machine learning, this paper highlights how machine learning innovates for better analysis and decision

[1] Faiza Rusrianti Tajul Arus is currently Senior Assistant Director of Core Team Big Data Analytics (CTADR), Department of Statistics Malaysia and [2] Mohamad Nor Hakkim Nor Azlan is currently Data Scientist of Keysight Tecnologies Sdn. Bhd.

making in the implementation of the school rankings in Selangor by using the Average Grade of The School (Gred Purata Sekolah or GPS) score developed by Malaysia's Ministry of Education. Besides, the paper also shows how machine learning algorithms trained on only 917 respondents that are sufficient to analyze the GPS with an accuracy ranging from 71.5% to 87.3% in the State of Selangor. The results showed that Logistic Regression model made the most accurate prediction other than regression models. Thus, this study proposed to use a machine learning approach to identify important variables in the formation of GPS. The main contribution of this study is to integrate socioeconomic characteristics into a better interpretation of the school ranking result in Selangor.

Keywords: Machine Learning, Logistic Regression, Average Grade of the School

1. INTRODUCTION

1.1 Education Systems in Malaysia

Education in Malaysia aims to develop the potential of individuals in a holistic and integrated manner. Malaysia plan to produce individuals who are intellectually, spiritually, emotionally and physically balanced and harmonious, based on a firm belief in and devotion to God. This effort is designed to produce Malaysian citizens who are knowledgeable and competent, with high moral standards. They should also be responsible and capable of achieving a high level of personal well-being, contributing to the harmony and betterment of the family, the society and the nation.

In Malaysia, primary school is compulsory which starts at the age of seven. The children will spend 6 years in primary schools and in the 6th year, they will sit for a national standardised test known as the Ujian Pencapaian Sekolah Rendah (UPSR, Primary School Achievement Test). Public primary schools are divided into two categories namely Malay-medium National Schools (Sekolah Kebangsaan, SK) and non-Malay-medium National-type Schools (Sekolah Jenis Kebangsaan, SJK), the SJK then divided into two which is National-type School (Chinese) (Sekolah Jenis Kebangsaan (Cina), SJK(C)), Mandarin-medium and simplified Chinese writing and National-type School (Tamil) (Sekolah Jenis Kebangsaan (Tamil), SJK (T)), Tamil-medium.

Malaysian national secondary schools are divided into several types: National Secondary School, Religious Secondary School, National-Type Secondary School, Technical Schools, Boarding Schools and MARA Junior Science College. The National Education System at school level under the category of government education institutions consists of lower and upper secondary education.

The National and National-Type Secondary School are daily basis high school focuses on providing equitable educational rights to the population in Malaysia. Technical and vocational secondary schools provide opportunities for students with a tendency in science and technology education to meet semi-professional and professional workforce in technical and engineering fields. Meanwhile, MARA Junior Science College is a group of boarding schools created by the People's Trust Council, a

Malaysian government agency. The institution provides learning facilities for bright students in local schools throughout Malaysia.

Boarding School or Fully Residential School (Sekolah Berasrama Penuh or SBP) is a school system established to nurture outstanding students to excel in academics and extracurricular activities. To date, there are 69 SBPs, section into 11 premier SBP's, 43 Science School, 12 Integrated School and 3 Federation Religious Secondary School in Malaysia. SBP's offers a well-organized, controlled and excellent schooling and learning environment to foster and develop the potential of outstanding students.

Sekolah Menengah Kebangsaan Agama (SMKA) or National Islamic Secondary School is a type of institutional group of education established and managed by the Malaysian Ministry of Education (MOE). As at 2018, there are 58 SMKAs all over Malaysia. The establishment of SMKA was in line with the effort to modernise Malaysian education system on that time. Improvements of Islamic education system in Islamic schools were in accordance with current developments as well. SMKA is a place for those who want to learn and practice Islamic culture, not only through the teaching and learning of Arabic language, Jawi and Quranic skills, but also applying Islamic values in daily lives. SMKA students, upon their graduation, have a wide chance to further their studies locally and internationally in various fields. In 5th year, the students will sit for a national standardised examination known as the Sijil Pelajaran Malaysia (SPM, Malaysian Certificate of Education). SPM is equivalent to the O-Level prior to entry into a tertiary level education at a universities or other higher education institutions.

With the aspiration to identify and create High Prestige School (Sekolah Berprestasi Tinggi or SBT), there is a special section in one of the six main parts in NKRA which is "Improving Student Outcome". The rational appreciation and give recognition to the SBT are to lift the best schools by improving the quality of performance through implementation of innovation in education. Secondly, by producing outstanding students of international calibre to further their education in educational institutions around the world and become superior personalities in all fields of endeavor. Finally, is to bridge the gap between schools through inspirations to others for developing excellence to high levels through benchmarking, mentoring and networking.

SBP is the most prestigious group of schools in Malaysia. The SBP's have ethos, character and a unique identity to excel in all aspects of education. The schools have a tradition of high culture and excellent work in developing human capital and continue to grow holistically and are competitive in the international arena. However, other types of school have also strengthen their capability by obtain good GPS. This research will be focused more on GPS obtained from UPSR and SPM results for the year 2017 in the state of Selangor.

Generally, this study aims to identify and determine which machine learning analytical methods that provide the most accurate value for GPS prediction. More specifically this study wants to:

- To identify important characteristics affecting the GPS such as school location, school type and type of area of study; and

- To use machine learning approach by using regression method to achieve the general and specific objectives mentioned above.

1.2 Research Goals

The general aim of this study is to produce a comparison analysis using the Machine Learning approach. This study also predicts the type of school that will be able to keep up the excellence results in the SPM examination in 2018.

The study is expected to explain the relationship between GPS and other related variables such as school location, type of school and type area of study taken by students such as science, art, vocational, or Islamic stream. This study is expected to solve two questions of the research:

- Which model of analysis can give more accurate results; and
- What are the major issues probably encountered during the analysis?

2. LITERATURE REVIEW

2.1 Introduction

Kane and Staiger (2002) distinguish three types of performance measure that are used to rank schools namely test score levels, test score gains and changes in test score levels. This view is supported by Neves, Pereira and Nata (2014) who writes that the power of rankings derives not from their ability to analyse and/or explain educational processes, but rather from the fact that they have become instrumental in the promotion of competition and the development of educational markets. This statement shows that rankings gained an increasing role in promoting competition between schools and educational accountability in the world. From the rankings, policy and improvement can be organised to achieve global standard level of education system.

This study uses a qualitative case study approach to investigate Malaysia's school achievement by analysing the average grade of the school (GPS) based on the results of the Malaysian public exams, Sijil Pelajaran Malaysia (SPM) for the year of 2017. In this study, researchers have identified school achievement based on the following values in Table 2.1.

Table 2.1: Level based on Average Grade of the School (GPS)

Average grade of the school (GPS)	Level
0 - 1.6	Excellent
1.7 - 3.2	Very good
3.3 - 4.9	Good
5.0 - 6.6	Satisfactory
6.7 - 8.3	Poor

Source: Ministry of Education, Malaysia

2.2 Machine Learning

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention. The processes involved in machine learning are data mining and predictive modelling. Both require searching through data to look for patterns and adjusting program actions accordingly. Many people are familiar with machine learning from shopping on the internet and being served ads related to their purchase. This happens because recommendation engines use machine learning to personalize online ad delivery in almost real time. Beyond personalized marketing, other common machine learning use cases include fraud detection, spam filtering, network security threat detection, predictive maintenance and building news feeds.

Machine learning is one of the proven analytic tools in harnessing the power of big data. Currently, Machine Learning is the advancement of analytical techniques for big data analysis. The results may empower decision-makers by helping them to better understand current situations and to predict future more precisely and accurately. The analytical approaches use computer algorithms to repeatedly learn from the given data. The result prediction by the model will improves accuracy over time as the analytical tool continues to learn from the data.

2.3 Random Forest

Many researchers have utilized Random Forests to measure classification algorithm. Random forest is a non-parametric machine learning technique with supervised classification algorithm. As the name suggest, this algorithm creates the forest with several trees. In general, the more trees in the forest the more robust the forest looks like. Random Forest is a flexible, easy to use and it can be used for both classification and regression tasks. The higher number of trees in the forest gives the high accuracy results.

Atkinson et al. (2018) identifies the accuracy of random forest prediction process is measured as Out-Of-Bag (OOB) error. This is a generalization error based upon the ability for the Random Forest classifier to correctly classify sets of test samples. Random Forests do not require manual separation of data into training and test sets. Each tree selects approximately 2/3 of samples for construction at random and leaves aside 1/3.

The approach used in this research is similar to that used by other researchers. Yoo (2018), use Random Forest to gives estimates variables are important in the classification and to identify the relative importance of urban physical and socioeconomic characteristics to the formation of spatially varying land surface temperature across Marion County. On the other hand, Breimen (2001) use Random Forest because there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error.

2.3 Other Regression Models

In improving the methodology in terms of decision making, various statistical models are practiced such as logistic regression, linear discriminant analysis (LDA), neural network (NN), decision tree, cluster analysis etc. The use of modern statistical methods and data mining techniques can contribute substantially to information science and be able to build models that measure the level of risk based on the characteristics and variables in the data that lead to the differences between each of their records.

West (2000) uses a neural network model in investigating the accuracy of credit scoring. He also compared the model with a commonly used model in determining lending approval such as linear discriminant analysis (LDA), logistic regression, k-nearest neighbourhood (k-NN) etc. There are five models that are covered by neural networks: mixture of experts (MOE), radial base function (RBF), learning vector quantization (LVQ), fuzzy adaptive resonance (FAS) and multilayer perception (MLP). However, only RBF and MOE models can be considered as models in determining approval. Traditionally, logistic regression is a more precise method and gives high precision value to determine loan approval.

Fan et al. (2013) review the implementation of principal component analysis (PCA) in reducing the variables and factors affecting loan approval. They found that the combination between the PCA method and the improved tree augmented naïve Bayesian classification (ITANC) model could produce a more accurate model in predicting the probability of the loan approval. In addition, PCA could improve the existing model to be applied in the loan application process.

3. DATA AND ANALYSIS

3.1 Research Scope

This research will use the existing data available in Ministry of Education Malaysia with the GPS data provided by Selangor State Education Department. The data include all primary and secondary school in Selangor.

3.2 Testing Data

The data used as the testing data is the profile data of each primary and secondary school in Selangor. This data has 917 school records based on 13 attributes including average grade of the school (GPS). Table 3.1 shows data references for each attribute.

Table 3.1: Data References for Each Attribute

Attributes	References
Level of school	1: Primary school 2: Secondary school
Type of school	1: National school 2: National-type school (Chinese) 3: National-type school (Tamil) 4: Vocational college 5: Religious secondary school 6: Fully residential school 7: National-type secondary school 8: Special secondary school 9: Primary and secondary school (Special model) 10: Religious national-type secondary school
Location of school	1: Urban 2: Rural
Sponsorship of school	1: Fully government school 2: Government-aided school
Stream of school	1: Nation 2: Chinese 3: Tamil 4: Special 5: Islamic 6: Academic (Science/Art) 7: Technical
Students	Number of students in school
Teachers	Number of teachers in school
Average grade of the school	Average grade of the school

Source: Ministry of Education, Malaysia

3.3 Data Exploration and Visualization

Data exploration and visualisation is an important element in recognising data in more detail. Figure 3.1 shows the number of schools by the average grade of the school (GPS). In this figure, there are 617 or approximately 67.0% of schools in Selangor have an excellent GPS of 1.7 to 3.3. The GPS achieved is in range between 1.42 and 8.21.

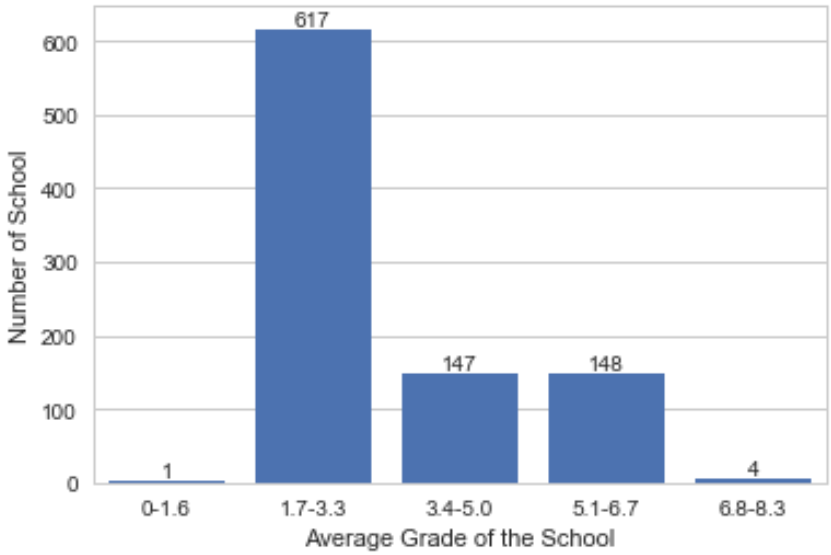


Figure 3.1: Number of Schools by the Average Grade of the School (GPS)

Figure 3.2 shows the scatter plots describing the relationship between the total number of students and teachers by GPS. This following diagram indicating that there is a significant overlap between different GPS groups. This implies that this figure does not illustrate the relationship between the number of students and teachers to GPS.

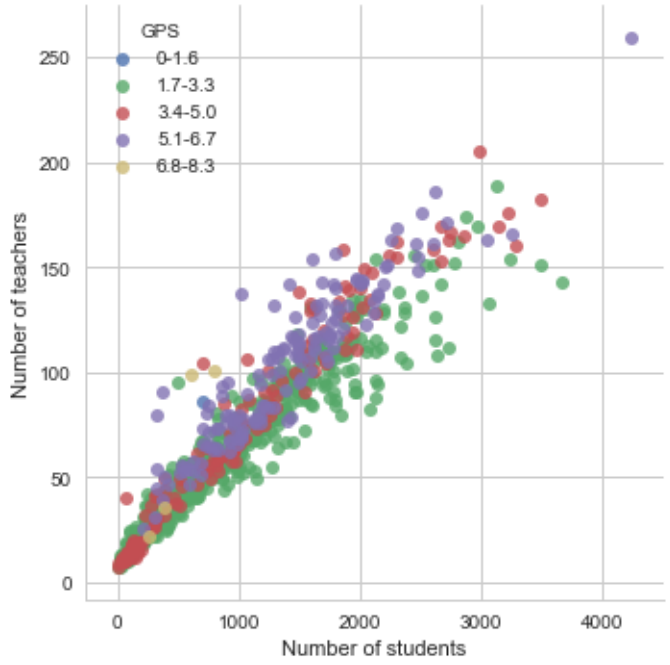


Figure 3.2: Number of Students and Teachers by GPS

Figure 3.3 shows the chart of school frequencies by their level. There are 653 primary schools and 264 secondary schools in Selangor. It is found that most of secondary schools have a satisfactory GPS of 5.1 to 6.7. In addition, the number of primary schools with GPS of 1.7 to 3.3 exceeds the number of such secondary schools. This clearly shows that level of school has significant relationships with GPS.

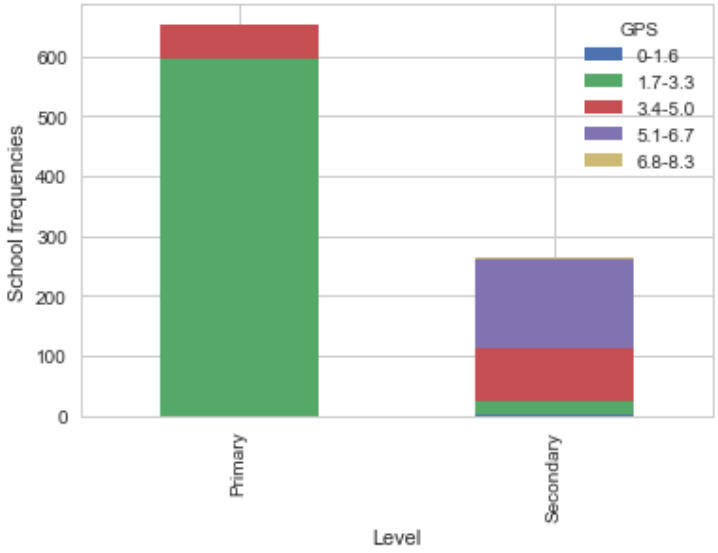


Figure 3.3: School Frequencies by Level of School

Figure 3.4 shows the school frequencies by type of school. This graph shows that most national-type school (including Chinese and Tamil) have excellent GPS of 1.7 to 3.3. However, most types of schools with satisfactory GPS of 5.1 to 6.7 are from the national-type secondary schools. Therefore, different proportions by type of school indicating that GPS may be influenced by the type of school.

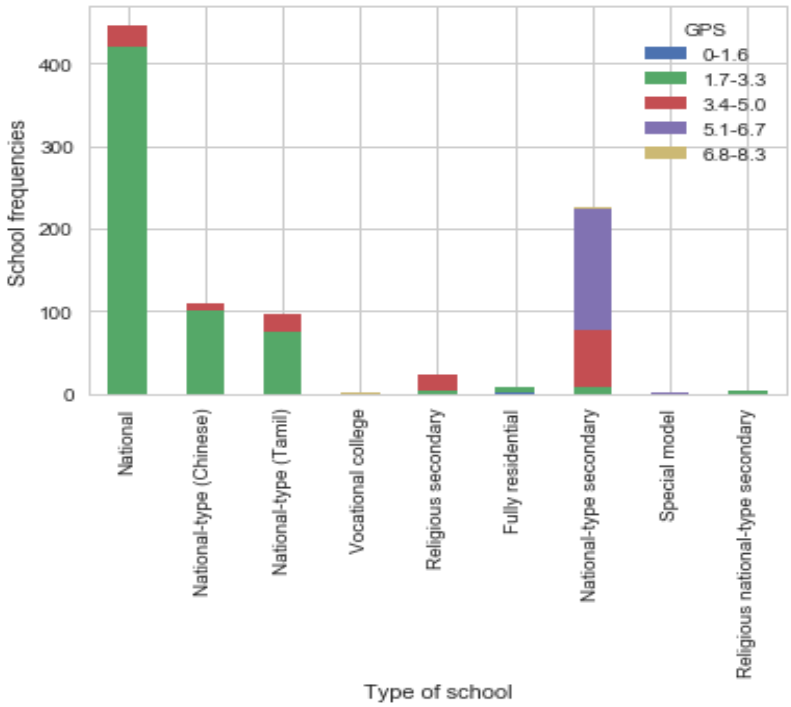


Figure 3.4: School Frequencies by Type of School

Figure 3.5 shows the school frequencies by the location of a school. A total of 54.3% of schools in Selangor are schools in the urban area and the rest of them are in the rural area. However, this graph shows the proportion of the group according to the GPS is almost similar based on the location of the school. This means that the location of a school may not affect the GPS either in urban or rural areas.

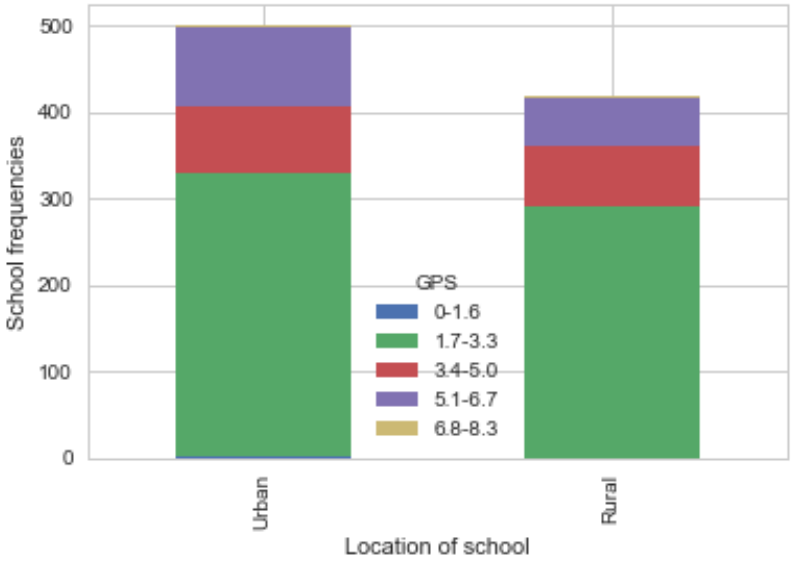


Figure 3.5: School Frequencies by Location of School

Figure 3.6 shows the school frequencies through the sponsorship of a school. A total of 81.7% of schools in Selangor are fully sponsored schools by the government and the rest of them are government-aided schools. However, schools that are fully sponsored by the government have a satisfactory GPS (5.1 to 6.7) and their number are more than the numbers of government-aided schools. This means that GPS may be influenced by government assistance to the schools whether fully sponsored or otherwise.

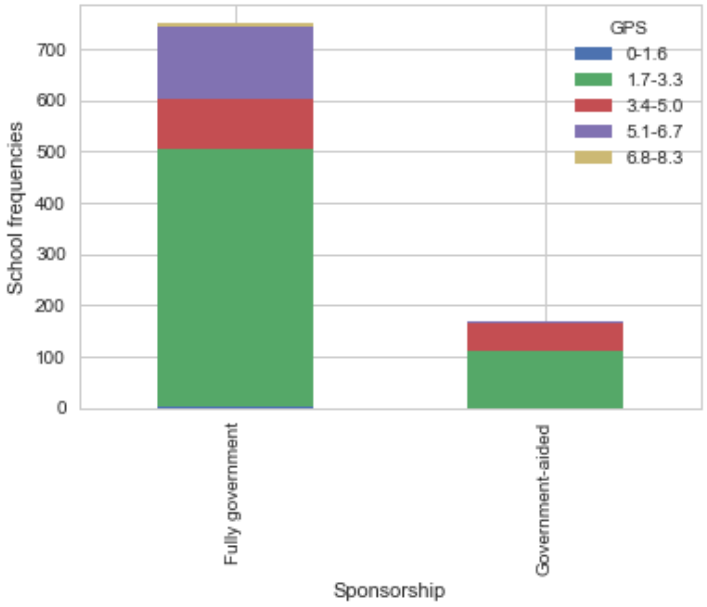


Figure 3.6: School Frequencies by Sponsorship of School

4. RESULT

4.1 Machine Learning Approach

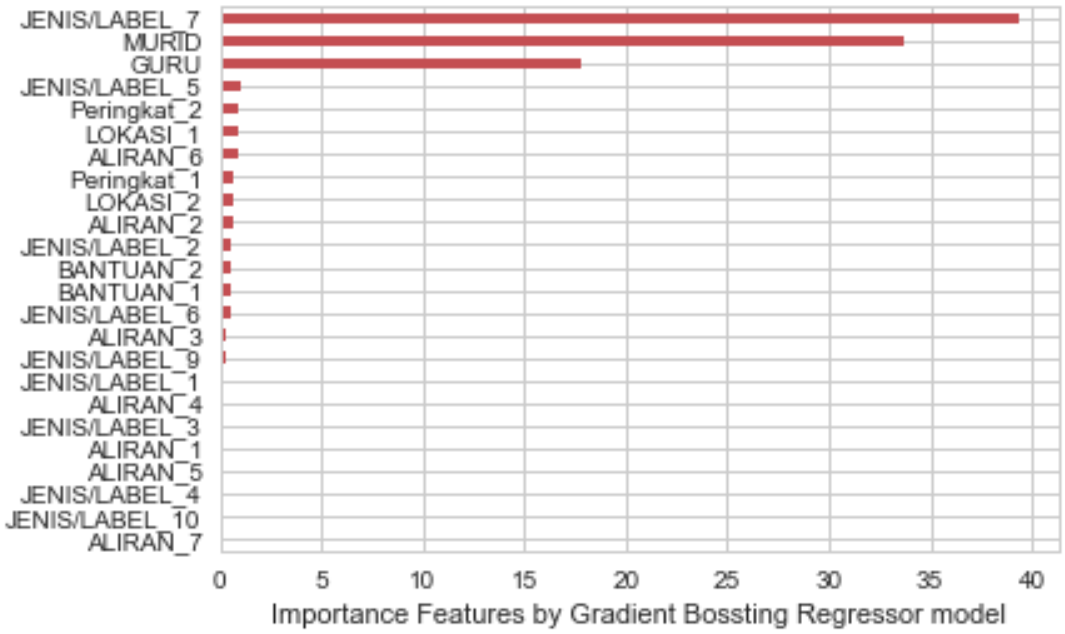
The machine learning used in building a model that classifies GPS from data is Logistic Regression, Linear Discriminant Analysis (LDA), k-Nearest Neighborhood (k-NN), Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression and Support Vector Regression (SVR). The accuracy and mean squared error (MSE) of each model are the indicators in determining the best model in predicting GPS based on the given profile. Table 4.1 shows the accuracy and MSE in each model. It was found that the k-NN and SVR models was not suitable as a model for predicting the important variables for GPS due to its low accuracy and high MSE. The model that provides high accuracy is the LDA model, while the model that provides a low MSE is the Logistic Regression model. However, overall, all models except k-NN and SVR models can provide good prediction of GPS based on the given accuracy and MSE.

Table 4.1: Accuracy and Mean Squared Error (MSE) of Regression Models

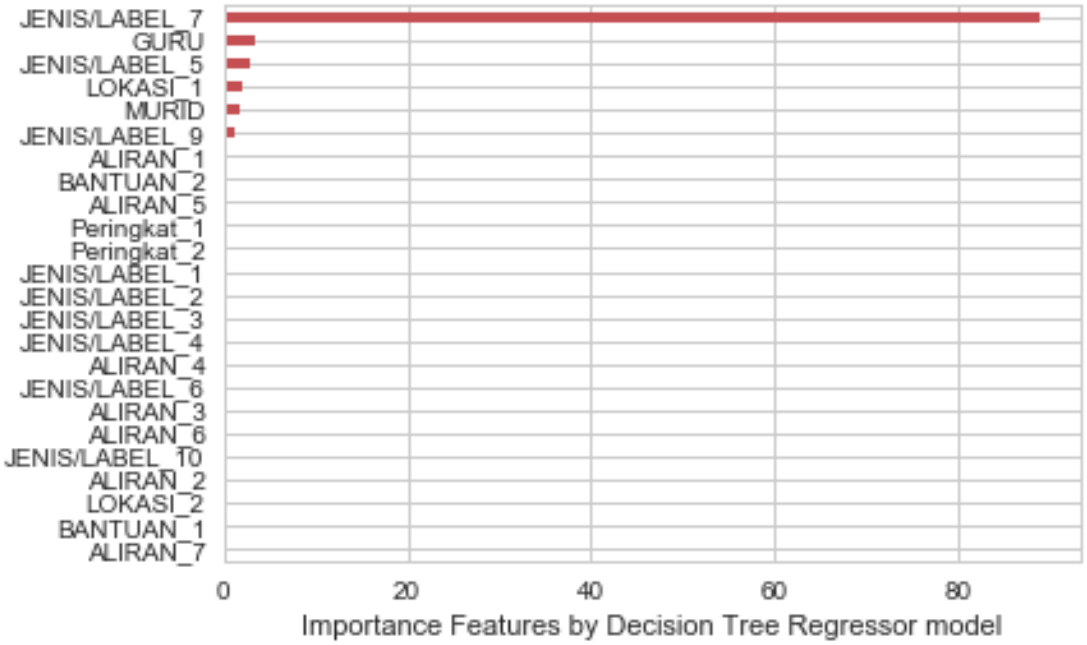
Model	Accuracy	Mean squared error (MSE)
Logistic Regression	0.869565	0.130435
Linear Discriminant Analysis (LDA)	0.873188	0.148551
k-Nearest Neighbourhood (k-NN)	0.076741	0.598732
Decision Tree Regression	0.733960	0.172526
Random Forest Regression	0.714893	0.184891
Gradient Boosting Regression	0.737191	0.170431
Support Vector Regression	0.081768	0.595472

4.2 Importance Features of GPS

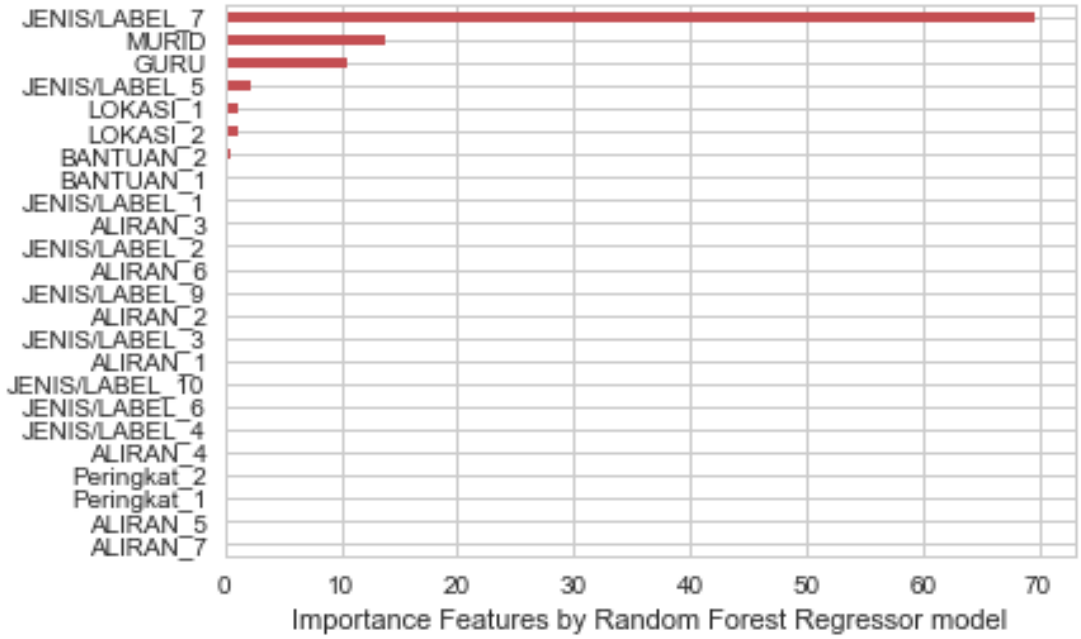
This study aims to identify important characteristics that are affecting the GPS. The importance features were measured using three models, which are Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression. Among these models, the significant features to classify and predict the GPS are type of school (whether it is national-type Secondary school or not), the number of students and teachers, and location of the school. Figure 4.2 shows the important features by Gradient Boosting Regression, Decision Tree Regression and Random Forest Regression models.



i. JENIS/LABEL_7 = Type of school_7: National-type secondary school
ii. MURID = Student
iii. GURU = Teacher
iv. JENIS/LABEL_5 = Type of school_5: Religious secondary school
v. Peringkat_2 = Level of school_2: Secondary school
vi. LOKASI_1 = Location of school_ Urban
vii. ALIRAN_6 = Stream of school_6: Academic (Science/Art)



i. JENIS/LABEL_7 = Type of school_7: National-type secondary school
ii. GURU = Teacher
iii. JENIS/LABEL_5 = Type of school_5: Religious secondary school
iv. LOKASI_1 = Location of school_ Urban
v. MURID = Student
vi. JENIS/LABEL_9 = Type of school_9: Primary and secondary school (Special model)



i. JENIS/LABEL_7 = Type of school_7: National-type secondary school
ii. MURID = Student
iii. GURU = Teacher
iv. JENIS/LABEL_5 = Type of school_5: Religious secondary school
v. LOKASI_1 = Location of school_ Urban
vi. LOKASI_2 = Location of school_ Rural
BANTUAN_2 = Sponsorship of school _2: Government-aided school

Figure 4.2: Important Features by Gradient Boosting Regression, Decision Tree Regression and Random Forest Regression model

5. CONCLUSION

The purpose of the current study was to determine that the overfitting problem will never happen when we use the random forest algorithm in any both classification and regression task. The results of this research support the idea that machine learning plays a main role in improving the accuracy, timeliness and relevance of statistics at a lower cost than expanding existing data collections such as validated surveys and censuses. The results of this study show that across all the characteristics affecting the GPS considered in the random forest, the total of students and the type/level of school are the most important variables to GPS level achievement. We have shown both scenario and confusion matrix to get an overview of the GPS performance. This approach will prove useful in expanding our understanding of how to use machine learning approach in data analytics. Notwithstanding these limitations, the study suggests that to undertake another analysis by using principal component analysis approach. In response to the dynamic and changing global technology environment, the implementation of data analytics development of data analytics expertise will pave the way to the desired level and improving the ability to extract knowledge and insights from large and complex collection of digital data.

REFERENCES

Atkinson, J. S., Mitchell, J.E., Rio, M. and Matich, G. (2018). Your WiFi is leaking: What do your mobile apps gossip about you? *Future Generation Computer Systems, 80*, 546-557.

Breiman, L. (2001). Random Forests Machine Learning. 45: 5–32. *View Article PubMed/NCBI Google Scholar*.

Fan, Y. Q., Yang, Y. L., & Qin, Y. S. (2013). Credit scoring model based on PCA and improved tree augmented Bayesian classification.

Kane, T. J. and Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic perspectives, 16*(4), 91-114.

Neves, T., Pereira, M. J. and Nata, G. (2014). Head teachers' perceptions of secondary school rankings: Their nature, media coverage and impact on schools and the educational arena. *Education as Change, 18*(2), 211-225.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research, 27*(11-12), 1131-1152.

Yoo, S. (2018). Investigating important urban characteristics in the formation of urban heat islands: a machine learning approach. *Journal of Big Data, 5*(1), 2.