Predictive Model of Graduate-On-Time Using Machine Learning Algorithms

Chapter · September 2019

DOI: 10.1007/978-981-15-0399-3_11

Nurafifah Mohammad Suhaimi
2 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Shuzlina Abdul Rahman
Universiti Teknologi MARA
89 PUBLICATIONS **380** CITATIONS
SEE PROFILE

Sofianita Mutalib
Universiti Teknologi MARA
63 PUBLICATIONS **235** CITATIONS
SEE PROFILE

Nurzeatul Abdul Hamid
Universiti Teknologi MARA
26 PUBLICATIONS **138** CITATIONS
SEE PROFILE

Predictive Model of Graduate-On-Time Using Machine Learning Algorithms

Nurafifah Mohammad Suhaimi[3], Shuzlina Abdul-Rahman[1,2,3](✉), Sofianita Mutalib[1,2,3], Nurzeatul Hamimah Abdul Hamid[1,2,3], and Ariff Md Ab Malik[1,3,4]

[1] Research Initiative Group of Intelligent Systems, Universiti Teknologi MARA,
40450 Shah Alam, Selangor, Malaysia
{shuzlina, sofi, nurzea}@tmsk.uitm.edu.my,
ariff215@puncakalam.uitm.edu.my

[2] Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

[3] Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
afifahsuhaimi01@gmail.com

[4] Faculty of Business and Management, Universiti Teknologi MARA,
42300 Puncak Alam, Selangor, Malaysia

Abstract. In most universities, the number of students who graduated on time reflect tremendously on their operation costs. In such cases, the high number of graduate-on-time or GOT students achievement will indirectly reduce the university's annual operation cost per student. Not as trivial as it seems, to ensure most of the students able to GOT is challenging. It may vary in the perspective of university practises, academic programmes, and students' background. At the university's level, students' data can be used to identify the achievement and ability of students, interests, and weaknesses. To build an accurate predictive model, it requires an extensive study on significant factors that may contribute to students' ability to graduate on time. Consequently, this study aims to construct a predictive model that can predict students' graduation status. We applied five different machine learning algorithms (classifiers) namely Decision Tree, Random Forest, Naïve Bayes, Support Vector Machine (PolyKernel), and Support Vector Machine (RBFKernel). These classifiers were evaluated with four different *k* folds of 5, 10, 15, and 20. The performance of these classifiers was compared based on different measurement subject to accuracy, precision, recall, and F-Score. The results indicated that Support Vector Machine (PolyKernel) outperformed other classifiers and the best numbers of *k* folds for this experiment are 5 and 20. This predictive model of GOT is hopefully will beneficial to university management and academicians to devise their strategies in helping and improving the weakness of students' academic performance and to ensure they can graduate on time.

Keywords: Data mining · Graduate-On-Time · Machine learning · Predictive model · Supervised algorithm

M. W. Berry et al. (Eds.): SCDS 2019, CCIS 1100, pp. 130–141, 2019.
https://doi.org/10.1007/978-981-15-0399-3_11

1 Introduction

The decreasing pattern of the Graduate-On-Time (GOT) students has become a major issue by The Malaysian Ministry of Education, professors, academicians and related parties despite the increasing number of student enrolment. Ensuring the students to GOT has become the biggest challenge to the university since the position of a university in the education industry relies on this indicator that acts as one of the metrics to measure institutional effectiveness [1]. The measurement of academic productivity is measured in many ways, depending on the range of academic input parameters and outcomes. The two productivity indicators used by the Malaysian universities are the Intake Graduation on Time (iGOT) and the annual cost per Full-Time Student Equivalent (Cost per FTSE) [2]. The iGOT measures the productivity based on the number of students (based on intake cohort) that graduates within the programme stipulated duration. Rather than looking at the annual graduation rate, iGOT highlights the number of students who graduated in comparison to the intake numbers. Based on this measurement, though a student may take a longer period to finish his degree, the university will bear more cost. The improvement of iGOT rate also ties with the Cost per FTSE [2]. Based on the 2013/2014 statistics, the average iGOT for public universities is 74%. The statistics also show a large gap between the universities of 27%. Though in some cases, delayed graduation is unavoidable, analysis from historical data could provide beneficial insights [1]. Hence, Education Data Mining (EDM) application in such areas could pave a way to understand the issues at hand better.

EDM devises a new exploration paradigm to analyse students' learning behaviour to gain insights. It uses machine learning techniques to explore data from educational settings such as online logs, teaching approaches, teaching resources, interim tests and examination results to predict and learn the patterns that characterize students' behaviours that affect their performance. Subsequently, the goal of the exploration is to understand and improve the educational outcomes. For example, the prediction model of the students' performance by forecasting the students' grade allows the academic management of the university to devise a proper warning mechanism for students who are at risk. Hence, able to help them to overcome difficulties in their study. Therefore, the prediction model could provide useful insights for strategic programmes to plan a suitable measure to improve the students' performance [3]. In a similar vein, research findings reveal that among the factors that contributed to students' poor performance are gender issue [4], marital status and age issue [5], family issues [6], previous academic record [7], demographics, personal, educational background, psychological, academic progress, and other environmental variables [8]. Even though most of the significant factors have been identified, the prediction of students' performance is dynamic as it varies from universities, programmes, and students' background [9]. In the EDM method, predictive modelling anticipates students' graduation time. In order to build the predictive model, there are several tasks used, which are classification, regression, and categorization. For example, classification task creates predictive models for target variable prediction based on several input variables [10]. Classification techniques are frequently applied to ease the decision-making process [11]. This paper aims to discover the ideal classifier that performs the best to predict students' graduation status or simply GOT. There are several classifiers under classification task that have been applied to predict students' graduation time. Among the classifiers used are Decision tree, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor, and Support Vector Machine [10, 12–14]. However, the performance of these classifiers is varied, as they perform differently according to the data type. Several experiments need to be carried to find which classifier works the best. We present this paper as follows: The first section summarizes the overview of this research while section two describes the related works. The third section reviews the methodology involved to carry out this research. The fourth section presents the results and discussions. Lastly, we conclude this paper in section five.

2 Related Work

The increasing number of students who are unable to graduate on time or GOT significantly affect the institution to produce the quality outputs each year and contribute the low score on the graduation rates [2]. Consequently, it gives adverse impact to the university's productivity, hence affect the university's ranking. The iGOT measures whether students finish their studies within the required time. Shariff et al. [14] defined iGOT as a state where students accomplished their studies in particular time that has been set by the university. In which most of the institutions in other countries in Europe have set the time for undergraduate students to graduate is four years, but regrettably, many students delay in finishing their degree (40% completed within four years while 60% in six years) [15].

Students who take a longer time to graduate affect the university's budget as the university has to spend more money to provide extra resources such as extra classroom to cater number of students [2]. There are about 25% of students in Malaysian universities overstay their course duration [16]. Among the suggestions proposed is to penalize students who fail to GOT. This action could act as a reminder to others to always take their study seriously. This challenging scenario worries many stakeholders especially the university's management as they have to think outside of the box and come out with a sturdy plan in solving such T as well as improving the number of graduation rates. They need to handle this issue brilliantly and proactively as the university's achievement depends highly on the graduation rate. One of the solutions to handle this issue is by analysing students' performance since it can be the indicator to predict the students' graduation time. However, analysing the performance of students is very complicated and tedious as it involves with many data that is continuously increasing year by year [17]. Alternatively, data mining is applicable to perform analysis in solving this issue.

Data mining (DM) is a process that can convert massive amount of data and turn into meaningful information or knowledge. It is about analysing and categorizing the information as well as summarizing the knowledge from different kind of data stored in database and data warehouse [9, 18]. Nowadays, the application of data mining is widely prevalent in the education system and beneficial to analyse students' performance, forecast students' graduation time and other related issue [19–21]. A predictive model to predict the graduation status of doctorate students was performed by [14]. The result shows that there is a total of 79 students who are predicted to GOT. Based on their research, they outlined that female tend to GOT compared to male as the number of female students who are predicted to GOT is 56% higher compared to male. Findings from [14] can be intensified by [22] as they underlined that from their research, the achievement level of female students in software education was higher than male students. The accuracy of DM model can be improved as well as performing a more in-depth analysis of the mined data with the use of machine learning algorithms. Machine learning algorithms are the methods usually employed by researchers to discover patterns from data sets by letting them learn on their own [23]. Besides with the advancement of extensive computing capabilities, learning through a tremendous amount of data is now seems possible.

3 Methodology

There are six phases involved in this research methodology following Cross-Industry Process for Data Mining (CRISP-DM) [24] where each phase is associated with the necessitate activities. The detail of each phase is discussed briefly in the next sections.

Business Understanding. Business Understanding is the first phase of this research. In this phase, the main area that will be examined is issues that are related to students' graduation status and the factors that contribute to students' timely graduation.

Data Understanding. The second stage of the CRISP-DM process is Data Understanding which requires the researcher to obtain the required data as well as transformed it into a format that can be mined using Data Mining Tool. In this research, the technique applied in gathering the data was through document and records, by examining UiTM students' database that contains the historical data of UiTM's students of cohort 2013. This data was then been examined carefully to identify the distribution and its range values. Based on the examination, the raw data from UiTM's database consist of 31 attributes with 74,670 instances.

Data Preparation. This is the crucial phase in CRISP-DM process as the competency of the research's model is highly depends on the quality of the dataset. Several activities involved in this phase which include data selection & cleaning, data construction, and data integrating & formatting. Several attributes from the raw dataset were selected based on the relevance to the goals of this research. The remaining of the unselected attributes were discarded as it did not give any meaning and contribution to achieve the goals of this research. Table 1 shows the list of the selected attributes, 13 attributes with GOT status as the target or class label.

Table 1. Attribute selection

Attribute	Value	Description
Prog_desc	Science programme, Engineering programme	Programme taken by the students
Study_mode	Full-Time, Extended Full-Time	The mode of study
Sponsor	Yes, No	Loan or scholarship taken by the student
Disability	Yes, No	Student's disability
Gender	Male, Female	Student's gender
Marital_status	Single, Married	Student's marital status
Race	Dusun, Iban/Sea Dayak, Jawa, Bidayuh, Melayu	Student's race
Age	18, 20, 25, 30	Student's age
Permanent_state_desc	Johor, Kedah, Kelantan, Perak, Perlis, Pulau Pinang, Sabah, Sarawak	Student's permanent address
Entry_rquirement	Diploma, Matrikulasi KPM	Student's intake mode
CGPA	0–1.99, 2.00–2.49, 2.50–2.99, 3.00–3.49, 3.5–4.0	Student's CGPA
Family_income	Nil Income RM1 - RM499.99 RM500 - RM999.99 RM1000 - RM1999.99 RM2000 - RM2999.99 RM3000 - RM3999.99 RM4000 - RM4999.99 RM5000 - RM7999.99 RM8000 - RM9999.99 RM10000 dan keatas	Student's family income
GOTstatus	GOT, Non-GOT	Student's graduation status

This research catered students' data from Science and Engineering Programme. By using "filter" task in Microsoft Excel, for *prog_desc* attribute, all programmes other than Science Programme and Engineering Programme were discarded. The total of instances of Engineering Programme is 1160 instances while Science Programme is 2575 instances.

Modelling. Modelling phase is the part to search for useful patterns in data. Within machine learning process, dataset needs to undergo the modeling process to identify the patterns from the datasets. In data mining, there are numbers of modelling algorithm or commonly known as classifier, but not all of them suit with this research's project. The list of the classifiers was narrowed, based on few rules particularly based on the business questions and the type of variables involved. The prepared data derived from data preparation process was trained on five different classifiers, namely Decision Tree, Random Forest, Naïve Bayes, SVM with PolyKernel, and SVM with RBFKernel.

Evaluation. In model evaluation phase, these classifiers were evaluated based on four performance measurements which are the value of accuracy score, precision, recall, and F-measure. The classifier that scored the best is appointed as the best classifier for this research.

Deployment. Deployment phase is the phase where all the research's progress, outcomes, results and findings as well as any problem or limitations are deployed in a report form.

The predictive models are developed using a data mining tool, called Waikato Environment for Knowledge Analysis (WEKA) [25].

4 Results and Discussions

This section presents the performance of the five classifiers as mentioned in the previous section. These classifiers were evaluated based on cross-validation with 5, 10, 15, and 20 folds. The performance of each classifier is interpreted using different performance measures such as accuracy score, precision, recall, and F-Score. The next subsections present and analyze the results gained from these classifiers.

4.1 Accuracy Score Analysis

Accuracy score is the correct number predictions made divided by the total number of predictions made, multiplied by 100 to turn it into a percentage. In other word, accuracy is the possibility that the classifier can correctly predict the positive and negative instances. Figure 1 shows the overall accuracy score for all the five classifiers.

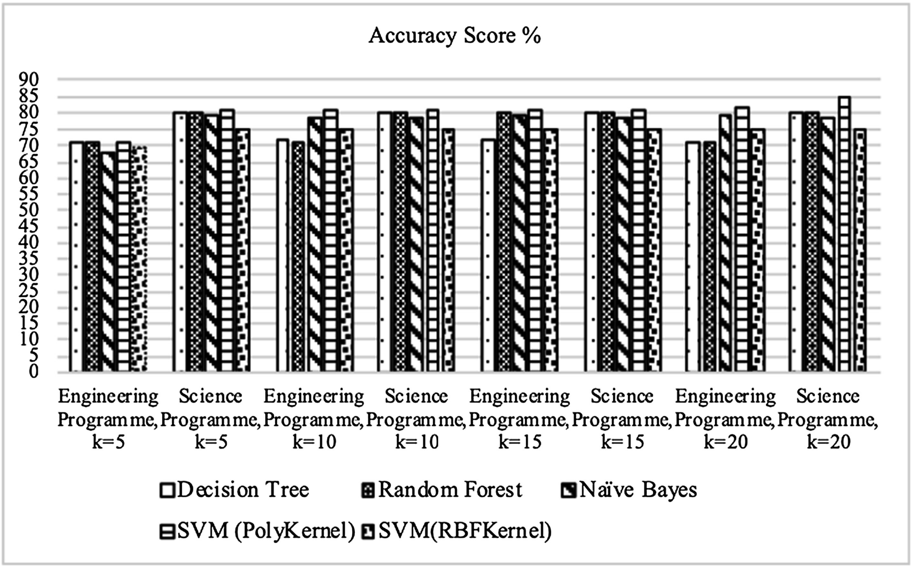

Fig. 1. Accuracy score of different classifiers

According to the chart in Fig. 1, the results of each classifier are illustrated, according to the type of data and the cross-validation fold. From the graph, the highest bar chart can be seen in Science Programme is when k = 20, which is from SVM (PolyKernel). This classifier has outperformed the others with an accuracy score of 84.78%. On the other hand, the lowest accuracy score was achieved by Naïve Bayes with k = 5, in Engineering Programme, with 68.13% only. When k = 5, we can clearly see that classifiers performed better on Science Programme, compared to Engineering Programme. However, on average, the accuracy score gained from all classifiers on Engineering Programme increased when k = 10, but there are no big differences occurred in term of accuracy score on Science Programme. The accuracy score of Random Forest rose dramatically when k = 15 for Engineering Programme, but for the other classifiers, the score remained constant for both type of data. Nevertheless, the accuracy score for Random Forest dropped back when k value is changed to 20. SVM (PolyKernel) on the other side went up steadily in the term of accuracy score when k value is changed from 15 to 20. Moreover, we can see a clear upward trend on the accuracy score of SVM (PolyKernel) for both type of data, when being tested on k = 5, 10, 15, and 20 and it reached the maximum score when k value is 20. Overall, from the graph above, we can conclude that on average, all classifiers gave better accuracy result for Science programme in the range of 75 to 85% and the ideal k value for determining the accuracy score is when k = 20.

4.2 Precision Score Analysis

In machine learning, precision is the number of positive values that are predicted correctly to the total predicted positive. Here, the correct positive prediction is highlighted out of all the positive predictions. For example, in this research, we want to highlight the real number of students who are actually GOT out of all the students who are predicted to GOT. The high precision score indicates that there is low false positive. False positive is when the classifier labels data that is actually negative with positive. In this research, non-GOT students are labeled with GOT. The bar chart in Fig. 2 shows the comparison of the precision score for all the classifiers.

Overall, from the graph above, we can see that the precision score of these five classifiers remained steady throughout the experiments when k = 5, 10, 15, and 20. Here, when k = 5, the performance of SVM (PolyKernel) when being tested on Science Programme achieved the highest precision score compared to other classifiers. The precision score of SVM (PolyKernel) decreased about 1% when k = 10, 15, 20 and the score remained unchanged throughout the experiments which are 0.82, with Science Programme. From Science Programme, it can be seen that SVM (PolyKernel) and SVM (RBFKernel) shared the same precision score when k = 15 and k = 20, which show that these two classifiers performed identically with that k values. In a similar vein, for Science Programme, Decision Tree and Random Forest also achieved the same precision score throughout the experiments, which is 0.8. On the other hand, for Science Programme, the trend of the precision score for all classifiers are the same, except for Random Forest. When k = 5 and k = 10, Random Forest scored 0.71, but the score increased slightly to 0.72 for the next cross-validation folds, which are k = 15 and k = 20. Besides, SVM (PolyKernel) scored the lowest precision score, compared to

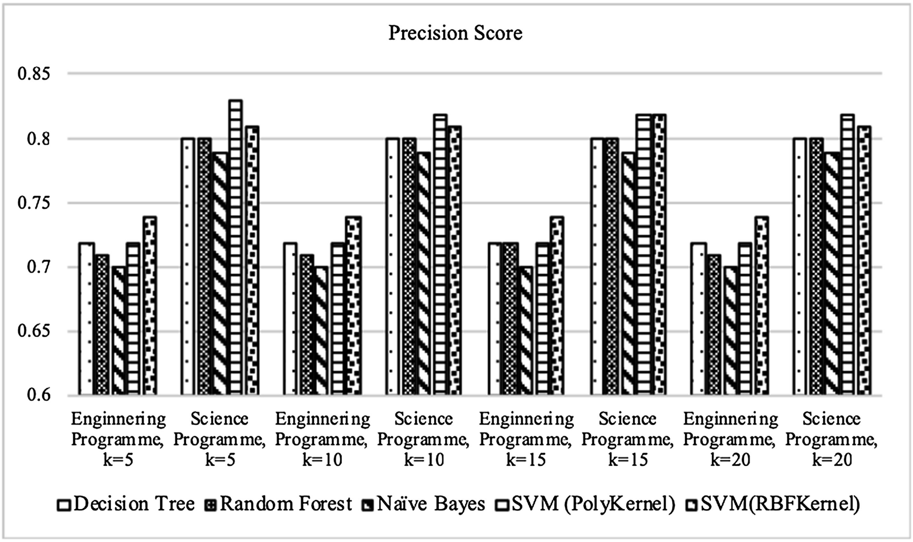

Fig. 2. Precision score of different classifiers

the other classifiers, which is 0.70. However, this classifier performed better on Science Programme, which shows that this classifier was good in classifying data in Science Programme, compared to Engineering Programme. Overall, from the graph, it can be concluded that all classifiers able to produce more precise result in predicting GOT with Science programmes and on average, the best number of cross-validation folds to measure the precision score of all the classifiers is 5. In the next section, the performance analysis on each fold is discussed in detail.

4.3 Recall Score Analysis

Recall or Sensitivity is the proportion of real positive value that is correctly predicted positive. It is also can be defined as the ratio of correctly predicted positive values to the actual positive values. Recall highlights the sensitivity of the algorithm i.e. out of all the actual positives of how many were caught by the classifier. Recall score is calculated by the real positives number divided by the real positives number plus with the false negatives number. Real positives are where data is classified as positive by the classifier that is actually positive, or in a simpler word, they are correct. False negatives on the other hand are data that the classifier marks as negatives are actually positive, or incorrect. In this research, real positives are correctly predicted GOT and false negatives are the students that the classifiers label as non-GOT that actually were GOT. Recall can be thought as of a classifier's ability to find all the data points of interest in a dataset. The bar chart in Fig. 3 shows the comparison of the recall score for all the classifiers.

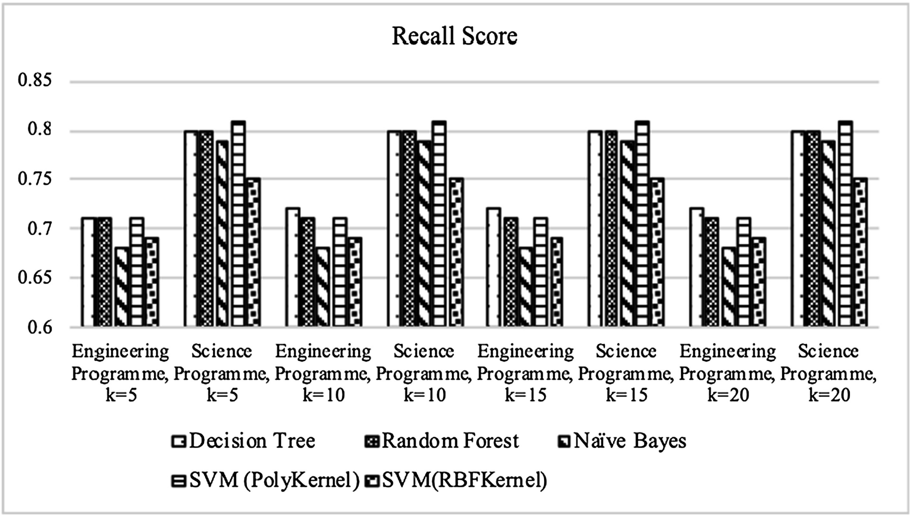

Fig. 3. Recall score of different classifiers

From the graph, for Engineering Programme, it can be seen that the performance of all classifiers was slightly low when the cross-validation folds are equal to 5. Here, Decision Tree, Random Forest and SVM (PolyKernel) shared the same level of recall score which is 0.71%, and the lowest recall score is achieved by Naïve Bayes, with only 0.68%. However, the performance of Decision Tree increased to 0.72% and remain constant for k = 10, 15 and 20. Moreover, the performance of all of the classifiers also remained constant for k = 10, 15 and 20. It can be concluded that when k = 5, the classifiers' performance was slightly lower, compared with the rest of the cross-validation folds. For Science Programme, the pattern of the recall score for all classifiers is similar when k = 5, 10, 15, and 20. Here, Decision Tree and Random Forest shared similar score, which is 0.80%, throughout the experiments. SVM (PolyKernel) has outperformed the other classifiers with a score of 0.81%, while SVM (RBFKernel) achieved the lowest recall score, which is 0.75%. Overall, from the graph, it can be concluded that all classifiers predicted GOT status better with Science Programme and the different number of folds did not affect the recall score of the classifiers.

4.4 Performance Analysis

The F score, also called as F measure, is a measure of a test's accuracy. The F score is defined as the weighted harmonic mean of the test's precision and recall. The score of F score takes the precision and recall of a test into account. As previously explained, precision is the ratio of positive results that are truly positive. This is also known as the positive predictive value. Recall on the other hand, or more known as sensitivity, is the ability of a test to correctly identify positive results to get the true positive rate. When a classifier achieved F score closest to 1, that's mean the classifier has a perfect balance of precision and recall. The F score that is closest to 0 is the worst, which means it has low score of precision as well as recall. The following Fig. 4 illustrates the trend of F score between Engineering Programme and Science Programme for all the classifiers.

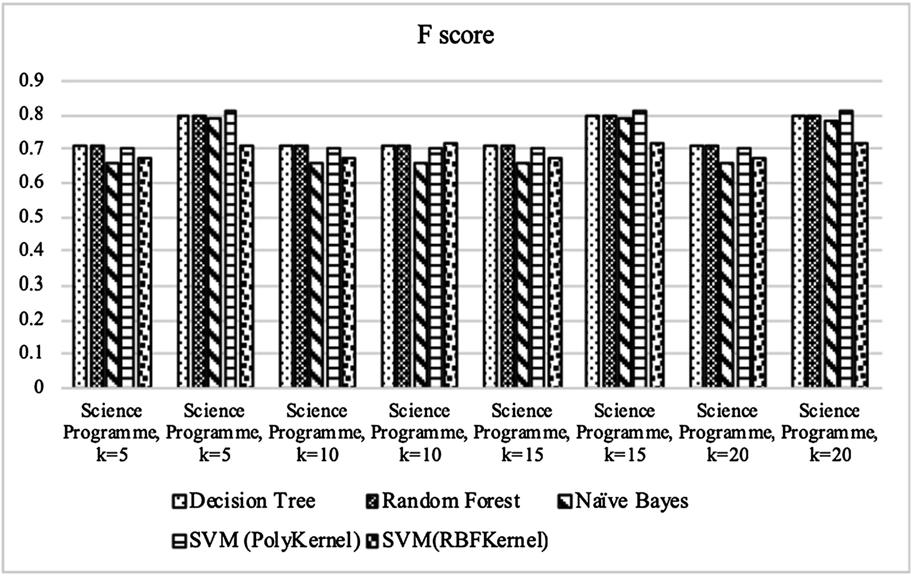

Fig. 4. F score of different classifiers

From the graph, the trend of the F score for all classifiers with Engineering Programme and Science Programme, when k = 5, 15, 20, is similar. Here, it can be interpreted that Decision Tree, Random Forest and SVM (PolyKernel) have dominated the F score for both types of data, if being compared according to the type of data. Specifically, these classifiers performed better with Science Programme, compared to Engineering Programme. When k = 10, the F score of Decision Tree, Random Forest, Naïve Bayes and SVM (PolyKernel) dropped significantly when being tested with Science Programme. However, in contra with the performance of those classifiers, the F score for SVM (RBFKernel)increased slightly when k = 10 and remained constant when k = 15 and 20. It shows that this classifier had difficulties to classify data when k = 5, but it gets better when the k values increased. For Engineering Programme, the trend of all classifiers is the same throughout the experiments. Decision Tree and Random Forest achieved the same level of F score when k = 5, 10, 15 and 20, which is the highest F score for this data type (0.71%). Naïve Bayes on the other hand achieved the lowest score, which is 0.66%, where the difference in term of F score between this classifier with Decision Tree and Random Forest is 0.05%, which is small. Overall, on the average, it can be said that classifiers performed better with Science Programme, when the cross-validation folds are equal to 15 and SVM (PolyKernel) is the best classifier that can be used to predict students' graduation status.

5 Conclusion

This paper demonstrated the use of five classifiers of machine learning algorithms to produce predictive model of Graduate-On-Time (GOT). These classifiers were modeled on Engineering Programme and Science Programme using cross-validation with 5, 10, 15, and 20 folds. The performance of each classifier was interpreted using different performance measurement such as accuracy score, precision, recall, and F-Score. The results showed that SVM (PolyKernel) outperformed other classifiers. However, Naïve Bayes had difficulties to predict students' graduation status as this classifier had produced the lowest average accuracy and score subject to precision, recall, and F Score. The model generated from Science Programme has contributed better classifiers performance compared to Engineering Programme due to its balance data distribution. The ideal *k* values for cross-validation folds for this research experiment were 5 and 20. This research is beneficial to many parties such as the university's academic management, academicians and students, as it can give alert about students' performance that are most likely fail to GOT and the actions can be taken to solve this problem. In addition, this approach also can improve the university's academic quality as the number of students who unable to GOT can be reduced significantly. In the future, this work can be enhanced by utilizing more datasets from different programmes particularly from non-science and technology fields.

Acknowledgment. The dataset for this research was provided by Center for Strategic Planning and Information (CSPI) Universiti Teknologi MARA. The authors would like to thank the Faculty of Computer and Mathematical Sciences and Institute of Quality and Knowledge Advancement (InQKA), UiTM for all the support given.

References

1. Ojha, T., Heileman, G.L., Martinez-Ramon, M., Slim, A.: Prediction of graduation delay based on student performance (2017)
2. Enhancing Academic Productivity and Cost Efficiency (University Transformation Programme Silver Book), Ministry of Education Malaysia (2016). http://mohe.gov.my/muat-turun/awam/penerbitan/university-transformation-programme/188-the-unitp-silver-book
3. Ibrahim, Z., Rusli, D.: Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. In: Proceedings 21st Annual SAS Malaysia Forum, pp. 1–6 (2007)
4. Dayioglu, M., Türüt-Asik, S.: Gender differences in academic performance in a large public university in Turkey. High. Educ. **53**(2), 255–277 (2007). https://link.springer.com/article/10.1007/s10734-005-2464-6
5. Amuda, B.G., Bulus, A.K., Joseph, H.P.: Marital status and age as predictors of academic performance of students of colleges of education in the Nort-Eastern Nigeria. Am. J. Educ. Res. **4**, 896–902 (2016). https://doi.org/10.12691/EDUCATION-4-12-7
6. Asif, R., Merceron, A., Ali, S.A., Haider, N.G.: Analyzing undergraduate students' performance using educational data mining. Comput. Educ. **113**, 177–194 (2017). https://doi.org/10.1016/j.compedu.2017.05.007
7. Herath, D.: Educational data mining to investigate learning behaviors : a literature review (2018). https://doi.org/10.13140/RG.2.2.20919.01446

8. Agrawal, R.S., Pandya, M.H.: Data mining with neural networks to predict students academic achievements. Int. J. Comp. Sci. Technol. **7**(2) (2016). http://www.ijcst.com/vol72/1/19-richa-shambhulal-agrawal.pdf
9. Mohammad Suhaimi, N., Abdul-Rahman, S., Mutalib, S., et al.: Review on predicting students' graduation time using machine learning algorithms. Int. J. Mod. Educ. Comput. Sci. **11**, 1–13 (2019). https://doi.org/10.5815/ijmecs.2019.07.01
10. Wah, Y.B., Ibrahim, N., Hamid, H.A., et al.: Feature selection methods: case of filter and wrapper approaches for maximising classification accuracy. Pertanika J. Sci. Technol. **26**, 329–340 (2018)
11. Akmal, E., Zaman, K., Farhan, A., et al.: Soft computing in data science. Soft Comput. Data Sci. **545**, 387–401 (2015). https://doi.org/10.1007/978-981-287-936-3
12. Cahaya, L., Hiryanto, L., Handhayani, T.: Student graduation time prediction using intelligent K-Medoids algorithm, pp. 263–266 (2017)
13. Pang, Y., Judd, N., O'Brien, J., Ben-Avie, M.: Predicting students' graduation outcomes through support vector machines. In: Proceedings - Frontiers in Education Conference FIE, October 2017, pp. 1–8 (2017). https://doi.org/10.1109/FIE.2017.8190666
14. Shariff, S.S.R., Rodzi, N.A.M., Rahman, K.A., et al.: Predicting the "graduate on time (GOT)" of Ph.D. students using binary logistics regression model. In: AIP Conference Proceedings (2016)
15. Mujani, W.K., Muttaqin, A., Khalid, K.A.: Historical development of public institutions of higher learning in Malaysia. Middle-East J. Sci. Res. **20**, 2154–2157 (2014). https://doi.org/10.5829/idosi.mejsr.2014.20.12.21113
16. Graduating on time is Malaysia's target. In: Afterschool.my (2015). https://afterschool.my/articles/graduating-on-time-is-malaysias-target. Accessed 14 Nov 2018
17. Ogwoka, T.M., Cheruiyo, W., Okeyo, G.: A model for predicting students' academic performance using a hybrid of K-means and decision tree algorithms. Int. J. Comput. Appl. Technol. Res. **4**, 693–697 (2015)
18. Jing, L.: Data mining and its applications in higher education. New Dir. Inst. Res. **2002**, 17 (2002)
19. Ma, X., Zhou, Z.: Student pass rates prediction using optimized support vector machine and decision tree. In: 2018 IEEE 8th Annual Computing Communication Workshop Conference CCWC 2018, Janua 2018, pp. 209–215 (2018). https://doi.org/10.1109/CCWC.2018.8301756
20. Athani, S.S., Kodli, S.A., Banavasi, M.N., Hiremath, P.G.S.: Student academic performance and social behavior predictor using data mining techniques. In: Proceeding - IEEE International Conference Computing Communication Automation ICCCA 2017, Janua 2017, pp. 170–174 (2017). https://doi.org/10.1109/CCAA.2017.8229794
21. Al-Shehri, H., Al-Qarni, A., Al-Saati, L., et al.: Student performance prediction using support vector machine and K-Nearest neighbor. In: Canadian Conference on Electrical and Computer Engineering, pp. 17–20 (2017). https://doi.org/10.1109/CCECE.2017.7946847
22. Lee, S.J., Kim, J.M., Lee, W.G.: Analysis of factors affecting achievement in maker programming education in the age of wireless communication. Wirel. Pers. Commun. **93**, 187–209 (2017). https://doi.org/10.1007/s11277-016-3450-2
23. Asyraf, A.S., Abdul-Rahman, S., Mutalib, S.: Mining textual terms for stock market prediction analysis using financial news. Commun. Comput. Inf. Sci. **788**, 293–305 (2017). https://doi.org/10.1007/978-981-10-7242-0_25
24. Marbán, Ó., Mariscal, G., Segovia, J.: A data mining & knowledge discovery process model. In: Ponce, J., Karahoca, A. (eds.) Data Mining and Knowledge Discovery in Real Life Applications, February 2009, pp. 438–453. I-Tech, Vienna, Austria (2009). ISBN 978-3-902613-53-0
25. Frank, E., et al.: The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques, 4th edn. Morgan Kaufmann Press, Burlington (2016)