FINAL GRADE PREDICTION MODEL BASED ON STUDENT'S ALCOHOL CONSUMPTION

M. Rangga Ramadhan Saelan [1*]; Siti Masturoh [2]; Taopik Hidayat [3]; Siti Nurlela [4]; Risca Lusiana Pratiwi [5]; Muhammad Iqbal [6]

[1,3]Sains Data, [2,4,5]Sistem Informasi, [6]Teknik Informatika
[1,2,3,4,5]Universitas Nusa Mandiri, [6]Universitas Bina Sarana Informatika
[1,2,3,4,5]www.nusamandiri.ac.id, [6]www.bsi.ac.id
rangga.mgg@nusamandiri.ac.id[1*], siti.uro@nusamandiri.ac.id[2], taopik.toi@nusamandiri.ac.id[3], siti.sie@nusamandiri.ac.id[4], risca.ral@nusamandiri.ac.id[5], iqbal.mdq@bsi.ac.id[6]

Abstract— *The influence of alcohol consumption and several other factors that are estimated to have a role on the level of learning performance of adolescents who are still in school, in this work, research is carried out on public data that has been obtained. This work trains a model on a dataset that provides information about student grades from the first year to the final year. In this work, we will focus on the final year mark in Portuguese as a label, with several variables as factors that influence the final grade as a reference for learning performance. on a group of students who are the target of research. Using machine learning techniques by training several models to predict the final grade as a reference for student learning performance through the final grade results obtained. By training several machine learning models to predict final year grades (G3) from Portuguese lessons by doing a comparative method comparing the Support Vector Regressor (SVR) and Random Forest (RF) models. All models have hyperparameters that must be adjusted using Cross Validation. The best models for predicting G3 are SVR and RF, and have a mean absolute error (MAE) of about 2.24 and 2.25, respectively. Through the MAE plot, the SVR and RF models work well. However, By analyzing the distribution of errors made by both models. Through this work, it can be concluded that SVR is more balanced, i.e. has a better ratio between underestimation and overestimation, while RF performs better on outliers.*

Keywords*: Model, MAE, Parameter, Machine, Learning, Value*

Intisari— Pengaruh konsumsi alcohol dan dan beberapa faktor lainnya yang diperkirakan memiliki peran terhadap tingkat kinerja belajar remaja yang masih bersekolah, pada pekerjaan ini dilakukan penelitian terhadap data publik yang telah didapatkan. Pekerjaan ini melatih model terhadap dataset yang menyediakan informasi mengenai nilai-nilai pelajar dari tahun pertama hingga tahun akhir, Pada pekerjaan ini akan berfokus kepada nilai tahun akhir bahasa portugis sebagai label, dengan beberapa variable sebagai faktor yang berpengaruh terhadap nilai akhir yang menjadi acuan kinerja belajar pada sekumpulan pelajar yang menjadi sasaran penelitian. Menggunakan teknik *machine learning* dengan melatih beberapa model untuk memprediksi nilai akhir sebagai acuan kinerja belajar pelajar melalui hasil nilai akhir yang didapatkan. Dengan melatih beberapa model *machine learning* untuk memprediksi nilai tahun akhir (G3) dari pelajaran bahasa portugal dengan melakukan metode komparatif membandingkan model *Support Vector Regressor* (SVR) dan *Random Forest* (RF). Semua model memiliki hyperparameter yang harus disesuaikan menggunakan *Cross Validation*. Model terbaik untuk memprediksi G3 adalah SVR dan RF, dan memiliki *mean absolute error* (MAE) masing-masing sekitar 2,24 dan 2,25. Melalui plot MAE, model SVR dan RF bekerja dengan baik. Tetapi, Dengan menganalisis distribusi kesalahan yang dibuat oleh kedua model. Melalui pekerjaan ini, dapat disimpulkan bahwa SVR lebih seimbang, yaitu memiliki rasio yang lebih baik antara nilai yang diremehkan dan ditaksir terlalu tinggi, sementara RF berkinerja lebih baik pada *outlier*.

Kata Kunci*:* Model, MAE, Parameter, *Machine, Learning*, Nilai

Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3056

INTRODUCTION

The seriousness, totality and focus of adolescents in the learning period at school are the main problems at this time to organize the younger generation for the future. On the other hand, alcohol consumption needs to get great attention from the relevant authorities, from the government, social institutions, and families. This is because alcohol consumption is currently not only a target for consumption among adults, but also a target for teenagers who are still relatively young where they should focus more on their learning period. Thus, with the increasing percentage of consumption of alcohol by adolescents who are studying this will become a disorder or obstacle in the process of receiving knowledge gained from educational institutions. In 2016, the National Institutes of Health reported that 26% of 8th graders, 47% of 10th graders, and 64% of 12th graders had experience consuming alcoholic beverages (Palaniappan et al., 2017).

Data mining (DM) has been applied in the field of education, and is an emerging interdisciplinary research field also known as Educational Data Mining (EDM). One of the goals of the EDM is to better understand how to predict the academic performance of students given personal, socio-economic, psychological and other environmental attributes. Another goal is to identify factors and rules that affect the academic outcomes of education (Satyanarayana & Nuckowski, 2016).

To determine the effect of alcohol consumption and several other factors that are estimated to have a role in the level of learning performance of adolescents who are still in school, a study is currently being carried out on public data that has been obtained using machine learning techniques by training several models to predict the final value as a reference for student learning performance. Alcohol consumption is reported by 85% of students. About 70% of first- and third-year students and 47% of sixth-year students are motivated by socializing with peers. Alcohol consumption is widespread in those who engage in physical activity (93%) and live with family (89%) (Freire et al., 2021). Of course, there is a correlation between the level of alcohol consumption and environmental factors and so on, therefore it is necessary to conduct research to find out the relationship between one factor and another, as well as the influence of certain factors that play a very large role in the level of alcohol consumption by students so that it has an impact on learning performance in the form of final grade results (G3). Alcohol consumption can be influenced by several factors such as heredity/habits, environment, mental health. Alcohol consumption in learners has a long-term effect on students' brain and learning performance (Nur'artavia, 2017). Alcoholic beverages themselves Ethanol is a psychoactive ingredient that can reduce the consciousness of its consumers (Wijaya, 2016). In particular, students who have consumed excessive alcohol tend to have difficulties related to brain memory and the ability to focus. Alcohol addiction can be influenced by several factors such as genetics, social environment, and mental health. Alcohol consumption at a young age (learners) has a long-term effect on students' brain and academic performance (Sagala & Tampubolon, 2018).

The purpose of the following work is to train some machine learning models so that later it can be used to predict the final Portuguese grades of students attending two schools in Portugal. In the Portuguese undergraduate system, the value is between 0 and 20. By analyzing exploratory data (EDA) from the available data by focusing on the label, namely the end-of-year value (G3). Then in this work will train several models so that the best model will be known to later be able to predict (G3). Then the last goal is to analyze the distribution of errors committed by the two models, and the latter gives us a discriminant to choose one of the two models.

Some machine learning models will be trained on the training dataset, this is done to get the best model. This related research has been carried out by several research literacy, including research conducted by sagala et al (Sagala & Tampubolon, 2018) "This study aims to apply and perform performance analysis of data mining algorithms to predict alcohol consumption and analyze related factors in intermediate level students. The stages carried out are data pre-processing, feature selection, classification, and model evaluation. At the preprocess stage, some features are transformed into appropriate shapes to facilitate the classification process. The results showed that the classification model built using Naïve Bayes had the highest accuracy value using the 5 best features of the Gain Ratio. In addition, the use of the feature selection method is able to improve the performance of all classifiers in general."

Further research conducted by Pisutaporn et al. (Pisutaporn et al., 2018) "In this paper, we have studied student alcohol consumption and identified factors that have a significant impact on students' alcohol consumption. It was found that men tend to have more alcohol consumption rates than women. The high rate of going out with friends leads to high alcohol consumption. In addition, students who have a short weekly study

time, no school support for additional education and do not have the desire to go on to higher education are more likely to have more alcohol consumption. We also found that the random forest algorithm performed better than the decision tree algorithm for this classification problem. Finally, we have confirmed the negative relationship between the level of alcohol consumption and the value of students."

The purpose of this study is to train several machine learning models to predict the final year value of the Portuguese language by conducting a comparative method comparing the Support Vector Regressor (SVR) and Random Forest (RF) models so that the best model will be obtained to predict.

Literature Review

Machine Learning

ML is defined as a scientific field that gives machines the ability to learn without being strictly programmed (Liakos et al., 2018).

Machine learning (ML) is used to teach machines how to handle data more efficiently. Many industries are applying machine learning to extract relevant data. The purpose of machine learning is to learn from data. A lot of research has been done on how to make machines learn on their own without being explicitly programmed. Many mathematicians and programmers apply several approaches to find a solution to this problem that has a very large data set (Batta, 2020).

Support Vector Regression

The purpose of the SVR is to find a linear regression equation suitable for all sample points and minimize the total sample point variance of this regression hyperplane. here is an example of a training set

$E = \{(x_i, y_i) | i = 1,2,..n\}, x_i \in R^n, y_i \in R$. Function $f(x_i)$ investigated on R^n, in such a way, that $y_i = f(x_i)$, and there is always value y which is appropriate for each input x. Such a function $f(x_i)$ called the regression function, and $f(x_i)$ can be described as follows.

$$f(x_i) = \omega . \emptyset (x_i) + b, \quad (1)$$

where $\omega \in R^n$ is *Weight Vector*, $\emptyset(x_i)$ is a nonlinear mapping that serves to map data from space R^n to the higher feature space, and b is biased. Equation (1) can be attached to all sample points with precision.

$$|y_i - [\omega . \emptyset(x_i) + b]| \leq \varepsilon, \; i = 1,2, \ldots n \quad (2)$$

Because there is a certain installation error, the slack variable $(\xi_i, \xi_i *)$ and penalty parameter C is introduced. Regression adjustment issues are turned into optimization problems (Tang et al., 2022).

Random Forest

Random Forest (RF) is an algorithm that uses recursive binary separation methods to reach the final node in a tree structure based on classification and regression trees (Yoga Religia et al., 2021)

Derived from the theory of learning ensembles, RF combines several individual Decision Trees (DTs). Due to simplification and nonparametric behavior, classification and regression trees (CART) are commonly used as DT in RF Each DT relies on a random bootstrap dataset (Liu et al., 2021).

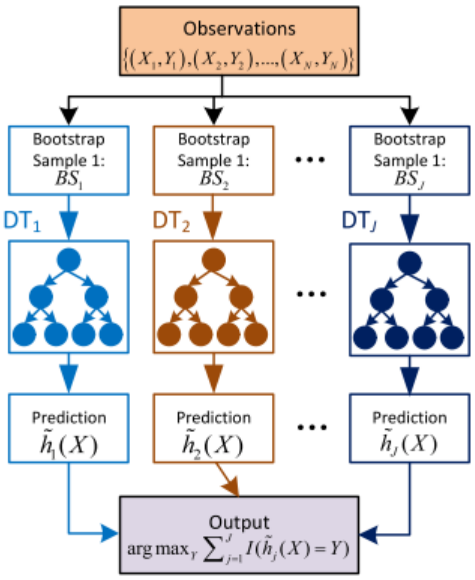

Figure 1. Classification Structure of Random Forest (Liu et al., 2021)

Breiman in 2001 introduced the RF algorithm by showing several advantages including being able to produce relatively low errors, good performance in classification, being able to overcome large amounts of training data efficiently, and effective methods for estimating missing data.

The main purpose of the RF training stage is to build many uncorrelated DT. To reduce variance associated with classification, an overlapping sampling solution named 'bagging' was adopted in RF (Liu et al., 2021).

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3056

METHOD

1. Data Collection

Data source is retrieved through a public site *https://www.kaggle.com/datasets/uciml/student-alcohol-consumption*, This dataset is from portugal country by P. Cortez and A. Silva. Data publik ini memiliki 33 atribut dengan jumlah record sebanyak 649 responden. Data were obtained in a survey of students of mathematics and Portuguese language courses in high school. It contains a lot of interesting social, gender and study information about students.

2. Research Methods

The cross-industry standards for data mining (CRISP DM) process is a framework for translating business issues into data mining tasks and implementing data mining projects independent of the application area and technology used (Huber et al., 2019).

CRISP-DM with a few general steps and more, tasks for each step: *business understanding, data understanding, data preparation, data modelling, results evaluation and deployment* (Cazacu & Titan, 2020).

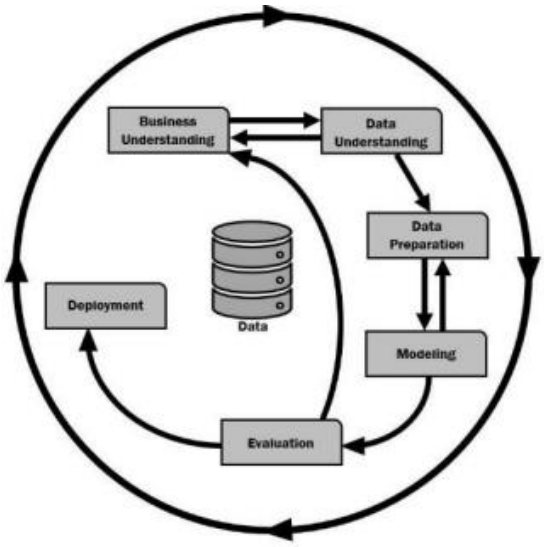

Figure 2. CRISP-DM (Cazacu et al., 2020)

1) Business Understanding

This stage explains the purpose of the research carried out. Exploratory data analysis (EDA) of the considered data set is provided. In particular, EDA is mainly focused on labels, but other insights from the data are also taken into account. The purpose of this study is to train several machine learning models to see the best model in predicting students' Portuguese scores in the final year (G3) of The Portuguese language, by training the Support Vector Regressor (SVR) and Random Forest (RF) model models so that later the best model will be obtained to predict the final value based on alcohol consumption by students.

2) Data Understanding

The following is a prospectus of all features into the dataset, explaining in detail each feature so that it can be understood and make it easier for research to be carried out.

Table 1. Feature description of the dataset

Atribute	***Record***
School	:(binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
Sex	:(binary: 'F' - female or 'M' - male)
Age	:(numeric: from 15 to 22)
Address	:address type (binary: 'U' - urban or 'R' - rural)
Famsize	:(binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	:(binary: 'T' - living together or 'A' - apart)
Medu	:mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Fedu	:father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
Mjob	:mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	:father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Reason	:reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
Guardian	:(nominal: 'mother', 'father' or 'other')
Traveltime	:(numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
Studytime	:(numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
Failures	:(numeric: n if 1<=n<3, else 4)
Schoolsup	:support (binary: yes or no)
Famsup	:family support(binary: yes or no)

Paid	:(Math or Portuguese) (binary: yes or no)
Activities	:extra-curricular(binary: yes or no)
Nursery	:attended nursery school (binary: yes or no)
Higher	:wants to take higher education (binary: yes or no)
Internet	:at home (binary: yes or no)
Romantic	:(binary: yes or no)
Famrel	:family relationships (numeric: from 1 - very bad to 5 - excellent)
Freetime	:(numeric: from 1 - very low to 5 - very high)
Gout	:(numeric: from 1 - very low to 5 - very high)
Dalc	:(numeric: from 1 - very low to 5 - very high)
Walc	:(numeric: from 1 - very low to 5 - very high)
Health	:(numeric: from 1 - very bad to 5 - very good)
Absences	:(numeric: from 0 to 93)
G1	:first period grade (numeric: from 0 to 20)
G2	:second period grade (numeric: from 0 to 20)
G3	:final grade (numeric: from 0 to 20, output/ target

Source: (Saelan et al., 2020)

3) *Data Preparation*

Making preparations to make the data ready to be propses at the modeling stage is important, including: trying to eliminate missing values, changing the names of some column values so that they can make plot visualization cleaner, making plots for categorical and numerical data. Before training some machine learning models to predict the final value (G3), it begins by defining the dataset, namely the x feature and the y label. In the current study, the final value (G3) becomes label data (y) so that the work will refer to a trained model that is later considered good for predicting label values. After several plots of each feature, it was discovered that the average score of portuguese (G3) was influenced by the alcohol consumption of 'Walc' and 'Dalc' learners, the higher level of alcohol consumption was associated with lower learning performance.

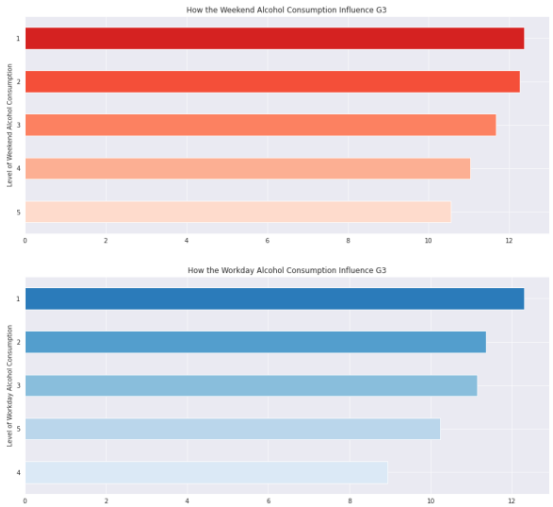

Figure 2. Walc and Dalc plot against G3 (Source: research, 2022)

In addition to the correlation of alcohol consumption with the final value, there are also some of the most influential features and it is considered important to predict the final value (G3).

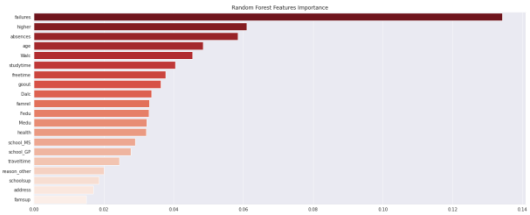

Figure 3. Feature Importance (Source: Research, 2022)

According to the EDA, it can be seen that there are ten most important features for the prediction of the value of G3 are: *'failures', 'higher', 'Absences', 'Age', 'Walc', 'StudyTime', 'freetime', 'goout', 'Dalc', 'famrel'*.

4) *Modeling*

Adjustment of the hyperparameter by using Cross Validation (CV) to obtain the optimal model at the time of training, so that when getting analysis with optimal adjustments can start calculating the predicted value and MAE by training the model that has been formed. The models to be trained to predict label values are the Random Forest and Support Vector Regressor.

5) *Evaluation*

Techno Nusa Mandiri: Journal of Computing and Information Technology
Vol. 19, No. 1 March 2022 | DOI https://doi.org/10.33480/techno.v19i1.3056

Based on the evaluation of the test, the predicted value and MAE of the trained model are obtained. The metrics used to evaluate the trained model are *Mean Absolute Error* (MAE).

6) *Deployment*

If needed, from the model that has been generated, it is tested using new data and re-evaluated for data accuracy. Using generated and represented models or knowledge representation processes.

DISCUSSION

The dataset taken is a collection of data that provides information about student values from the first-year end but in this work will focus on the final year values of the Portuguese language as label data, with several variables that can function as factors that affect the final score which is a reference for learning performance in a group of students who are the target of research.

The dataset was taken from a public site related to alcohol consumption by a group of students. Data is divided into 2 types: categorical data and Numerical data. Before processing by training the model against the data, the data is neated first, if there is empty data, or duplicates so that it will be ready to be processed.

Train multiple models to predict the final class (G3) value. Start by defining the dataset, i.e. features (x) and labels (y). For that it is necessary to preprocess the data. Specifically, convert the categorical data into dummy variables, by using a one-hot encoder. Then divide the dataset into 80% training data and 20% test data from all tests. All models have hyperparameters that must be adjusted. To tune this hyperparameter using cross validation.

Table 2. Description of the model hyperparameter set

MODEL	HYPERPARAMETER	MEANING
SVR	Kernel	the type of kernel to be used in the algorithm
	C	regularization parameters. The power of regularization is inversely proportional to C. It must be absolutely positive
	Epsilon	specifies an epsilon tube in which there is no penalty associated in the training loss function with the predicted points within the distance of epsilon from the actual value.
	Gamma	kernel coefficients for 'rbf', 'poly' and 'sigmoid
RF	n_estimators	number of decision trees used
	max_depth	maximum depth of one decision tree
	max_features	number of random features to consider on each split

Source: (Yang & Shami, 2020)

To form an optimal model it is necessary to make adjustments to the hyperparameter. The best parameters obtained are as follows:

Table 3. Set Hyperparameters

Model	Hyperparameters	Value
SVR	Kernel	Rbf
	Gamma	Scale
	Epsilon	0.0001
	C	2
RF	n_estimator	87
	max_feature	0.2
	max_depth	10

After obtaining the best set of hyperparameters, it can be tested the model and it can be seen how the model performs its best.

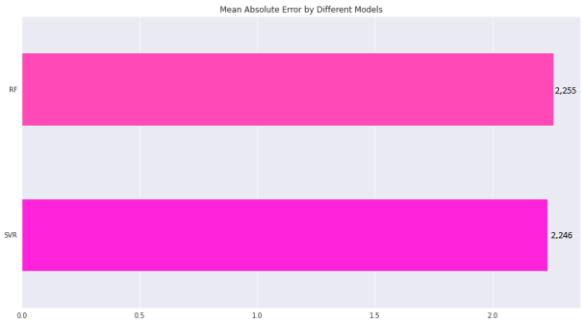

Figure 3. MAE SVR and RF

Source: (research, 2022)

To better understand how this model works, a plot is created between test and prediction data. So that the spread statistics can be known

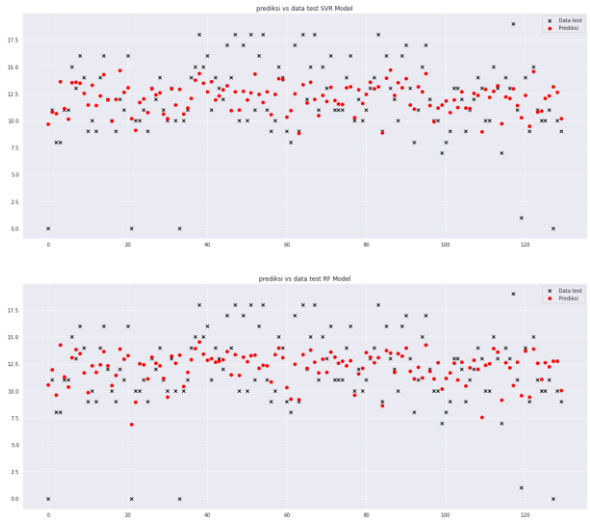

Figure 4. Plot Predictions against Test Data
Source: (research, 2022)

Based on the plot, both models tend to give predictions of G3 close to the average value of the latter. thus, both models perform well enough to predict values close to the average value, but perform poorly to predict values that are far from the latter, that is, the best student scores and the worst student scores. In addition, models tend to overestimate values below the average value and underestimate values above the average value.

Plots containing MAE associated with different models show that the model is RF and SVR works fine. Although the second performs better than the first, these two MAEs are very similar. To determine the best performance between the two models, errors distribution checks are carried out by these models.

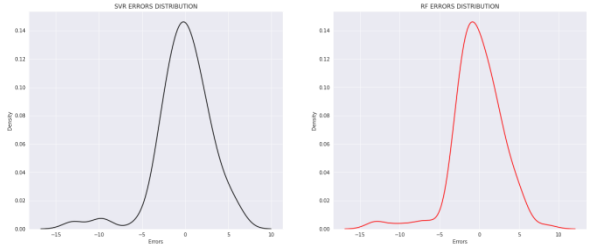

Figure 5. Errors Distribution
Source: (research, 2022)

Errors Distibution SVR is more symmetrical around the main peak and the latter is closer to 0, while the RF fault distribution has a less wide secondary peak. In other words, SVR tends to be more balanced in the ratio of below/exaggerated value with respect to RF, whereas RF performs better than SVR in outliers. Therefore, the choice between SVR and RF depends on what is the main focus: having a more balanced model or a model that performs better on the outlier.

CONCLUSION

At the end of this study, several conclusions were obtained stating that in predicting the final value (G3) of the Portuguese language SVR will be better than RF, looking at the two models trained by SVR and RF, the best model for predicting the final value of G3 is the Support Vector Regressor (SVR) by having an absolute mean error (MAE) of about 2.24 each, while Random Forest has a MAE of 2.25; By analyzing the distribution of errors made by both models, it can be concluded that the SVR is more balanced, that is, it has a better ratio between underestimated and overestimated values; then Based on the provided EDA, the most important feature for G3 predictions is *'failures', 'higher', 'Absences', 'Age', 'Walc', 'StudyTime', 'freetime', 'goout', 'Dalc', 'famrel'*. In reality some variables that were not previously considered the most important clearly affect the final value as shown in the EDA. However, based on several results of the plot analysis of each feature on the final value, Dalc and Walc have the highest influence on the final value (G3).

REFERENCES

Batta, M. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research (IJ*, *9*(1), 381–386. https://doi.org/10.21275/ART20203995

Cazacu, M., & Titan, E. (2020). Adapting CRISP-DM for Social Sciences. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *11*(2sup1), 99–106. https://doi.org/10.18662/brain/11.2sup1/97

Freire, B. R., de Castro, P. A. S. V., & Petroianu, A. (2021). Alcohol consumption by medical students. *Revista Da Associacao Medica Brasileira*, *66*(7), 943–947. https://doi.org/10.1590/1806-9282.66.7.943

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model. *Procedia CIRP*, *79*, 403–408. https://doi.org/10.1016/j.procir.2019.02.106

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors (Switzerland)*, *18*(8), 1–29. https://doi.org/10.3390/s18082674

Liu, K., Hu, X., Zhou, H., Tong, L., Widanage, W. D., & Marco, J. (2021). Feature Analyses and Modeling of Lithium-Ion Battery

Manufacturing Based on Random Forest Classification. *IEEE/ASME Transactions on Mechatronics*, *26*(6), 2944–2955. https://doi.org/10.1109/TMECH.2020.3049046

Nur'artavia, M. R. (2017). Karakteristik Pelajar Penyalahguna Napza Dan Jenis Napza Yang Digunakan Di Kota Surabaya. *The Indonesian Journal of Public Health*, *12*(1), 27. https://doi.org/10.20473/ijph.v12i1.2017.27-38

Palaniappan, S., A Hameed, N., Mustapha, A., & Samsudin, N. A. (2017). Classification of Alcohol Consumption among Secondary School Students. *JOIV : International Journal on Informatics Visualization*, *1*(4–2), 224. https://doi.org/10.30630/joiv.1.4-2.64

Pisutaporn, A., Chonvirachkul, B., & Sutivong, D. (2018). Relevant factors and classification of student alcohol consumption. *2018 IEEE International Conference on Innovative Research and Development, ICIRD 2018*, *May*, 1–6. https://doi.org/10.1109/ICIRD.2018.8376297

Saelan, M. R. R., Sahputra, D. A., Widiastuti, W., & Gata, W. (2020). Komparasi Algoritma Klasifikasi untuk Prediksi Minat Sekolah Tinggi Pelajar pada Students Alcohol Consumption. *Jurnal Sains Dan Informatika*, *6*(2), 120–129. https://doi.org/10.34128/jsi.v6i2.236

Sagala, N., & Tampubolon, H. (2018). Komparasi Kinerja Algoritma Data Mining pada Dataset Konsumsi Alkohol Siswa. *Khazanah Informatika: Jurnal Ilmu Komputer Dan Informatika*, *4*(2), 98. https://doi.org/10.23917/khif.v4i2.7061

Satyanarayana, A., & Nuckowski, M. (2016). *Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance Ashwin Satyanarayana, Mariusz Nuckowski*. https://academicworks.cuny.edu/ny_pubs

Tang, F., Wu, Y., & Zhou, Y. (2022). Hybridizing Grid Search and Support Vector Regression to Predict the Compressive Strength of Fly Ash Concrete. *Advances in Civil Engineering*, *2022*. https://doi.org/10.1155/2022/3601914

Wijaya, P. A. (2016). Faktor-Faktor Yang Mempengaruhi Tingginya Konsumsi Alkohol Pada Remaja Putra Di Desa Keramas Kecamatan Blahbatuh Kabupaten Gianyar. *Jurnal Dunia Kesehatan*, *5*(2), 15–23. https://media.neliti.com/media/publications/76931-ID-faktor-faktor-yang-mempengaruhi-tingginy.pdf

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Yoga Religia, Agung Nugroho, & Wahyu Hadikristanto. (2021). Klasifikasi Analisis Perbandingan Algoritma Optimasi pada Random Forest untuk Klasifikasi Data Bank Marketing. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(1), 187–192. https://doi.org/10.29207/resti.v5i1.2813