International Journal of Innovation and Industrial Revolution IJIREV
EISSN: 2637-0972

Volume 5 Issue 15 (December 2023) PP. 120-130
DOI 10.35631/IJIREV.515013

International Journal of Innovation and Industrial Revolution (IJIREV)
eISSN: 2637-0972

INTERNATIONAL JOURNAL OF INNOVATION AND INDUSTRIAL REVOLUTION (IJIREV)
www.ijirev.com

MACHINE LEARNING FOR PREDICTING STUDENTS' ACADEMIC ACHIEVEMENT BASED ON LEARNING STYLE AND ACADEMIC RESULTS

Nazrina Bakar[1*], Neelam Amelia Mohamad Rejeni[2], Anding Nyuak[3]

[1] Department of Information Technology and Communication, Politeknik Kuching Sarawak, Malaysia
Email: nazrina@poliku.edu.my
[2] Politeknik Kuching Sarawak, Malaysia
Email: neelam.amelia@poliku.edu.my
[3] Politeknik Kuching Sarawak, Malaysia
Email: andingnyuak@poliku.edu.my
[*] Corresponding Author

Article Info:

Article history:
Received date: 01.11.2023
Revised date: 05.11.2023
Accepted date: 01.12.2023
Published date: 31.12.2023

To cite this document:

Ahmad, M. N. (2023). Machine Learning For Predicting Students' Academic Achievement Based On Learning Style And Academic Results. *International Journal of Innovation and Industrial Revolution, 5* (15), 120-130.

DOI: 10.35631/ IJIREV.515013

Abstract:

The objective of this study is to explore the viability of utilizing machine learning methods in forecasting students' academic success by examining their unique approaches to learning. The data was obtained by administering a questionnaire that employed the Motivated Strategies for Learning Questionnaires (MSLQ), along with the academic records of Semester 4 students from Jabatan Teknologi Maklumat dan Komunikasi (JTMK) Kuching Sarawak. After pre-processing the data and devising relevant features, we proceeded to train and test our machine-learning model. By employing Linear Regression, Decision Tree Regressor, Random Forest Regressor, Lasso, and Ridge, we were able to accurately anticipate students' performance based on their learning styles. Our study shows that the Random Forest and the Ridge have the most accuracy in predicting students' performance with MSE = 0.11, MAE = 0.22, and RMSE 0.32 for both models. The findings demonstrate the potential of machine learning models in accurately forecasting academic achievement, thereby offering valuable insights to educators who wish to personalize their teaching methods and intervention strategies. In summary, this study underscores the capacity of machine learning techniques in education to optimize learning outcomes and ensure the academic success of students.

Keywords:

Machine Learning, Student's Academic Achievement

Introduction

The surge of big data provides novel approaches to managing it. Conventional approaches and productivity utilities are incapable of handling and assessing extensive data. Enterprise leaders need to embrace innovative technologies and methodologies like machine learning that align with their information requirements to leverage big data's potential. This knowledge can be applied to both educational and production settings to assist decision-makers in achieving the highest level of precision. To guarantee the timely graduation of students, educators must take the lead. Utilizing machine learning, this research examined the correlation between student achievements and outcomes, learning preferences, individual traits, and parental approaches.

This study aimed to explore the relationship between student achievement and a range of factors, including learning preferences, individual traits, and parental involvement. By harnessing the power of machine learning, educators can gain valuable insights from extensive data sets. These insights can inform decision-making processes and contribute to the timely graduation of students. Additionally, implementing innovative technologies and methodologies aligned with information requirements allows enterprises to unlock the full potential of big data and optimize productivity in both educational and production settings.

Literature Survey

Based on the previous research outcomes presented by Sandra and her team in 2021, it has been established that students' future academic performance can be predicted by analyzing their past academic records, including grades, and considering factors such as their unique learning styles, personalities, and parental involvement. Yac, in 2022, further delved into this area and devised a method to estimate students' final exam scores using midterm results. To validate the accuracy of these predictions, the author employed various machine learning techniques. Chauhan et al (2019) suggested an enhanced aspect, wherein accomplished individuals in a particular subject are designated as tutors. These tutors could be classmates or senior students. Along with tutor information, they considered the dataset encompassing previous academic outcomes for predicting performance. Their findings indicated that the Multiple Linear Regression Model offers the most favorable solution.

Zohair (2019) conducted a study aimed at assisting vulnerable students and ensuring their continued enrollment. The objective was to establish a reliable accuracy rate within a limited dataset. The research involved investigating the potential identification of significant factors within the small dataset. These factors would then be employed to construct a predictive model, leveraging visualization and clustering algorithms. The findings of the study demonstrated the effectiveness of support vector machine and discriminant analysis algorithms in training small datasets and achieving satisfactory rates of accuracy and reliability in classification tests. Dhilipan et al (2021) presented a model that utilizes students' 10th, 12th, and prior semester performance to assist them in identifying their ultimate grade and boosting their academic behavior. The study employed Binomial logistic regression, Decision tree and Entropy, and KNN classifier methods. Altabrawee et al. (2019) emphasized the importance of proactive support and suitable measures to enhance student performance and achieve their goals. They used four machine learning techniques to forecast students' performance in a computer science subject, considering the impact of time spent on social networks. The fully connected feed-forward multilayer ANN showed the most promising results, achieving a performance score of 0.807 and the highest classification accuracy (77.04%) among all models tested. Pintrich et al. (1993) conducted assessments of the MSLQ questionnaires' reliability and validity, revealing that both the general framework and the scales serve as valid and reliable instruments to evaluate student motivation and classroom learning strategies. In this study, the MSLQ and students' past academic performance were analyzed to forecast their future achievement.

Method and Dataset

The extracted data from the Sistem Pengurusan Maklumat Politeknik (SPMP) serves as a valuable resource for this research. It provides comprehensive student records that are stored on a server situated at Politeknik Kuching Sarawak. This centralized information hub enables efficient access to crucial data, facilitating the analysis and exploration of various educational aspects. With a wealth of records at our disposal, our investigation aims to delve deeper into pertinent factors that shape student outcomes and contribute towards enhancing the educational experience at Politeknik Kuching Sarawak. The data we possess consists of semester outcomes, Sijil Pelajaran Malaysian (SPM) results for English and Mathematics, and the MSLQ.

The process of feature selection entails reducing the number of variables employed in forecasting a specific result. The aim is to enhance the model's comprehensibility, decrease intricacy, improve algorithmic computational efficiency, and prevent overfitting. Furthermore, the selected features undergo a rigorous evaluation process to ensure their relevance and effectiveness in achieving the desired outcome. This evaluation involves various statistical techniques, such as correlation analysis to assess the relationship between each feature and the target variable. By identifying and retaining the most influential attributes, the overall predictive power of the model is maximized, allowing for accurate and reliable forecasts. Once the feature selection process is complete, attention shifts toward developing an appropriate modeling technique. Several algorithms are considered, each with its strengths and limitations, catering to different types of data and prediction tasks. Whether it be regression, classification, or clustering, the choice of algorithm heavily depends on the nature of the problem at hand. In order to ensure the chosen algorithm performs optimally, model training is conducted using a well-curated dataset. This involves splitting the available data into training and validation sets, allowing the model to learn from the training set while validating its performance on unseen data. Iterative refinement techniques, such as cross-validation, are often employed to fine-tune model parameters, prevent overfitting, and improve generalization capabilities. Once the model has been adequately trained, it is ready for deployment and utilization. Through the application of new or unseen data, the model's predictive capabilities are put to the test, providing valuable insights and enabling intelligent decision-making. Continuous monitoring and evaluation of the deployed model's performance ensure its accuracy and effectiveness in real-world scenarios as shown in Figure 1.

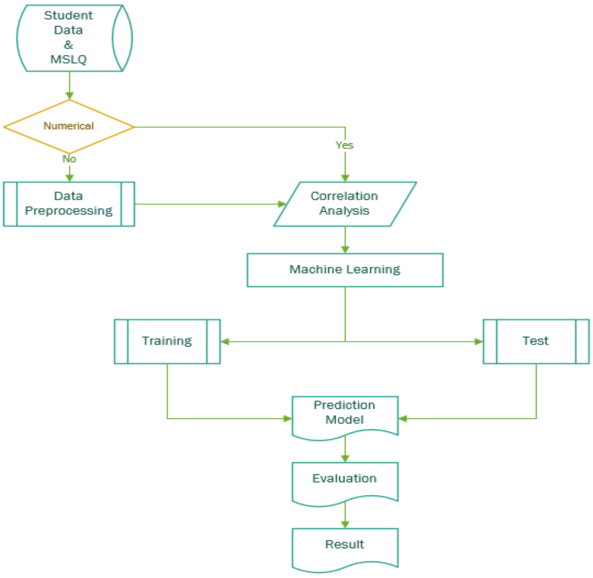


Figure 1: Steps in Machine Learning

Table 1: Sijil Pelajaran Malaysia Grade and Value

SPM GRADES	GRADE VALUE
A+	9
A	8
A-	7
B+	6
B	5
C+	4
C	3
D	2
E	1
G	0

Motivated Learning Strategy Questionnaire

The MSLQ was created for researchers to gauge student motivation and learning strategies, and for instructors and students to evaluate motivation levels and study techniques in a particular course, in our case study to assess students in Diploma of Digital Technology. Students assess themselves using a 7-point Likert scale, ranging from 1 (not applicable to me at all) to 7 (very applicable to me). To obtain scores for specific subcategories, the average of the items within that particular subcategory is calculated (Hilpert et al., 2013).

Table 2: Motivated Learning Strategies Questionnaire Components

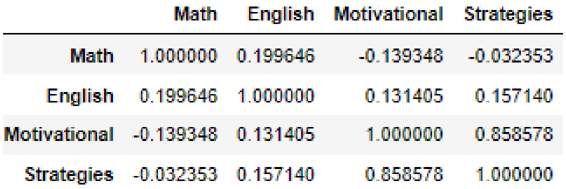
Motivational Beliefs	Self-Regulated Learning Strategies
Self-Efficacy	Cognitive Strategy Use
Intrinsic Value	Self-Regulation
Test Anxiety	

Source: Pintrich & De Groot, (1990).

Result and Analysis

Pearson Correlation

The Pearson product-moment correlation coefficient, often referred to as the Pearson correlation coefficient, is a metric that gauges the intensity of a linear connection between two variables. It is symbolized as 'r' (Zhou et al., 2016). If the association between the two variables is robust, the Pearson correlation coefficient 'r' will approach either +1 or -1, contingent on whether the relationship is positive or negative, respectively. Illustrated in Figure 2, specifically in the Pearson Correlation Matrix, there is a notable correlation between motivation and strategies. The previous outcomes for Mathematics in the SPM exams demonstrated an inverse relationship with both levels of motivation and employed strategies. According to the findings outlined in the study conducted by Polychroni et al. (2012), it is evident that students' attitudes towards their classroom and overall school environment play a pivotal role in influencing their drive to excel in the subject of mathematics. This emphasizes the significance of cultivating an environment that fosters positivity, engagement, and a sense of purpose within educational institutions. When students perceive their learning environment as supportive and conducive to their academic growth, it substantially bolsters their motivation to tackle mathematical challenges with vigor and determination. Furthermore, it highlights the importance of implementing effective teaching methods and strategies that not only transmit knowledge but also inspire a genuine interest in the subject matter (Liu et al., 2022). In light of these insights, educators and policymakers should consider how they can proactively shape the learning environment to nurture students' motivation and enhance their strategies for approaching mathematical concepts. By doing so, we can pave the way for improved outcomes in Mathematics education and equip students with the tools they need to succeed not only in exams but also in their future endeavors (Tuazon & Torres, 2022).



	Math	English	Motivational	Strategies
Math	1.000000	0.199646	-0.139348	-0.032353
English	0.199646	1.000000	0.131405	0.157140
Motivational	-0.139348	0.131405	1.000000	0.858578
Strategies	-0.032353	0.157140	0.858578	1.000000

Figure 2: Pearson Correlation Matrix

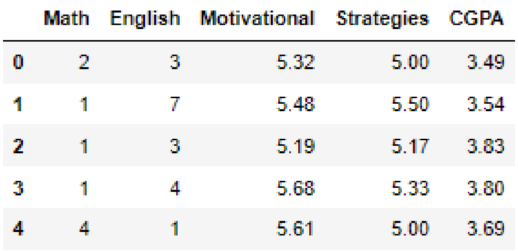
Data Preparation

Data pre-processing is a critical phase in the machine learning pipeline that lays the foundation for accurate and reliable model training. After gathering the dataset, the next step is to clean and prepare it for analysis. This involves a series of steps aimed at ensuring the data is in a format that the machine learning algorithms can effectively learn from. One of the initial steps in data pre-processing is handling missing values (Alboaneen et al., 2022). This can be done through techniques like imputation, where missing values are filled in using the mean, median, or mode of the respective feature. Alternatively, in some cases, dropping rows with missing values might be appropriate if the amount of missing data is negligible. Next, categorical variables need to be converted into a numerical format. This is achieved through techniques like one-hot encoding, which creates binary columns for each category, or label encoding, which assigns a unique numerical label to each category. This transformation ensures that the machine learning algorithms can process the data correctly. Standardization or normalization is another crucial step. This involves scaling the features so that they all have similar scales.

Outliers can significantly impact the performance of a machine-learning model. Identifying and handling outliers is essential. Finally, the dataset is typically split into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance on unseen data. This helps assess how well the model generalizes to new, unseen instances. Data pre-processing is a crucial step in preparing the dataset for machine learning. It involves handling missing values, converting categorical variables, standardizing or normalizing features, addressing outliers, selecting relevant features, and splitting the data into training and testing sets. A well-prepared dataset sets the stage for effective model training and ultimately leads to more accurate and reliable predictions.

Display The Dataset

A snapshot of the dataset is shown in Figure 3 below.



	Math	English	Motivational	Strategies	CGPA
0	2	3	5.32	5.00	3.49
1	1	7	5.48	5.50	3.54
2	1	3	5.19	5.17	3.83
3	1	4	5.68	5.33	3.80
4	4	1	5.61	5.00	3.69

Figure 3: Dataset For Predicting Students' Performance

Distribution or Density Plot

A distribution or density plot illustrates how data is spread across a continuous range. It resembles a smoothed version of a histogram and provides a visual representation of data distribution over a continuous interval. Consequently, a density plot offers insights into the likely distribution of the entire population as well. As shown in Figure 4, the dataset is not normally distributed for our case study.

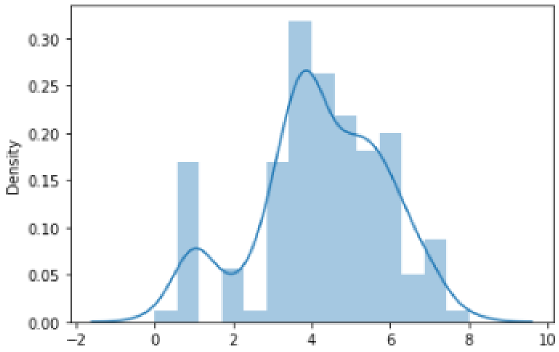


Figure 4: Dataset Distribution

Build A Machine Learning Model

Following this, the subsequent phase involves constructing a machine-learning framework. In our investigation, we employed various algorithms including Linear Regression, Decision Tree Regressor, Random Forest Regressor, Lasso, and Ridge, to forecast student achievement. Each of these models brings distinct strengths to the predictive process. Linear Regression offers simplicity and interpretability, while Decision Tree Regressor excel in capturing non-linear relationships within the data (El Aissaoui et al., 2019). Random Forest Regressor, on the other hand, harnesses the power of ensemble learning to enhance prediction accuracy. Additionally, Lasso and Ridge regularization methods introduce a layer of robustness against overfitting, further refining the predictive capabilities of our model. By employing this diverse set of algorithms, we aim to comprehensively analyze and understand the factors influencing student performance. This multifaceted approach not only enriches the accuracy of our predictions but also provides valuable insights into the complex dynamics at play in the realm of education.

Implementation and Evaluation

Numerous regression models utilize distance metrics to ascertain how closely they approach the optimal outcome. Additionally, it is imperative to quantify what constitutes the "best" result. The metrics commonly employed for this purpose include Mean Average Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics provide quantitative measures to evaluate the accuracy and performance of the model in predicting or fitting data. MAE represents the average of the absolute differences between predicted and actual values. MSE calculates the average of the squared differences, giving more weight to larger discrepancies. RMSE is the square root of MSE and provides a measure in the original units of the data, which can be easier to interpret. These metrics collectively assist in assessing and fine-tuning regression models for optimal performance.

Table 3 shows the comparison among the five selected machine learning models used in this study: Linear Regression, Decision Tree Regressor, Random Forest Regressor, Lasso, and Ridge. Among these models, Random Forest and Ridge models exhibited the lowest values for all three metrics: MSE (Mean Squared Error) = 0.11, MAE (Mean Absolute Error) = 0.22, and RMSE (Root Mean Square Error) = 0.32. This indicates that these models provided the most accurate predictions on average.

Table 3: Machine Learning Comparison

Machine Learning	MSE	MAE	RMSE
Linear Regression	0.11	0.22	0.33
Decision Tree	0.15	0.28	0.38
Random Forest	0.11	0.22	0.32
Lasso	0.13	0.24	0.36
Ridge	0.11	0.22	0.32

Visualize The Machine-Learning Model

Analyzing Figures 5, 6, 7, 8, and 9, it is apparent that there are not substantial differences among the four machine learning models in terms of graph visualization, with the exception of the Lasso model. The Lasso model might have a different pattern in its predictions, suggesting a distinct approach to capturing the underlying relationships within the data. To further validate the models, it's imperative to compare the actual and predicted values, as illustrated in Figure 10. This visual representation allows for a direct assessment of how well each model aligns with the observed data. A close match between the predicted and actual values indicates a robust predictive capability, while discrepancies may highlight areas for improvement or further investigation.

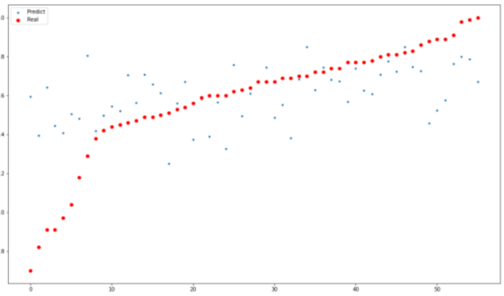


Figure 5: Linear Regression Model

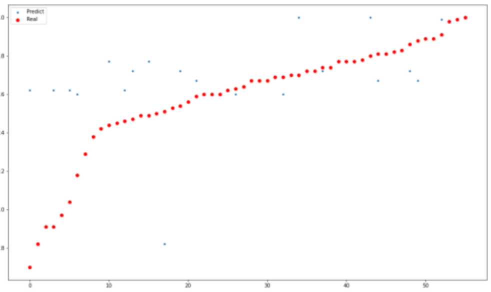


Figure 6: Decision Tree Model

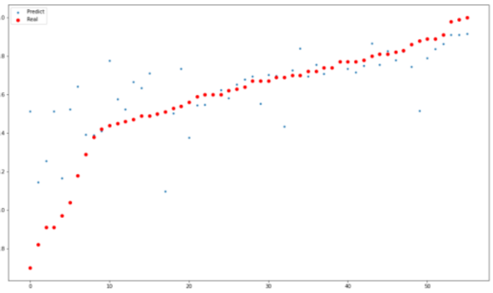


Figure 7: Random Forest Model

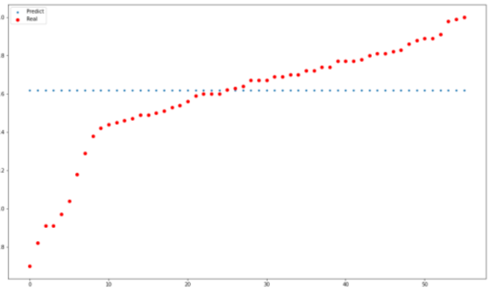


Figure 8: Lasso Model

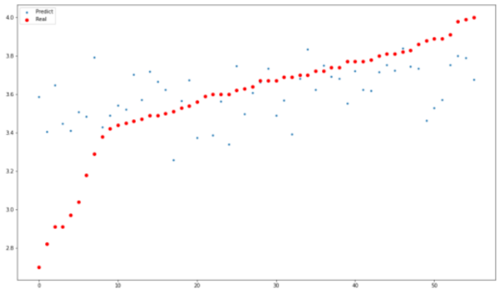


Figure 9: Ridge Model

Validating the Machine Learning Model

Random Forest	Ridge

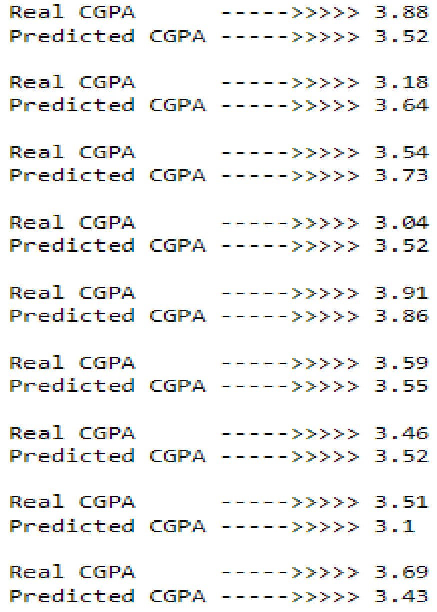
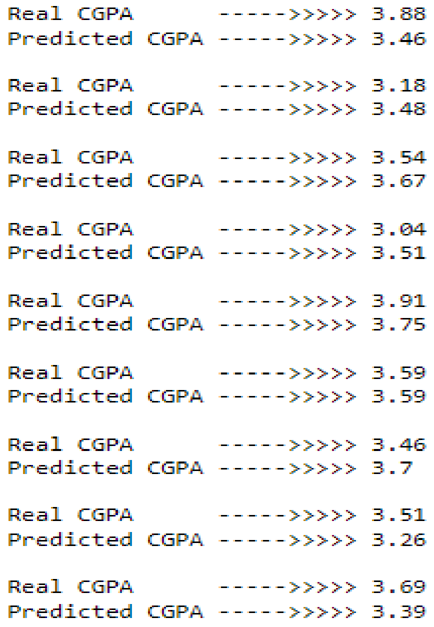

`Real CGPA      -----»»»» 3.88`
`Predicted CGPA -----»»»» 3.52`
`Real CGPA      -----»»»» 3.18`
`Predicted CGPA -----»»»» 3.64`
`Real CGPA      -----»»»» 3.54`
`Predicted CGPA -----»»»» 3.73`
`Real CGPA      -----»»»» 3.04`
`Predicted CGPA -----»»»» 3.52`
`Real CGPA      -----»»»» 3.91`
`Predicted CGPA -----»»»» 3.86`
`Real CGPA      -----»»»» 3.59`
`Predicted CGPA -----»»»» 3.55`
`Real CGPA      -----»»»» 3.46`
`Predicted CGPA -----»»»» 3.52`
`Real CGPA      -----»»»» 3.51`
`Predicted CGPA -----»»»» 3.1`
`Real CGPA      -----»»»» 3.69`
`Predicted CGPA -----»»»» 3.43`
	
	`Real CGPA      -----»»»» 3.88`
	`Predicted CGPA -----»»»» 3.46`
	`Real CGPA      -----»»»» 3.18`
	`Predicted CGPA -----»»»» 3.48`
	`Real CGPA      -----»»»» 3.54`
	`Predicted CGPA -----»»»» 3.67`
	`Real CGPA      -----»»»» 3.04`
	`Predicted CGPA -----»»»» 3.51`
	`Real CGPA      -----»»»» 3.91`
	`Predicted CGPA -----»»»» 3.75`
	`Real CGPA      -----»»»» 3.59`
	`Predicted CGPA -----»»»» 3.59`
	`Real CGPA      -----»»»» 3.46`
	`Predicted CGPA -----»»»» 3.7`
	`Real CGPA      -----»»»» 3.51`
	`Predicted CGPA -----»»»» 3.26`
	`Real CGPA      -----»»»» 3.69`
	`Predicted CGPA -----»»»» 3.39`

Figure 10: Validating Model with Real and Predicted Value

Validating And Visualizing Between Random Forest And Ridge Model.

We plotted a graph between the Random Forest and Ridge model to compare the two models. With all data points, we can see that the Ridge regression line fits the model more accurately than the Random Forest line, as depicted in Figure 11. The Ridge regression line hugs the data points closely, indicating a stronger linear relationship between the predictor variables and the target variable. Upon closer examination, it's evident that the Random Forest line exhibits more fluctuation, reflecting its inherent ability to capture complex interactions within the data. However, this complexity also leads to a slightly less precise fit to the data compared to the Ridge model. The trade-off between complexity and accuracy is a fundamental consideration in model selection.

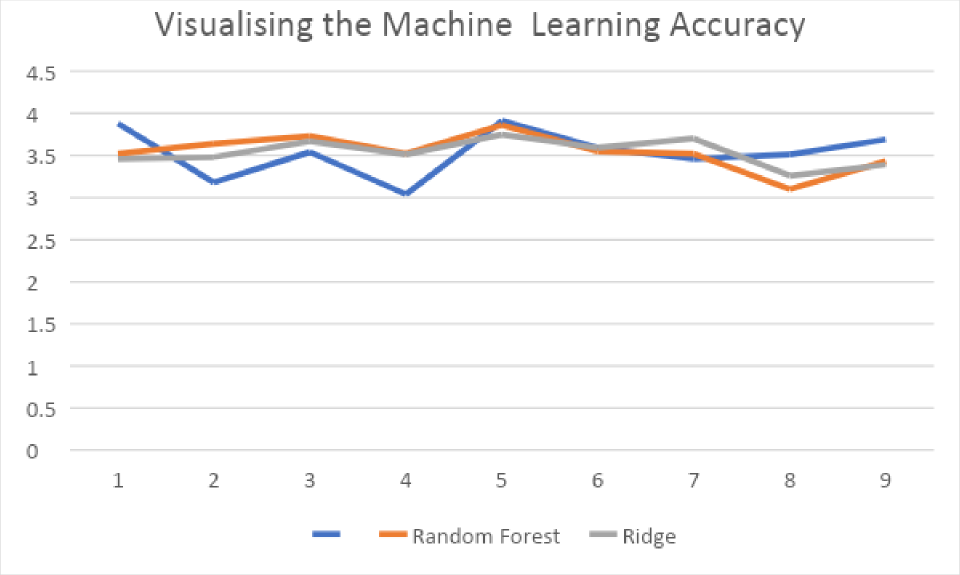


Figure 11: Comparison between Random Forest and Ridge Model

Conclusion

In conclusion, the comprehensive evaluation of these machine learning models suggests that Random Forest and Ridge models outperform the others in terms of predictive accuracy for predicting students' performance in JTMK. However, the choice of model should also consider other factors such as interpretability, computational efficiency, and the specific requirements of the problem at hand.

Moving forward, it may be beneficial to explore ensemble methods or fine-tune hyperparameters to further optimize model performance. Additionally, conducting cross-validation or assessing the models on an independent test set can provide a more robust evaluation of their generalization capabilities. These steps will contribute to building a reliable and robust predictive model for the given dataset.

Acknowledgement

We would like to extend our gratitude to students and staff of Department of Information Technology and Communication, Politeknik Kuching Sarawak whose contributions were instrumental in the completion of this case study. Special thanks to our families and colleagues for their understanding, encouragement, and support throughout the completion of this case study.

References

Abdul Rahim Hamdan & Hayazi Mohd Yasin (2010). Penggunaan Alat Bantu Mengajar (ABM) Di Kalangan Guru-Guru Teknikal Di Sekolah Menengah Teknik Daerah Johor Bahru, Johor. Fakulti Pendidikan, Universiti Teknologi Malaysia

Ahamad Sipon. (2005). Perutusan Hari Guru 2005. http://www.moe.edu.my/hariguru/perutusanKPPM.htm [20 Feb 2009].

Alizah Lambri (2015) Pelaksanaan Pendekatan Pembelajaran Berpusatkan Pelajar Di Sebuah Universiti Awam. Tesis Doktor Falsafah. Universiti Kebangsaan Malaysia.

A. Kulop Saad, A. Ahamad. (2000). Keberkesanan Bahan Pengajaran Multi Interaktif (BPMI) dalam Pengajaran. Jurnal BTP, 17–34.

Dale. (1969). Audiovisual Methods in Teaching. Third Edition. The Dryden Press, New York.

Hanifah Mahat, Mohamad Suhaily Yusri & Izza Ahmad. (2015). Kajian tahap amalan kelestarian dalam kalangan murid prasekolah kementerian pendidikan Malaysia.

Kamalian, A.Rashki, M & Arbabi, M.L (2011). Barriers to innovation among Iranian SMEs. Asian Journal of Business Management, 3(2):79-90

Mok Soon Sang. (2008). Pedagogi Untuk Pengajaran dan Pembelajaran. Selangor: Penerbitan Multimedia Sdn Bhd.

Mohamed Khaled Nordin. (2010). Pelancaran pelan tindakan fasa 2 PSPTN. Kementerian Pengajian Tinggi, Malaysia.

Noraziah Md Yusop & Latipah Sidek. (2010). Pendidikan alam sekitar dalam pendidikan Islam.

Noor Azlan, A. Z., & Nurdalina, D. (2010). Penggunaan Bahan Bantu Mengajar Di Kalangan Guru Pelatih UTM Yang Mengajar Mata Pelajaran Matematik. Universiti Teknologi Malaysia, 1-6.

Nur Fadzilah Othman. (2010). Tahap Penggunaan Aplikasi Web 2.0 dalam Kalangan Pelajar Institusi Pengajian Tinggi Awam di Malaysia. Tesis Sarjana. Universiti Teknologi Malaysia.

Nur Syafiqah, I., & Nurul Nazirah, M. I. M. (2018). Keberkesanan penggunaan grafik berkomputer sebagai alat bahan bantu mengajar dalam kalangan pelajar reka bentuk dan teknologi. Jurnal Sains Humanika, 10(3-3), 81-87

R. Hamdan, H. Mohd Yasin. (2010). Amalan Penggunaan Alat Bantu Mengajar (ABM) di Kalangan Guru-Guru Teknikal di Sekolah Teknik Daerah Johor Bharu. 1–8.

Saifuddin, M., & Muhammad Idham. (2017). Strategi belajar mengajar. Indonesia: Syiah Kuala University Press.

Siohng Tih dan Zuraidah Zainol. (2012). Minimizing waste and encouraging green practices. Jurnal ekonomi Malaysia. 46(1) 157-164.

Trilling, B., & Fadel, C. (2009). 21 st century skills: Learning for life in our times. San Francisco: Jossey-Bass.

Umi Nadiha Mohd Nor, Zamri Mahamod dan Jamaludin Badusah. (2011). Penerapan Kemahiran Generik Dalam Pengajaran Guru Bahasa Melayu Sekolah Menengah. Jurnal Pendidikan Bahasa Melayu, Vol. 1 (2) November 2011. ISSN:2180-4842.

UNESCO. (2002). Education for Sustainable Development. Retrieved from http://www.unesco.org/en/esd/ [July 15 2020].

Walker, S.E. (2003). Active Learning Strategies to Promote Critical Thinking. Journal of Athletic Training, 38 (3): 263-267