Classification modeling of support vector machine (SVM) and random forest in predicting pharmacodynamics interactions

N A Farhana1*, F M Afendi1, A Fitrianto1 and S H Wijaya2

1Department of Statistics, IPB University, Bogor 16680 Indonesia
2Department of Computer Science, IPB University, Bogor 16680 Indonesia

*E-mail: nurul_aifarhana95@apps.ipb.ac.id

Abstract. Drug-drug interaction (DDI) is a drug’s effectiveness that can affect the body’s response to the treatment process. DDI occurs when food, drinks, chemicals, and other drugs change the effectiveness of a drug that is given simultaneously. One type of DDI is pharmacodynamics interactions. This interaction is difficult to detect and is very dangerous to humans. Therefore it is necessary to do classification modeling to identify pharmacodynamics interactions based on the value of Side Effect Similarity (SES), Chemical Similarity (CS), and Target Protein Connectedness (TPC). The Support Vector Machine (SVM) and random forest classification method that can be used to predict pharmacodynamics interactions. This study aims to find the best classifications technique by first applying the scaling process, variables interaction, resampling technique, and binarization technique. Best on the analysis result obtained by the random forest is the best model with the highest accuracy and AUC value to other models. The accuracy and AUC values for the best models are 89.93% and 79.96%.

1. Introduction

Drug-drug interaction (DDI) is a factor that can affect the body's response to treatment. DDI is considered clinically important if it increases the toxicity and/or reduces the effectiveness of the interacting drug especially when it comes to drugs with narrow safety limits (narrow therapeutic index), such as cardiac glycosides, anticoagulants, and cytostatic drugs. Medicines can interact with food, drink, chemicals, or other drugs. DDI occurs when the food, drink, chemicals, and other drugs change the effect of a drug that is given simultaneously or almost simultaneously. Several drugs are often given concurrently at prescription, so the first drug can counteract, strengthen or weaken, prolong, or shorten the action of the second drug.

Based on the mechanism, DDI can be divided into three major group interactions, namely pharmaceutical (FS), pharmacokinetics (FK), and pharmacodynamics (FD). FS interactions occur due to changes or chemical and physical reactions between 2 or more drugs that can be recognized or seen and occur outside the body and can result in the loss of pharmacological activity of the drug. FK interaction is a pharmacological aspect that includes the fate of drugs in the body, namely their absorption, distribution, metabolism, and excretion to arrive at the workplace and have an effect. FD interactions occur when two or more drugs acting on the same receptor system, workplace, or physiological system result in additive effects, synergistic effects (mutually reinforcing), and antagonistic effects (mutually canceling out).
In a previous study used two different computational approaches in predicting DDI [7]. The first approach is based on the similarity of drug information and the second is based on the knowledge base. However, both approaches have limitations, such as the need to differentiate between drug classes and the inability to handle new drugs with limited information. A similar study was conducted by [9] who measured the strength of the tissue connection between drug targets or Target Protein Connectedness (TPC) and measured the relationship between drugs using side effect similarity (SES) by applying the Bayes approach to predicting DDI.

Currently, research focuses on FK interactions, experiments, and simulation approaches carried out to test for metabolic and transporter-based drug interactions [9]. However, a large amount of DDI cannot be explained in the FS and FK which should be a potential FD. Research has been conducted by [13] namely classifying drug pairs that interact with FD by applying the DP-Claus algorithm to minimize errors in determining non-interacting drug pairs. This grouping is done by measuring the distance of the relationship between drugs by looking at the similarity of drug side effects or Side Effect Similarity (SES), measuring the distance of structural similarity of drug compounds or Chemical Similarity (CS), and measuring the strength of the network connections between drug-protein targets or Target Protein Connectedness (TPC).

One method that can be used to analyze the relationship between FD interaction classes is classification modeling. The purpose of classification is to predict the category or class of an object. For example, in this study, a classification model was used to identify the interactions of FD with additive, synergistic, and antagonistic pharmacological effects. The additive effect has similarities with the synergistic effect because in practice it is very difficult to know the extent to which the effectiveness of a drug is increased i.e. whether the effect is greater or less than the sum of the effects of an individual drug [3]. Meanwhile, the antagonistic effect of one drug interferes with the effect of another which causes neutralization or a decrease in the effect of the other drug.

One method that is widely used in classification modeling is the Support Vector Machine (SVM) and random forest. The SVM method can classify data whose classes can be separated linearly or non-linearly. The SVM algorithm was chosen because in the study of [8] to identify pharmacodynamic interactions with the results obtained showed that the Area Under of Curve (AUC) value reached 68%. Meanwhile, according to [4] the random forest method is a classification method that forms a collection of classification trees that can produce low errors in predicting an event. In the research of Lingga et al. (2017) show that the random forest classification method has a higher level of accuracy in predicting an event than other methods. SVM and random forest algorithms can be directly used for classification modeling on data with response variables of more than two categories (multi-class). Multi-class classification modeling can be done with a binarization approach (binarization). There are two binary approaches, namely One vs One (OVO) and One vs All (OVA) [2].

Drug pair data consists of 3 explanatory variables, namely the distance of the relationship between drugs by looking at the similarity of side effects between drugs or Side Effect Similarity (SES), the distance of similarity in structure of drug compounds or Chemical Similarity (CS), and the strength of the network connection between drug-protein targets or Protein Connectedness (TPC). The data for this drug pair was obtained from the research of [8, 9]. Based on this, the researchers used the SVM and random forest methods to predict FD interactions between drugs based on the values of SES, CS, and TPC. The results of the classification modeling obtained the model accuracy in predicting class based on the accuracy and AUC values. The determination of the best model is selected based on the accuracy and AUC values.

2. Data and method
3. Data
The data used in this study are secondary data obtained from research by [13]. Drug pair data are taken from Drug Bank (https://www.drugbank.ca) on July 9, 2018, however, only drug data approved by the Food and Drug Administrations (FDA) and target proteins present in the human body will be further analyzed. Collecting data from various databases produces slices which are then used for research. The
data to be analyzed is the result of all data slices, namely 50,000 pairs of drugs. The data set used 39,454 drug pairs included in non-interacting and 10,546 drug pairs that interacted with FD with 1,599 drug pairs an antagonist effect and 8,947 drug pairs a synergistic and/or additive effect. The variables used in the classification are presented in table 1.

| Variables | Information | Type of variable |
|-----------|-------------|------------------|
| $y$       | Drug-drug interactions (not interact, antagonist, synergistic and/or additive) | Category          |
| $x_1$     | Side Effect Similarity (SES) | Numeric |
| $x_2$     | Chemical Similarity (CS) | Numeric |
| $x_3$     | Target Protein Connectedness (TPC) | Numeric |

4. Method

4.1. Side Effect Similarity

Side Effect Similarity (SES) is a value developed by [5] who focused on predicting specific protein targets generally associated with the main mechanism of drug action which could then be used to predict pharmacodynamic interactions. The SES value can be obtained from the calculation:

$$SES = \sum_{i \in X \cap Y} (-\log f_i) \cdot g_i$$

with $X$ and $Y$ are two different drugs, $f_i$ is the frequency of side effects, and $g_i$ is the weighted value of the correlated side effects. The higher the SES value, the more similar the side effects of the drug pair so that the stronger the drug interactions are indicated.

4.1.2. Chemical Similarity (CS)

The Klekota-Roth fingerprint is used in predicting drug interactions by measuring the chemical similarity (CS) distance of the chemical structure. This fingerprint involves converting the chemical structure into a sequence of bits (binary digits, value 0 or 1) of a certain length which can then be easily compared between molecules. One of the similarity functions that can be used to calculate the chemical structure similarity distance between drugs is the Tanimoto similarity function which is written in the following equation:

$$\text{Tanimoto similarity functin} = \frac{a}{a + b + c}$$

where $a$, $b$, and $c$ are the frequency of occurrence of events ($O_1 = 1$ and $O_2 = 1$), ($O_1 = 1$ and $O_2 = 0$), ($O_1 = 0$ and $O_2 = 1$), and ($O_1 = 0$ and $O_2 = 0$) where $O_1$ is the binary vector for drug 1 and $O_2$ is the binary vector for drug 2. The drug targets proteins through an enzyme lock-key process with substrates that are related to the molecular forms of drugs and proteins. Therefore, the closer the molecular geometric structure of a drug is, the stronger the indication for drug-pair interactions.

4.1.3. Target protein connectedness (TPC)

Target Protein Connectedness (TPC) is a value used to predict pharmacodynamic drug interactions. The TPC value is used to see how much the relationship between two drug-protein target systems centered on the Protein-Protein Interaction (PPI) network. TPC values are obtained using an algorithm over the PPI network. This algorithm was developed by [9], first mapping to the PPI network based on the target association of these drugs. Furthermore, because many drugs have therapeutic effects (side effects), the weighting is carried out by weighing the target protein in each drug where the protein to be weighted is the right protein one first step of the target protein in the PPI network. The weighting of the protein uses the Pearson correlation coefficient calculation. To calculate the TPC value, you can use the following formula [9]:

$$TPC = \frac{\sum_{i} w_i \cdot r_i}{\sum_{i} w_i}$$

where $w_i$ is the weight of the protein and $r_i$ is the correlation coefficient of the target protein.
\[ S - \text{score (TPC)} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \]

where \(\bar{x}\) is the mean, \(s\) is the standard deviation, and \(n\) is the number of cross-network expressions of points connecting two drug-centered systems. If the TPC value is high, it shows the tighter the protein connection targeted by the drug partner, thus indicating the stronger the drug pair interaction.

4.1.4. Preprocessing

Data preprocessing of drug pairs was observed to increase the accuracy of the model in predicting pharmacodynamic interactions. The values on the SES, CS, and TPC variables have a very large range, so it is necessary to do a scaling process so that the range on each variable is not too large. The scaling process used is to normalize all predictor variable values.

Drug pair data are class unequal. To deal with unbalanced data, the researchers applied resampling techniques by applying several methods, namely oversampling, undersampling, and SMOTE. The application of the oversampling method is to add artificial observations to the minority class (antagonist and synergistic and / or additive) so that the number of observations from the minority class is equal to the majority class (not interacting). Meanwhile, the application of the undersampling method is to use all observations from the minority class and randomly take observations from the majority class to become a minority class. The application of the SMOTE method is to increase the number of minority class observations to make it equal to the number of observations of the majority class by forming artificial observations based on the k-nearest neighbor concept.

4.1.5. Random forest

Random forest is one of the classification methods developed by the Classification and Regression Tree (CART) method by applying the boostrap aggregating (bagging) and random feature selection methods [4]. The random forest method forms as many as \(k\) trees to form a forest, then the analysis is carried out on the tree clusters by combining the predicted results from the \(k\) trees formed. Suppose the data set consists of \(n\) observations and the number of predictor variables is \(p\), while the stages of random forest development are as follows:

1. Random sampling (bootstrap) is performed by returning size \(n\) from the dataset. Trees formed at the bootstrap stage are formed without pruning.
2. Selection of explanatory products randomly as many as \(m\), with \(m < p\). This stage is carried out during the selection process on the formation of a single tree. In this study, 3 explanatory variables were used for each sorting. This stage is called the random feature selection stage.
3. Steps 1 and 2 are carried out \(k\) times to obtain \(N\) estimation trees. The response to observation is predicted by aggregating the prediction results of the \(N\) trees based on the majority vote.
4.1.6. Support vector machine (SVM)
Support Vector Machine (SVM) is a classification method included in supervised learning classes. SVM is a method that aims to find the best dividing plane (hyperplane) that separates two classes in the input space [10]. The basic principle of SVM is a classification case which can be separated linearly.

The concept of SVM can be explained simply as an effort to find the best hyperplane that functions as a separator of the two classes. The best separator hyperplane is a hyperplane that lies in the middle between two sets of objects from two classes. The hyperplane is the best separator between the two classes which can be found by measuring the hyperplane margin and finding its maximum point. Margin is the distance between the hyperplane and the closest pattern from each class. The pattern that is closest to the hyperplane is called a support vector. SVM separates data using a hyperplane with the largest margin between classes.

Determining the optimal hyperplane for data whose class cannot be separated linearly can be solved by entering kernel functions. The kernel function is a function that maps data to a larger dimensional space in the hope that the data will have a better structure so that it is easier to separate. Commonly used kernel functions based on [1] include the following

1. Linear : \( K(x_i, x_j) = x_i^T x_j \)
2. Polynomial : \( K(x_i, x_j) = (x_i^T x_j + 1)^2 \)
3. Gaussian : \( K(x_i, x_j) = \exp (-\gamma \|x_j - x_j\|^2) \)

4.1.7. Model evaluation
Measurement of the performance of a model needs to be done to find out how well the model is in classifying the data correctly. The confusion matrix is a classification table obtained from the amount of accuracy of the prediction results with actual / training data on each data validation observation [6]. Classification accuracy is the ratio of the number of observations that are correctly classified (predictions according to actual) to the total number of observations tested. The equation for getting the accuracy value is as follows [2]:

\[
\text{Accuracy} = \frac{\text{Predicted Correct}}{\text{Total Observations}}
\]

Figure 1. A general illustration of random forest
\[
Accuracy = \frac{h_{11} + h_{22} + h_{33}}{h}.
\]

where \( h_{ij} \) is the number of observations of the \( j \)-class that predicts in the \( i \)-class.

Area Under the Curve (AUC) is the value of the area under the curve that provides an overview of the overall measurement of the suitability of the model used. The AUC value is a part of the area of a square unit whose value is between 0.5 to 1. The AUC value calculation uses the following equation:

\[
AUC = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} (f(x_i^+) > f(x_i^-))
\]

where \( m \) and \( n \) are the number of positive and negative samples, \( f(x_i^+) \) and \( f(x_i^-) \) are the values a function for positive and negative samples.

According to [11] the greater the AUC value, the better the classification model used. The AUC has also been shown to be a reliable performance measure for class imbalance problems and shows how successful it is at separating positive and negative observations. The closer the AUC value is to one, the higher the accuracy of the model or classification.

5. Result
6. Data exploration

The data provided has been selected and calculated the values of each variable \( x \). In previous studies, the data used were 50,000 drug pairs consisting of pharmacodynamic interaction class (FD) and non-interacting class. However, in this study, the pharmacodynamic interaction class (FD) was divided into two classes, namely the antagonist class and the synergistic and/or additive class. So that there are three classes or three response variables, namely the non-interacting class, the antagonist class, and the synergistic and/or additive class. The data set used consists of three explanatory variables that are built based on the results of certain calculations. Based on these calculations, the values of SES, CS, and TPC were obtained for each drug pair. A summary regarding the values of the 3 explanatory variables in the data set is presented in table 2.

|             | Min  | Q1  | Q2  | Mean | Q3  | Max  |
|-------------|------|-----|-----|------|-----|------|
| SES         | 0.000| 0.000| 0.000| 6.575| 0.000| 652.358|
| CS          | 0.000| 0.100| 0.153| 0.166| 0.217| 1.000|
| TPC         | -24.703| 0.324| 2.770| 2.343| 4.633| 400.904|

Based on table 2, information related to the explanatory variables used in the analysis can be obtained. If the drug pair has a strong similarity in drug structure, the CS value is close to 1. Whereas if the drug pair has a strong similarity of drug side effects, the SES value is greater. The greater the value of the TPC variable, meaning that the drug pairs were closer to the similarity seen based on the target protein. Table 3 provides information related to the relationship between predictor variables.

|    | SES | CS  | TPC |
|----|-----|-----|-----|
| SES| 1   |     |     |
| CS | 0.022| 1   |     |
| TPC| -0.007| -0.029| 1   |
Based on table 3, it can be concluded that the correlation between SES and CS variables is 0.022. The correlation between CS and TPC was -0.029, while the correlation between TPC and SES was -0.007. This shows that the relationship between the predictor variables is very weak. In addition, because the response variables are categorical and the predictor variables are numeric, to see the relationship between variables using cross tabulation analysis (cross tabulation). The initial stage of cross tabulation analysis is to change the numerical predictor variables into categorical predictor variables by applying the entropy-based discretization method. The entropy-based discretization method is a technique used to discretize numerical variables by selecting the value of the variable with the minimum entropy as the dividing point. In this study, cross tabulation analysis was carried out by means of the Chi square test. The results of the chi square test result in a p-value where:

P-value $> 0.05 = H_0$ is accepted
P-value $< 0.05 = H_0$ is rejected

The initial hypothesis used at the cross tabulation calculation stage is that $H_0$ states there is no significant relationship between variables, while $H_1$ states that there is a significant relationship between variables. The results of the Chi square test for each predictor variable and response variable are presented in table 4.

**Table 4. Chi square test**

| Variables  | Chi square Value | P-value | Correlation |
|------------|------------------|---------|-------------|
| SES*DDI    | 2073.5           | 0.000*  | Significant |
| CS*DDI     | 1457.7           | 0.000*  | Significant |
| TPC*DDI    | 60545.5          | 0.000*  | Significant |

*significant on $\alpha = 5\%$

The results of the Chi-squared test are shown in table 4 showing that all predictor variables obtained p-values less than the real level so that it can be concluded that there is a significant relationship between the predictor variables and the response variables. Furthermore, to see the relationship between the predictor variables with the strong, moderate, or weak response variables by looking at the eta square value (eta square). According to [13], the eta squared value is used to measure the level of relationship between two different variables, namely the interval scale variable (numeric) and the nominal scale variable (categorical). The eta squared value is presented in table 5.

**Table 5. The eta coefficient test**

| Predictor variables | Eta squared |
|---------------------|-------------|
| SES                 | 0.0003      |
| CS                  | 0.0002      |
| TPC                 | 0.0005      |

Based on table 5, it is found that the eta quadratic value on each predictor variable is very small, meaning that the relationship between each predictor variable and the response variable is very weak. Furthermore, to see the distribution pattern of each predictor variable for each class using the boxplot.
Based on Figure 2, it can be seen that the distribution for the values of the SES variable, the CS variable, and the TPC variable tend to be the same, namely sticking out to the right. This shows that the mean value of each variable is smaller than the median value. The CS variable and the TPC variable have several outliers. Researchers conducted modeling by ignoring outliers data, but the results obtained by the model without outliers were no better than the model with outliers. This means that outliers add unique patterns to the data so that model performance increases with outlier data. Thus, for extreme conditions (outliers) the model can explain these conditions.

There are 18 classification models formed and to evaluate the best model using accuracy and AUC values. The AUC value determines the best model selection, because the greater the AUC value, the better the classification model used. This also includes an accuracy value. Table 6 is a scenario of the formed classification model.

| Scenarios | Model |
|-----------|-------|
| Model 1   | SVM   |
| Model 2   | Random forest |
| Model 3   | SVM + OVO |
| Model 4   | Random Forest + OVO |
| Model 5   | SVM + OVA |
| Model 6   | Random Forest + OVA |
| Model 7   | OVO + Oversampling + SVM |
| Model 8   | OVO + Oversampling + Random Forest |
| Model 9   | OVO + Undersampling + SVM |
| Model 10  | OVO + Undersampling + Random Forest |
| Model 11  | OVO + SMOTE + SVM |
| Model 12  | OVO + SMOTE + Random Forest |
Scenarios | Model |
---|---|
Model 13 | OVA + Oversampling + SVM |
Model 14 | OVA + Oversampling + Random Forest |
Model 15 | OVA + Undersampling + SVM |
Model 16 | OVA + Undersampling + Random Forest |
Model 17 | OVA + SMOTE + SVM |
Model 18 | OVA + SMOTE + Random Forest |

7. SVM classification model and random forest

The application of the SVM method in classification modeling in this study uses the package "e1017". Meanwhile, the application of random forest uses the package "randomForest". The parameters of the two methods are the default parameters available in the package without setting any parameters. Table 7 shows the evaluation results for the SVM and random forest classification models.

Table 7. Evaluation results of the SVM and random forest classification model.

| Scenarios | Accuracy (%) | AUC (%) |
|---|---|---|
| Model 1 | 78.91 | 50.00 |
| Model 2 | 75.96 | 51.62 |

Based on table 7, it is found that the accuracy value and AUC value for model 1 are 78.91% and 50.00%, respectively. While the accuracy and AUC values for model 2 were 75.96% and 51.62%, respectively. Of the two models, model 2 has an AUC value greater than model 1, so for this classification model 2, the random forest model, is the best model.

8. Binarization technique classification model

The binarization technique is a classification method by breaking multi-class response variables into two class classifications (binary class). The multi-class response variable split into two class classifications was carried out using the One vs One (OVO) and One vs All (OVA) approaches.

The OVO approach is to build training data sets by building a binary classification model for all possible class pairs. If the number of classes is K, then the binary classification will be K (K-1) / 2. While the OVA approach is to build training data clusters by building a binary classification model between each class and all other classes. If there are K classes, then the binary classification model that is built in as many as K classes. Table 8 is the evaluation result of the binarization technique classification model.

Table 8. Evaluation results of the binarization technique classification model

| Scenarios | Accuracy (%) | AUC (%) |
|---|---|---|
| Model 3 | 78.93 | 50.02 |
| Model 4 | 76.29 | 50.39 |
| Model 5 | 78.92 | 50.02 |
| Model 6 | 76.87 | 50.22 |

Based on table 8, it is obtained that the classification model with the highest AUC value is model 4 at 50.39%. This shows that model 4, namely the random forest + OVO model, is the best model for the binarization technique classification model.
9. Model classification of binarization technique and resampling technique

The data on drug pairs used in this study have an unbalanced observation. Therefore we need a resampling technique to handle class imbalance. There are several methods of resampling technique used, namely oversampling, undersampling, and SMOTE. This resampling technique is combined with the binarization technique as well as the SVM and random forest classification methods so that 12 classification models can be formed from this combination. Evaluation of the classification model uses the calculation of the accuracy value and the AUC value. Figure 3 is the evaluation result of the binarization and resampling technique classification model.

![Figure 3. Result of the evaluation of the binarization and resampling technique classification model](image)

Based on Figure 3, it is found that the classification model with the highest AUC value is the OVA + Oversampling + Random Forest model of 79.96%. This shows that this model is the best model for the classification model binarization technique and resampling technique.

10. Determination of the best model

The combination of classification models formed by 18 models has selected the 3 best classification models based on the classification method. Figure 4 is the selected classification model.
Based on figure 4, it can be seen that model 14 has the greatest accuracy and AUC value compared to other models. The accuracy value is 89.93% and the AUC value is 79.96%, this shows that the accuracy of the model is moderate. Model 14 is an OVA + Oversampling + random forest model. This shows that the binarization technique with the OVA approach gives a good performance in classification modeling for multi-class response cases. The oversampling method can also handle class imbalances in this drug pair data, and the random forest classification method is better than the SVM classification method for drug pair data cases with multi-class response variables.

11. Comparison of classification models

Research that has been conducted by Hasnita et al. (2019) was to predict pharmacodynamic interactions on drug pairs data. The data used consisted of 2 response variables, namely non-interaction and pharmacodynamic interactions. The best classification model for predicting pharmacodynamic interactions is the SVM method with an accuracy value of 62.24% and an AUC value of 67.91%. However, data preprocessing must be carried out, namely scaling, resampling techniques, the interaction between variables, and variable discretization.

Based on the research of Hasnita et al. (2019), the pharmacodynamic interaction class is divided into 2 classes, namely antagonistic effects and synergistic and / or additive effects. So that there are 3 response variables used, namely the non-interacting class, the antagonist class, and the synergistic and / or additive class. However, the preprocessing carried out was slightly different from previous studies, namely, the scaling process was carried out by normalizing data, resampling techniques, the interaction between variables, and binarization techniques. In this study, the best model for predicting pharmacodynamic interactions was the random forest method with an accuracy value of 89.93% and an AUC value of 79.96%. This suggests that identifying pharmacodynamic interactions to be more specific provides an increase in the accuracy and AUC values of the model.
12. Conclusion

Based on the results and discussion in this study, it can be concluded that binarization techniques can be applied in classification modeling with multi-class responses. In this case, the One vs All (OVA) approach is better than the One vs One (OVO) approach. To deal with class imbalance you can apply several resampling techniques. The oversampling approach, in this case, is better than the two approaches, namely undersampling and SMOTE. Meanwhile, the random forest classification method is better than the SVM classification method. However, data preprocessing must be done because it can increase the ability of a classification method to predict data.

References

[1] Abe S 2010 Support vector machine for pattern classification 2nd ed New York (US): Springer
[2] Adnan M N and Islam M Z 2015 One-vs-all binarization technique in the context of random forest Computational Intelligence Machine Learning 4(1):22-24
[3] Baxter K 2008 Stockley's Drug Interactions 8th ed London (UK): Pharmaceutical Pr
[4] Breiman L 2001 Random forests Machine Learning 45(1):5-32
[5] Campillos M, Kuhn M, Gavin A, Jensen L J, and Bork P 2008 Drug Target Identification Using Side-Effect Similarity Vol 321 doi: 10.1126/science.1158140
[6] Fawcett T 2006 An introduction to ROC analysis Pattern Recognition Letters 2(1):861-874
[7] Gottlieb A, Stein G Y, Oron Y, Ruppin E, and Sharan R 2012 INDI: a computational framework for inferring drug interactions and their associated recommendations Molecular System Biology 8(1)
[8] Hasnita, Afendi F M, and Fitrianto A, 2019 Perbandingan beberapa metode klasifikasi dalam memprediksi interaksi farmakodinamik Indonesian Journal of Statistics and Its Applications 4(1): 11-21
[9] Huang J, Chaoqun N, Christoper D G, Lun Y, Hongkang M, and Jing-jing D J H 2013 Systematic Prediction of Pharmacodynamic Drug-Drug Interactions through Protein-Protein-Interaction Network PLOS Computational Biology 9(3) doi:10.1371/journal.pcbi.1002998
[10] James A 2007 Sistem Informasi Akuntansi Edisi Ketiga Terjemahan Dewi Fitriasari Jakarta: Salemba Empat
[11] Jian C, Jian G, and Ao Y, 2016 A new sampling method for classifying imbalanced data based on the support vector machine ensemble Neurocomputing 193(6):115-122
[12] Lingga R.D, Fatichah C, and Purwitasari D 2017 Deteksi gempa berdasarkan data twitter menggunakan decision tree random forest dan SVM Jurnal Teknik 6(1):2337-3539
[13] Richardson J T E 2011 Eta squared and partial eta squared as measures of effect size in educational research. Educational Research Review 6(2011): 135-147
[14] Winata H M, Afendi F M, and Fitrianto A 2019 Peningkatan akurasi klasifikasi interaksi farmakodinamik obat berbasis seleksi pasangan obat takberinteraksi Indonesian Journal of Statistics and Its Applications 3(3):247-259