ACSSE
APPLIED COMPUTER SCIENCE AND SOFTWARE ENGINEERING
E-ISSN : 0000000000

Comparison of Classification Algorithm Accuracy Levels in Phishing Detection: Decision Tree, Random Forest, and Support Vector Machine

Dewi Fortuna, Master of Computer Science Program, Amikom University Purwokerto, dewifortuna202411@gmail.com

ABSTRACT

Phishing attacks have become a serious threat in the realm of cybersecurity, with both direct and indirect impacts on individuals and organizations. This study aims to compare the performance of three machine learning classification algorithms, namely Decision Tree, Random Forest, and Support Vector Machine, in detecting phishing cases using relevant datasets. The experimental results showed that Random Forest achieved the highest accuracy (96.675%), followed by Decision Tree (93.57%), while Support Vector Machine showed lower performance (54.724%). Cohen's Kappa also indicated that Random Forest had the highest correlation value. In conclusion, for this dataset, Random Forest and Decision Tree are more effective at detecting phishing attacks compared to Support Vector Machine.

Keywords : *Phishing Detection, Machine Learning Algorithms, Random Forest Accuracy, Decision Tree Accuracy, Support Vector Machine*

1. INTRODUCTION

Phishing attacks are a global threat that leverages social engineering to steal sensitive information. This phenomenon has been rampant with the rapid growth of the internet, attacking various sectors such as online banking, social networking, and healthcare (Pranggono & Arabo, 2020). Despite the existence of detection tools, phishing incidents continue to increase every year, posing a serious threat (Kassim, 2019). The success of these attacks is highly dependent on the human factor, reinforcing the need for effective anti-phishing training and decision support systems (Sumner et al., 2021). With the impact evenly distributed on individuals, organizations, and economies, the importance of tackling phishing attacks is increasingly evident globally.

New approaches in phishing detection, such as anomaly-based methods, have been

developed to block suspicious activity on websites (Fariadi & Islami, 2022). E-commerce also prioritizes phishing threat analysis to protect online transactions (Ramadhan & Nurnawati, 2022). In addition, the utilization of advanced threat modeling methods such as stide and dread in e-health security promises a better response to phishing attacks (Faridi et al., 2021). This approach offers an adaptive solution to address the ever-evolving phishing attacks, highlighting the importance of a proactive response in mitigating risk. Based on a previous study entitled "Comparison of Support Vector Machine and Modified Balanced Random Forest in the Detection of Diabetic Patients". Diabetes, or diabetes, occurs when blood sugar levels are high due to disorders in the pancreas and insulin. According to data from the Indonesian Ministry of Health, diabetes is the number 3 cause of death in Indonesia with a percentage of 6.7%. To address this high mortality rate, research is conducted for early detection. This study uses Machine Learning to classify data of diabetic patients. The focus is on the comparison between the Support Vector Machine (SVM) and the Modified Balanced Random Forest (MBRF) as both have been shown to have high accuracy in previous studies. The preprocessing stage is applied to prepare the data so that it can be classified.

All stages of preprocessing are applied to both methods to ensure the datasets used are the same. The evaluation was carried out using the Confusion Matrix. The results of the experiment showed a maximum performance of 87.94% with SVM and 97.8% with MBRF (Ye et al., 2021). This study aims to evaluate and compare the accuracy of three machine learning algorithms (decision tree, random forest, and support vector machine) in detecting phishing attacks. It is hoped that this research can provide in-depth insights into the advantages and disadvantages of each method, as well as understand how these algorithms can overcome specific challenges in detecting phishing attacks.

2. METHODOLOGY

Decision Trees are suitable for phishing detection because they provide easy interpretation and do not require data normalization. Random Forest improves the stability of the model with ensemble learning, effectively overcoming overfitting on complex data. Meanwhile, SVM is effective in high dimensions and tolerant of overfitting, providing good performance. In this study, variables such as URL length and HTTPS presence are used as features in model training and testing. The selection of the algorithm should consider the characteristics of the dataset and the

purpose of the study, where Decision Tree provides interpretability, Random Forest improves stability, and Support Vector Machine performs well in high dimensions, creating a comprehensive approach to phishing detection.

2.1 Problem Identification

In this step, problems are identified and analyzed against previous studies, as well as techniques applied to identify phishing websites.

2.2 Data Collection

At the data collection stage, this study utilizes secondary data obtained from the kaggle database. In dataset_phishing consisted of 11430 records with 89 attributes (88 attributes and 1 target attribute). The target attribute has ligmate and phishing. These attributes will be grouped into 2 groups, namely 80% for training data and 20% for testing test data. The following is the number of datasets by class, which can be seen in table 1:

It	Status	Number of Record datasets
1	legitimate	5715
2	phishing	5715
Sum		11430

2.3 Pre-Processing Stage

Pre-processing is a step aimed at selecting data by overcoming missing values, resulting in clean and ready to use data. In this study, missing values are used.

2.4 Use of Decision Tree, Random Forest, and Support Vector Machine Algorithms

After the Pre-Processing Stage, there was the use of Decision Tree, Random Forest, and Support Vector Machine algorithms. In several metrics that can be used to evaluate classification models, including accuracy.

2.5 Drawing conclusions

The conclusion drawing process aims to summarize the research results obtained through the application of Decision Tree, Random Forest, and Support Vector Machine, which results in the best level of accuracy in identifying and classifying phishing websites based on accuracy. In the

context of the classification level in data mining for this algorithm, the assessment scale is applied as follows:

- Excellent classification: 90 - 100
- Good classification: 80 – 89
- Enough classification: 70 - 79
- Low classification: 60 - 69

The following is the framework of the methods used in this study, namely:

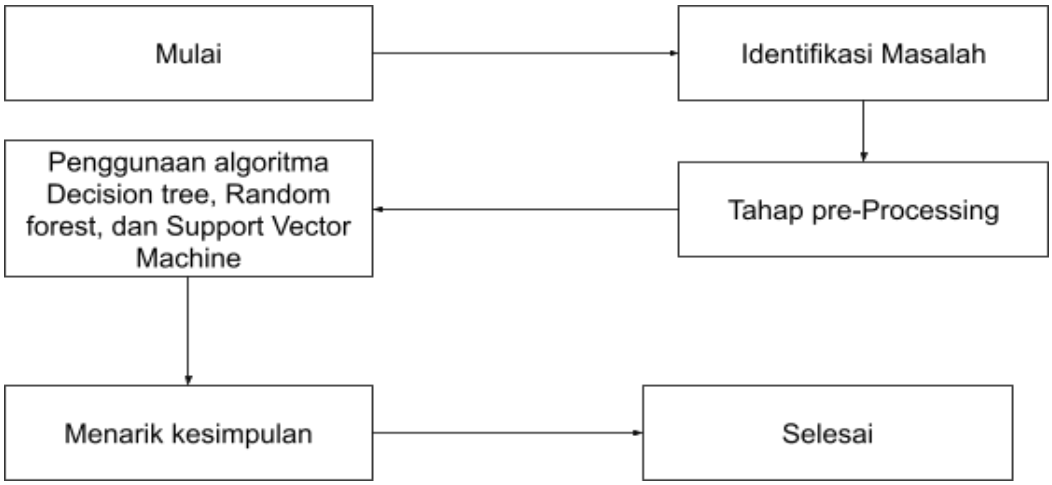
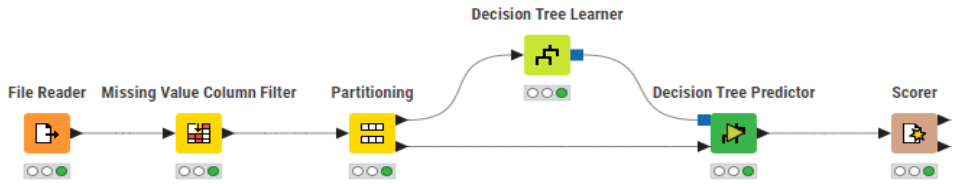


Figure 3.1 Method framework

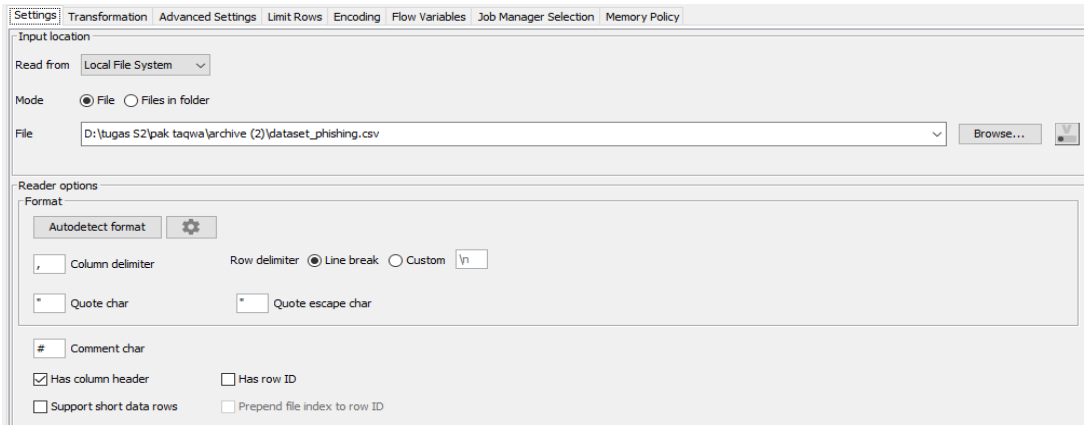
3. RESULTS AND DISCUSSION

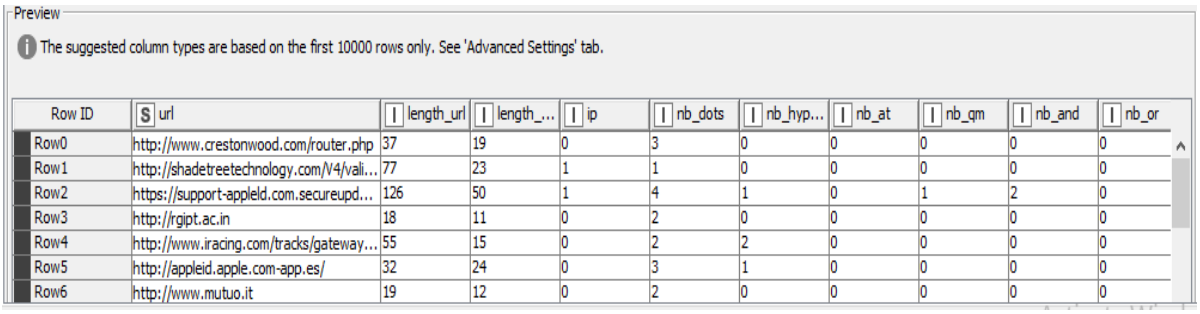
The dataset was obtained from a phishing web survey. Consisting of 89 Columns, the tested attribute is the status attribute.

3.1 Decision Tree



In the image above is the workflow in the training application calculating the decision tree.

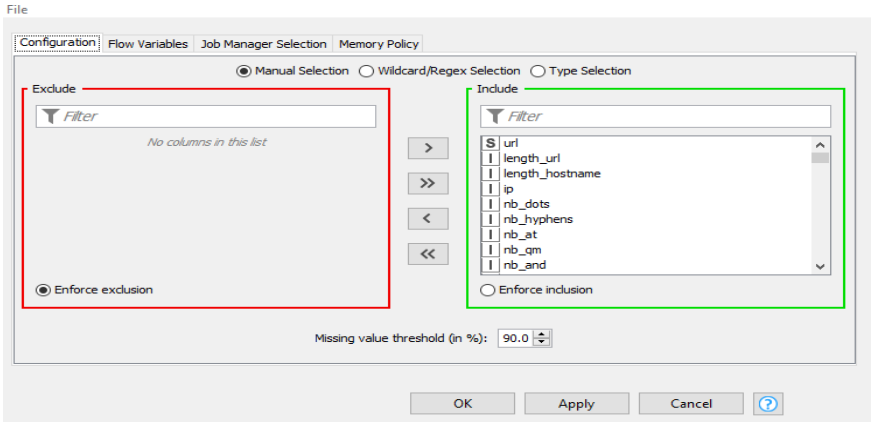




Row ID	url	length_url	length_...	ip	nb_dots	nb_hyp...	nb_at	nb_qm	nb_and	nb_or
Row0	http://www.crestonwood.com/router.php	37	19	0	3	0	0	0	0	0
Row1	http://shadetreetechnology.com/V4/vali...	77	23	1	1	0	0	0	0	0
Row2	https://support-appleld.com.secureupd...	126	50	1	4	1	0	1	2	0
Row3	http://rgipt.ac.in	18	11	0	2	0	0	0	0	0
Row4	http://www.iracing.com/tracks/gateway...	55	15	0	2	2	0	0	0	0
Row5	http://appleid.apple.com-app.es/	32	24	0	3	1	0	0	0	0
Row6	http://www.mutuo.it	19	12	0	2	0	0	0	0	0

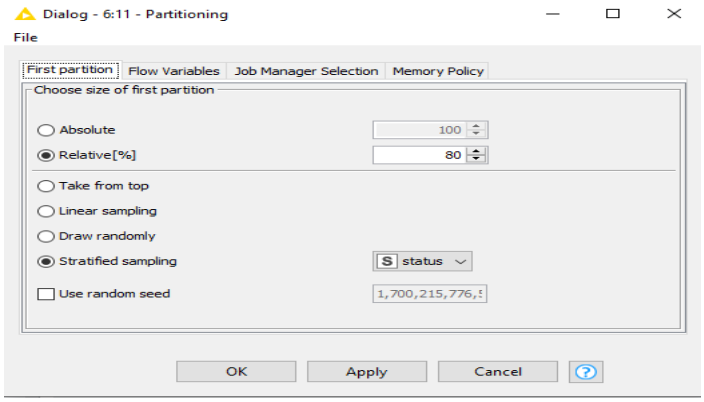
In File Reader, select a dataset to read, the way is by clicking

Browse



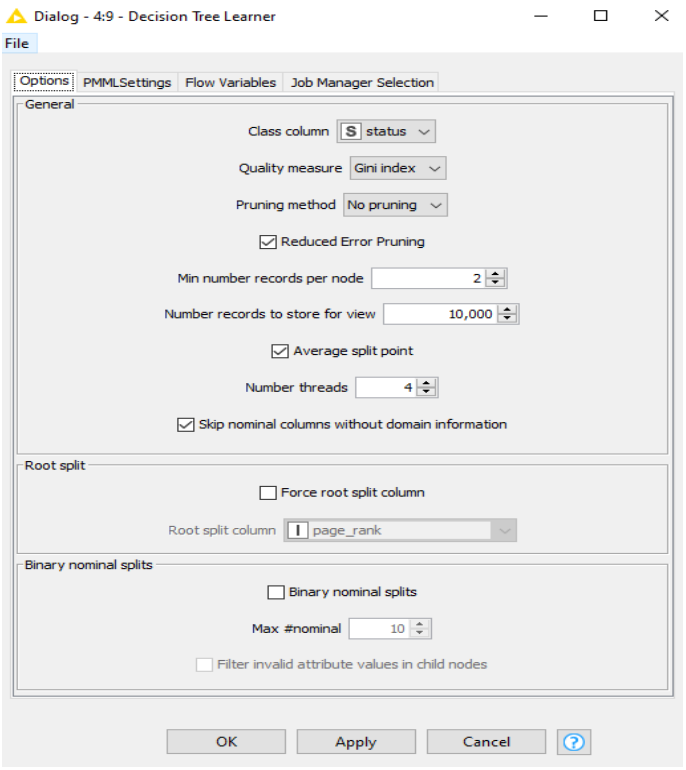
In the picture is the missing value, here is the explanation:

- Configuration Flow Variables are the names of files, columns, or other data used for system configuration or data projects.
- Missing value threshold (in %): 90.0, the percentage level of the limit that indicates the amount of missing data that can be included in the processing. In this case, the limit percentage rate is 90%.

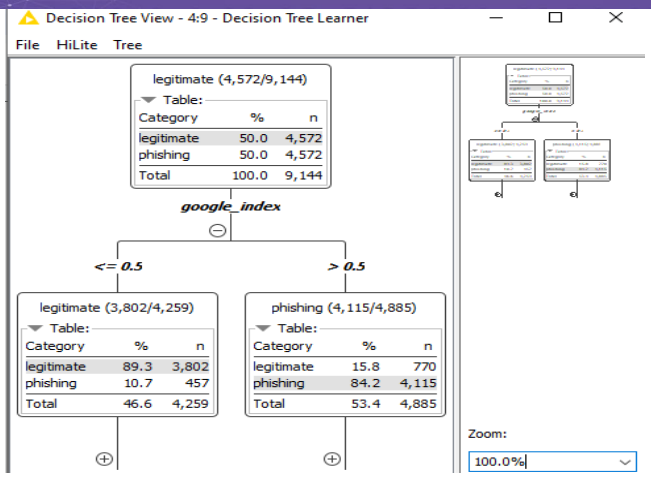


ACSSE
APPLIED COMPUTER SCIENCE AND SOFTWARE ENGINEERING
E-ISSN : 0000000000

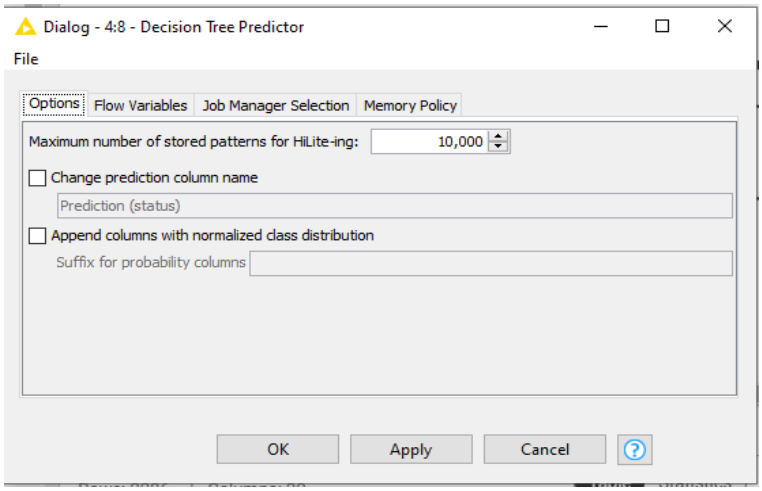
In Relative Partitioning, it is filled with 80 to divide the data into two groups where 80 data will go into one group, and 20 data will go into the other group. For stratified sampling, it is filled with status.



Fill in the Decission Tree Learner as shown in the image, here is the explanation: Decision Tree Learner is one of the nodes or tools used to create a decision tree model. This node enables the analysis and training of decision tree-based models within the KNIME Analytics platform. By using Decision Tree Learner in KNIME, it is possible to solve many problems such as classification and prediction based on the selected features. These nodes make it possible to customize settings such as separation criteria, maximum depth of trees, and many other options according to analysis needs.

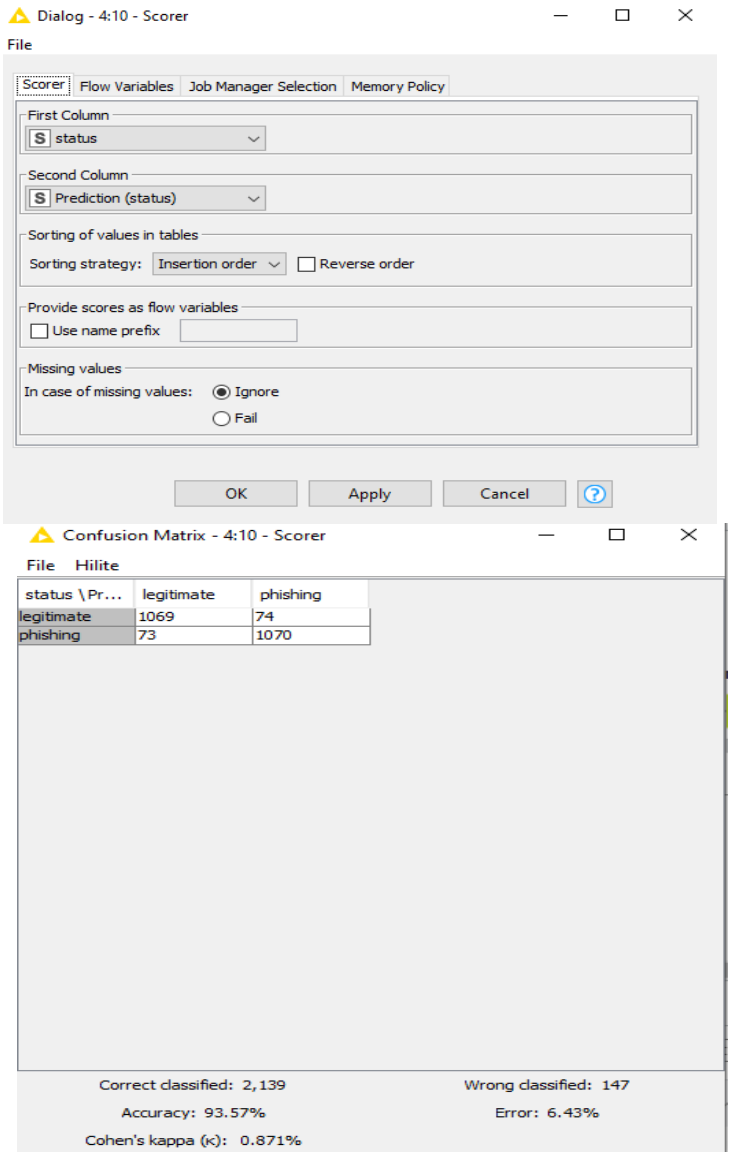


In the figure is a picture of Decision Tree Learner is one of the machine learning algorithms that is often used to calculate the probability of input variables and evaluate how well a model is against training data and testing data. In this case, the Decision Tree Learner algorithm is used to predict whether the received email is phishing or not. In its presentation, the final result is calculated through the amount of training and testing data that is below or above the previously set level of uncertainty (threshold). Based on the data that has been processed, the Decision Tree Learner algorithm calculates the probabilities for each category at each level of the tree hierarchy. For example, for the legitimate category, the amount of data that is below 0.5 indicates reliable certainty in the prediction. Overall, a tree is needed to predict whether the email is phishing or not. To achieve this, the tree is built by selecting the variable that has the highest probability of recalculating the amount of data on each node.



ACSSE
APPLIED COMPUTER SCIENCE AND SOFTWARE ENGINEERING
E-ISSN : 0000000000

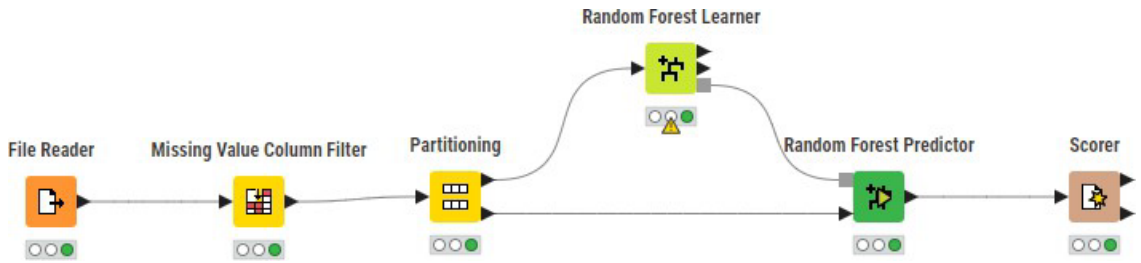
Maximum number of stored patterns for HighLighting -> Sets the number of patterns to be stored for hilite-ing. The value entered in this column will affect the amount of data stored for the hiLite-ing process. Here it is filled with 10,000. In the image Scorer is configured first, the first column is filled with "status" and for *the second column* is filled with "*prediction* ", then in *case of missing values* select the igrone.



The scorer shows how the score can differentiate between the two classes (in this case, phishing and legitimate) in several different categories. In this matrix, two numbers are displayed for each row and column, the number of examples that have been given to each class and the forecast results that have been generated by the model.

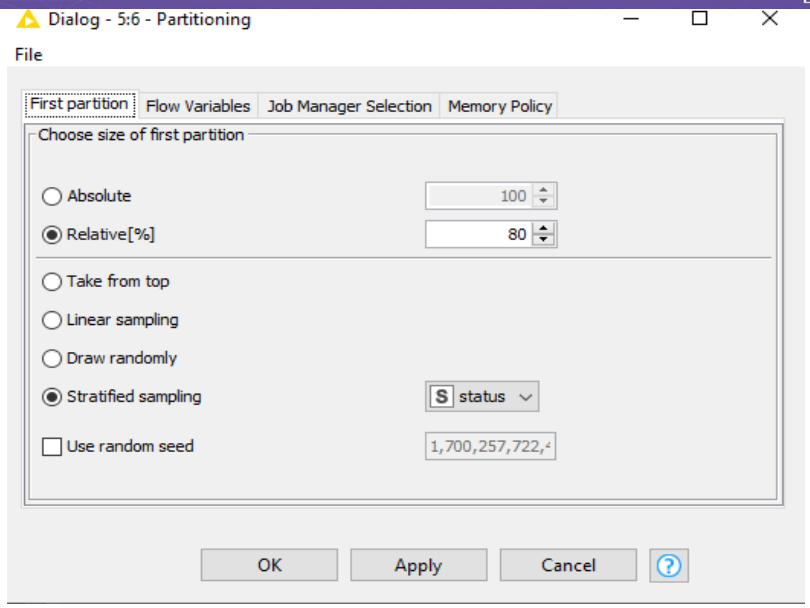
- In the legitimate class and the forecasting results show them as legitimate. This number (1069) indicates the number of instances that the model considers correct.
- On legitimate classes and forecasting results that show them as phishing. This number (74) indicates the number of correct instances that the model considers wrong.
- On the phishing class and the forecast results show them as legitimate. This number (73) indicates the number of instances that the model incorrectly considers correct.
- On the phishing class and the forecast results show them as phishing. This number (1070) indicates the number of instances that are considered wrong by the model.
  Correctly classified: 2,210 Accuracy: 93.57%
  Cohen's kappa (K): 0.871%
  Wrong classified: 147 Error: 6.43%

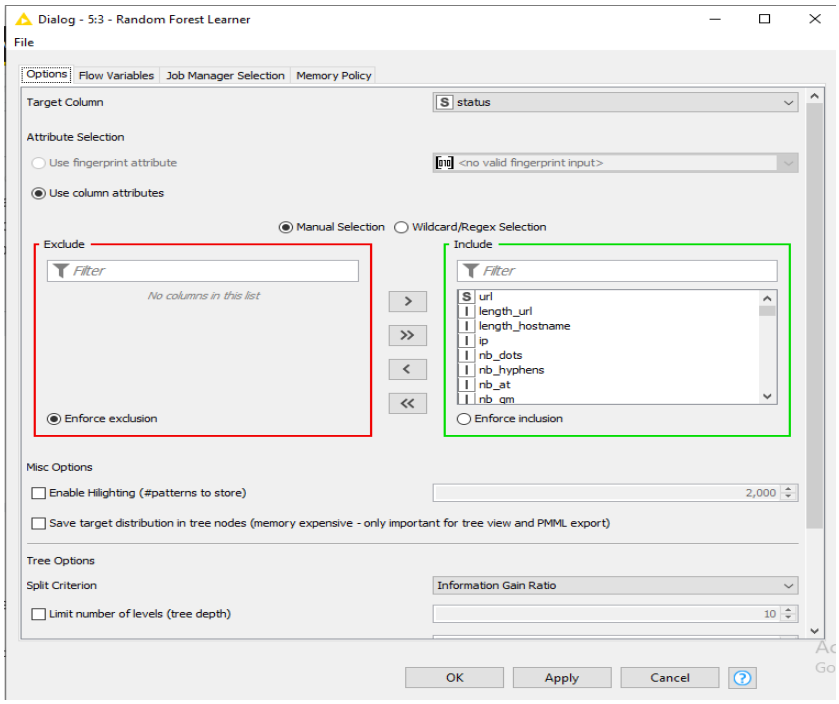
3.2 Random Forest



Like the first decision tree is File Reader, then select the missing value column filter then select partitioning. In Relative Partitioning, it is filled with 80 to divide the data into two groups where 80 data will go into one group, and 20 data will go into the other group. For stratified sampling, it is filled with status.

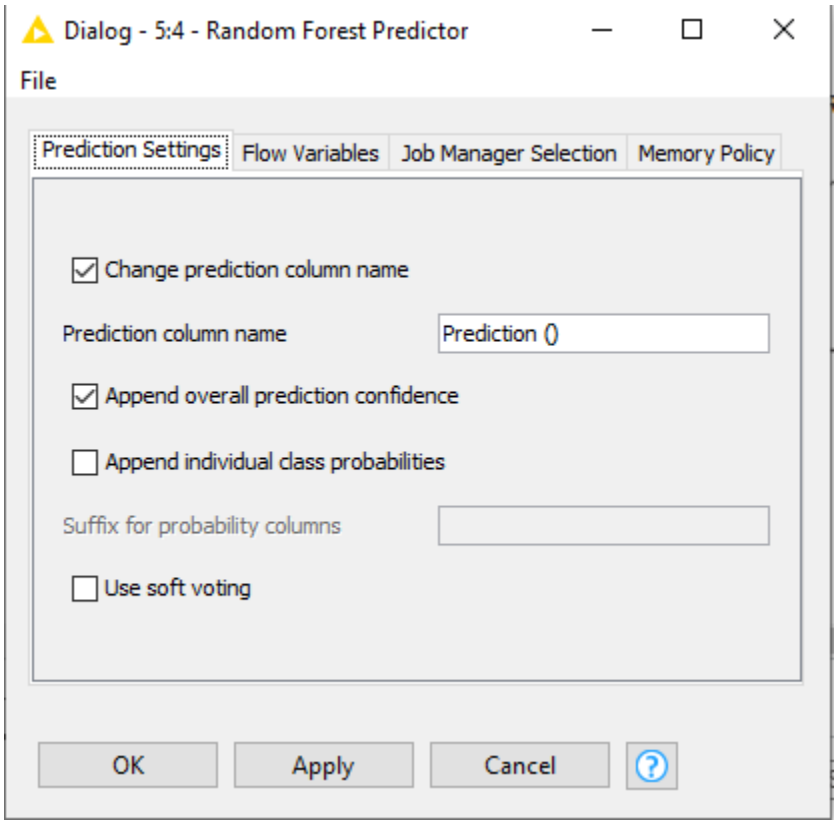
ACSSE
APPLIED COMPUTER SCIENCE AND SOFTWARE ENGINEERING
E-ISSN : 0000000000



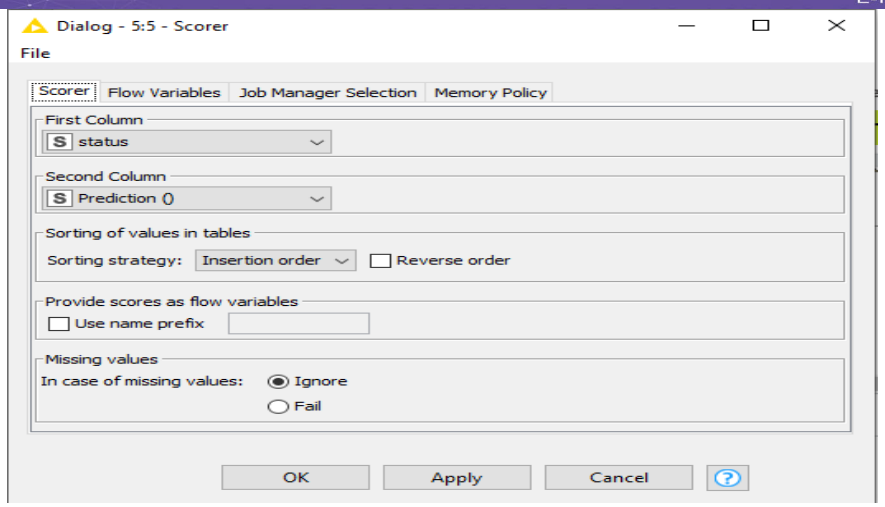


Here is an explanation related to Random Forest Learner:

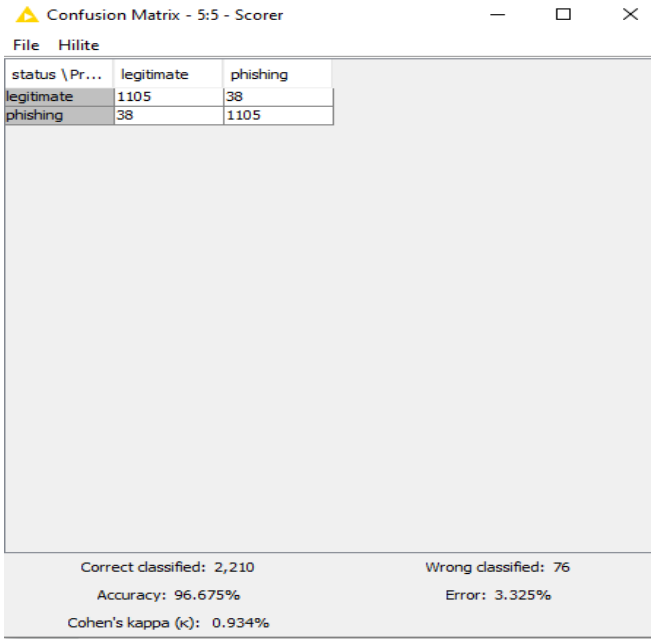
1) Target column is the state
2) Attribute Selection, "Use column attributes": Uses the attributes present in each data column as a representation of the data input .



To help users understand the forecast results, several options are available in the "Prediction Settings" area. Users can choose to rename prediction columns, add prediction depth (prediction certainty level), or add individual class probabilities. In addition, users can choose to use the soft voting method, which allows for forecasts that occur when classes compete.



In the scorer, the configuration first, namely in the first column is filled with "status" and for the second column is filled with "prediction ()", then in case of missing values select the igrone.



The scorer shows how the score can differentiate between the two classes (in this case, phishing and legitimate) in several different categories. In this matrix, two numbers are displayed for each row and column, denoting the number of examples that have been given to each class and the forecast results that have been generated by the model.

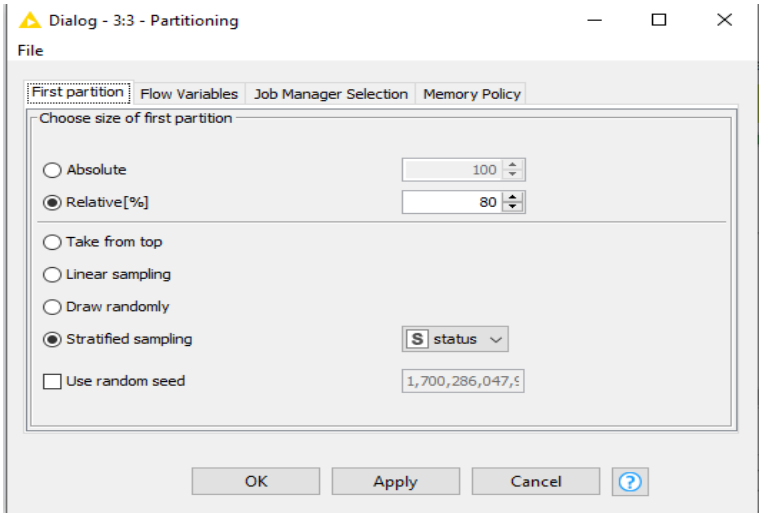
- In the legitimate class and the forecasting results show them as legitimate. This number (1105) indicates the number of instances that the model considers correct.
- On legitimate classes and forecasting results that show them as phishing. This number (38) indicates the number of correct instances that the model considers wrong.

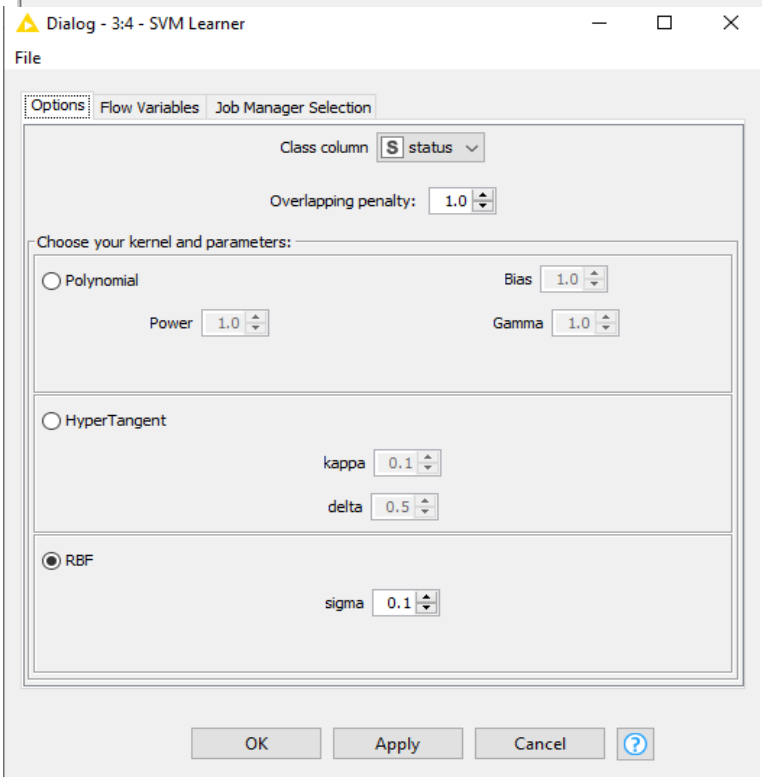
- On the phishing class and the forecast results show them as legitimate. This number (38) indicates the number of instances that the model incorrectly considers correct.
- On the phishing class and the forecast results show them as phishing. This number (1105) indicates the number of instances that are considered wrong by the model.
  Correctly classified: 2,210 Accuracy: 96.675%
  Cohen's kappa (K): 0.934%
  Wrong classified: 76 Error: 3.325%

3.3 Support Vector Machine

Like the decision tree and random forest, the first is File Reader, then select the missing value column filter and then select partitioning. In Relative Partitioning, it is filled with 80 to divide the data into two groups where 80 data will go into one group, and 20 data will go into the other group. For stratified sampling, it is filled with status.

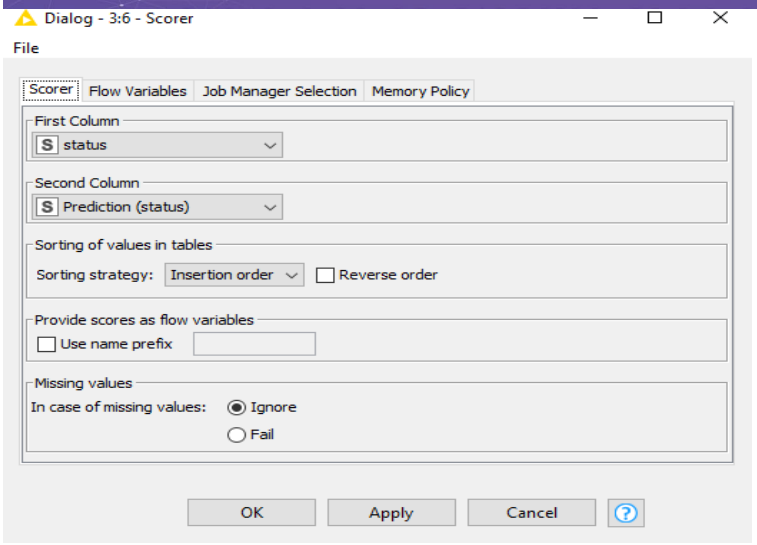
ACSSE
APPLIED COMPUTER SCIENCE AND SOFTWARE ENGINEERING
E-ISSN : 0000000000



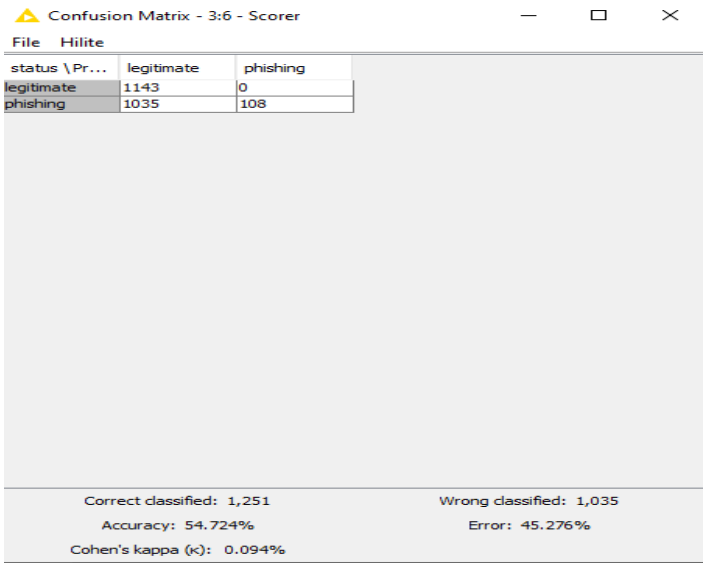


In SVM Learner, use a column class, namely status, then an overlapping penalty, which is 1.0. RBF: This setting uses the Gaussian Radial Basis Function (RBF) kernel which contains sigma 0.1.

ACSSE
APPLIED COMPUTER SCIENCE AND SOFTWARE ENGINEERING
E-ISSN : 0000000000



In the scorer, the configuration first, namely in the first column is filled with "status" and for the second column is filled with "prediction ()", then in case of missing values select the igrone.



The scorer shows how the score can differentiate between the two classes (in this case, phishing and legitimate) in several different categories. In this matrix, two numbers are displayed for each row and column, denoting the number of examples that have been given to each class and the forecast results that have been generated by the model.

- In the legitimate class and the forecasting results show them as legitimate. This number (1143) indicates the number of instances that the model considers to be correct.
- On legitimate classes and forecasting results that show them as phishing. This number (0) indicates the number of correct instances that the model considers incorrect.

- On the phishing class and the forecast results show them as legitimate. This number (1035) indicates the number of instances that the model incorrectly considers true.
- On the phishing class and the forecast results show them as phishing. This number (108) indicates the number of instances that are considered incorrect by the model.

Correctly classified: 1.251 Accuracy: 54.724%
Cohen's kappa (K): 0.094%
Wrong classified: 1,035 Error: 45,276%

4. CONCLUSION AND RECOMMENDATION

Of the 3 algorithms that have been implemented, Random Forest, and Support Vector Machine provide the highest accuracy results (96.675%), followed by Decision Tree (93.57%), while Support Vector Machine shows lower performance (54.724%). Cohen's Kappa, which provides a correlation value between the observation results and the prediction results adjusted for the likelihood of success occurring by chance, also shows that Random Forest has the highest value, followed by the Decision Tree and the Support Vector Machine. In terms of the number of errors, Decision Tree and Random Forest produce better results than Support Vector Machine. In conclusion, for this dataset, Random Forest and Decision Tree have better performance than the Support Vector Machine in terms of accuracy and correlation between prediction and observation results.

REFERENCES

Agustia, N. C. (2023). Systematic Literature Review: The Utilization of e-Learning in Indonesia Universities. *International Journal of Learning and Education*, *1*(1), Article 1.

Castaño, F., Fernañdez, E. F., Alaiz-Rodríguez, R., & Alegre, E. (2023). PhiKitA: Phishing Kit Attacks Dataset for Phishing Websites Identification. *IEEE Access*, *11*, 40779–40789. https://doi.org/10.1109/ACCESS.2023.3268027

Cesarelli, G., Donisi, L., Amato, F., Romano, M., Cesarelli, M., D'Addio, G., Ponsiglione, A. M., & Ricciardi, C. (2023). Using features extracted from upper limb reaching tasks to detect Parkinson's disease by means of machine learning models. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 1056–1063. https://doi.org/10.1109/TNSRE.2023.3236834

Deng, W., Guo, Y., Liu, J., Li, Y., Liu, D., & Zhu, L. (2019). A missing power data filling method based on improved random forest algorithm. *Chinese Journal of Electrical* Engineering,5 (4), 33–39 https://doi.org/10.23919/CJEE.2019.000025

Fillbrunn, A., Dietz, C., Pfeuffer, J., Rahn, R., Landrum, G. A., & Berthold, M. R. (2017). KNIME for reproducible cross-domain analysis of life science data. *Journal of Biotechnology,261*, 149–156. https://doi.org/10.1016/j.jbiotec.2017.07.028

Hsu, C.-H. (2021). Optimal Decision Tree for Cycle Time Prediction and Allowance Determination. *IEEE Access*, *9*, 41334–41343. https://doi.org/10.1109/ACCESS.2021.3065391

Li, C., Zhou, J., Du, K., & Dias, D. (2023). Stability prediction of hard rock pillar using support vector machine optimized by three metaheuristic algorithms. *International Journal of Mining Science and Technology*, *33*(8), 1019–1036. https://doi.org/10.1016/j.ijmst.2023.06.001

Li, W., Manickam, S., Laghari, S. U. A., & Chong, Y.-W. (2023). Uncovering the Cloak: A Systematic Review of Techniques Used to Conceal Phishing Websites. *IEEE Access*, *11,71925* –71939. https://doi.org/10.1109/ACCESS.2023.3293063

Ye, J., Yang, J., Yu, J., Tan, S., Luo, F., Yuan, Z., & Chen, Y. (2021). A chi-MIC based adaptive multi-branch decision tree. *IEEE Access*, *9*, 78962–78972. https://doi.org/10.1109/ACCESS.2021.3077125

Zieni, R., Massari, L., & Calzarossa, M. C. (2023). Phishing or Not Phishing? A Survey on the Detection of Phishing Websites. *IEEE Access*, *11*, 18499–18519. https://doi.org/10.1109/ACCESS.2023.3247135