ISSN 2278-3091

Volume 9 No.2, March - April 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse199922020.pdf

https://doi.org/10.30534/ijatcse/2020/199922020

Risk Level Prediction of Life Insurance Applicant using Machine Learning

B. Junedi Hutagaol[1], **Tuga Mauritsius**[2]

[1]Information Systems Management Department, BINUS Graduate Program-Master of Information Systems Management, Bina Nusantara University, Jakarta, Indonesia, b.hutagaol@binus.ac.id

[2]Information Systems Management Department, BINUS Graduate Program-Master of Information Systems Management, Bina Nusantara University, Jakarta, Indonesia,tmauritsus@binus.edu

ABSTRACT

This study aims to examine how machine learning can help life insurance companies like PT. XYZ to determine the level of risk of prospective customers. One of the main objectives of PT. XYZ is to achieve customer satisfaction by issuing insurance policies to prospective customers with a clean data profile according to Underwriting more quickly. The research will be carried out by implementing the Support vector machine (SVM) algorithm using the kernel, Random Forest and Naive Bayes. Determination of the best model will be evaluated using Model Evaluation like confusion matrix, accuracy, precision, and Recall of the developed model. This research will release the design and the impacting of machine learning to current business processes and systems and how to deploy and integrated the model with existing systems at PT. XYZ.

Key words : Machine Learning, Model Evaluation, Naive Bayes, SVM Using Kernel, Random Forest.

1. INTRODUCTION

In general, insurance business is a way of providing protection to the insured for some risk in the future. According to Executive Director of the Indonesian Life Insurance Association (AAJI) Togar Pasaribu, Life Insurance is a business that still has a very broad target market in Indonesia. In 2018 Indonesian population who have a policy is only about 6.6 percent of the total population of Indonesia. This means that there are still 93.4 percent that can be contested by life insurance companies in Indonesia [1].

Indonesia has many companies that compete and are engaged in life insurance. The quantity of this company creates intense competition and forces each company to keep innovating and providing the best service for its customers. If not, the company can be left by the customer to find the desired service provider.

Underwriting is a division of an insurance company that helps assess policy proposals risk and prospective customer data provided by insurance agents [2]. Company need to invest human and time resources to do underwriting. The manual assessment process to get the right policy premium, with the right product and the right risk will take 30-60 days according to insurance company policy[2].

PT. XYZ is one of the largest life insurance companies in Indonesia. One of the company's main missions is customer centricity where every service rendered is seen from the customer's perspective. They do the Underwriting process manually, so it takes a maximum of 40 days after the proposal goes to the core system. it requires more than 50 people to run the business processes.

Spending much time to do underwriting is one of the main problem to PT.XYZ. they can lose opportunity to get the best customers, especially for prospective customers who should not have any adverse risks to PT. XYZ.Most of life insurance companies in Indonesia offers many choices for customers. With the length of the policy proposal evaluation, it will be possible for customers to find life insurance service providers with more effective and efficient services.

In this information technology era, AI (Artificial Intelligent) is growing rapidly. The presence of AI has consistently helped human life, AI has become part of everyday life. AI offers a way to work faster, save time, be more accurate and be able to surpass what humans can do [4].

Machine Learning (ML) is one of the latest technology parts of AI. ML was founded on the basis that machines must be able to learn and adapt through experience. When applied properly, technology can enable organizations to make use of data collection to get business benefits [5].

This Research consist of five chapters. Chapter one explain about introduction, this chapter explain about the problem and the purpose of the research. Chapter two explain about related work, consist of supporting research and theory as references to the research. Chapter three explain about research methodology. Chapter four about discussion and research result. Chapter five will find out the conclusion of the research.

2. RELATED WORK

2.1 Underwriting

According to Lionel Macedo Underwriting is a process to assess the risks that will occur to insurance companies, while the Underwriter is a professional who has the ability to understand these risks [3].

One business unit that must be owned by a life insurance company is underwriting. Underwriting will examine each proposal submitted and then will assess the risk in accordance with the guidelines and rules set by the company[3].

2.2 Artificial Intelligence (AI) and Machine Learning (ML)

According to Arthur Samuel who is famous for his chess game program, ML is defined as a field of study that gives computers the ability to learn without being explicitly programmed[6]. AI can be a new tools for banking industry to analyze customer's data easier in marketing perspective [7].Some researcher using AI for classification of mushroom fungi [8].

Global technology service providers such as Google, Microsoft, Facebook and Bing also use machine learning algorithms to provide the best service in their products. Google provides the best page rank in every search on owned search engines, Facebook with features to recognize photos from users. On the other hand, every spam or ham feature that is owned by the email service is also an implementation of ML [6] .

According to Mohssen Mohammed, et al. as shows in figure 1, Machine learning methods are generally divided into 4 parts: supervised, unsupervised, semi-supervised, and reinforcement learning methods as illustrated [7].

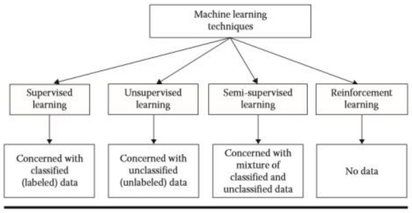
Machine learning techniques
Supervised learning
Unsupervised learning
Semi-supervised learning
Reinforcement learning
Concerned with classified (labeled) data
Concerned with unclassified (unlabeled) data
Concerned with mixture of classified and unclassified data
No data

Figure 1: Categories of Machine Learning Techniques [7].

2.3 Support Vector Machine (SVM) Using Kernel

SVM is one of the first machine learning methods introduced by Vladimir Vapnik and colleagues and can be used for classification types [8].

As described in figure 2, Svm aims to find the best hyperplane that separates two classes (the distance between classes) at the specified input [8].

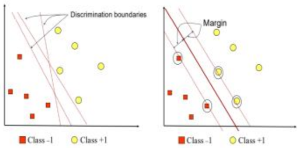

Figure 2: SVM Graph Description[9]

In general, the kernel functions that are often used are Linear, Polynomial and Radial Base Function (RBF) kernels[9].

2.4 Random Forest

The Random Forest Model is an ensemble learning method that constructs a series of decision trees at training time and produces a class that is a class mode (classification) or average prediction (regression) of each tree.

Rabia Emhamed Al, et al. Compare three algorithms for predicting the severity of events that occur in traffic. The main objective of this research is to achieve accuracy and identify the factors behind the Traffic Accident Rate. With this machine learning, it can help predict the level of accidents that will occur from the factors in the field.

Researchers concluded that random forest produces predictable models with 75 percent accuracy better than Logistic Regression (LR) with 74.5 percent accuracy, Naïve Bayesian Classifier (NB) with 73.1 percent accuracy, and AdaBoost algorithms with 74.5 percent accuracy[10].

There are several important parameters that will be used for the experimental development of models as : n_estimators: integer, optional (default = 100) is the number of trees in the forest[11].

2.5 Naive Bayes (NB)

The Naïve Bayes method is used in many real-world applications as a powerful solution for classification and prediction problems. Naïve Bayes finds use in various application domains [12].

Naive Bayes classification is a fast, accurate and reliable algorithm. Naif Bayes classification has high accuracy and speed on large datasets. NB can also be used for high data dimensions as the probability that each variable used is independent [13].

2.5 Model Evaluation

In this research, model evaluation Consist of three methods :

Confussion Matrix. Table 1 explain about confussion matrix value thatcontains the actual classification information and predictions made by the classification system. The following matrix is used for two different classifications[14].

Table 1 : Confusion matrix table [14]

		Predicted	
		Negative	Positive
Actua l	Negative	a	b
	Positive	c	d

a is the number of correct predictions that the actual data is negative or often called True Positive (TN). b is the number of false predictions that a positive instance is often called a false positive (FP). c is the number of incorrect predictions that a negative instance or what is often called a false negative (FN). is the number of true predictions of a positive example or often called true positive (TP).

Precision is one way to measure the level of accuracy between the predicted results provided by the model and the actual data that is already available. Precision will help to calculate the percentage of truly positive data from all positive data that has been predicted by the model built [15].

The formula used to determine the value of precision as shown below [15]:

$$precision = \frac{TP}{TP + FP}$$

Recall (sensitivity) is another technique to determine the best model from several models that have been tested. Recal calculates true positive prediction ratio compared to overall true positive data [15].

The formula used to determine the Recall value is shown below [15]:

$$recall = \frac{TP}{TP + FN}$$

Accuracy is the result of calculating the accuracy of a model in classifying data to be predicted correctly. Accuracy can also illustrate the closeness of predictive values with actual values[15].
Formula :

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

3. METHODOLOGY

Table 3 describes about the research methodology. This research was conducted based on the need for prediction of the level of risk proposals for life insurance prospective customers.

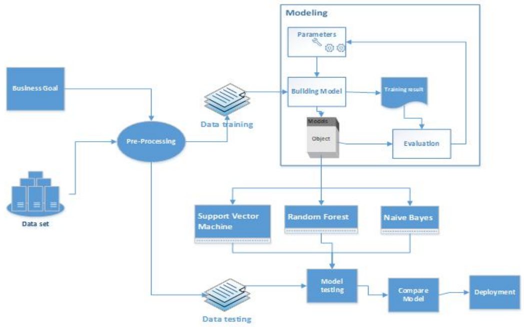

Figure 3 : Research Methodology

3.1 Business Goal Understanding

At this stage an understanding of how the insurance business develops in the world, especially in Indonesia. This business understanding is carried out to find out the parts of the business that can be developed to run the business effectively and efficiently.

One important part in the business insurance process need to be developed is the Underwriting section. So it is important to learn how the Underwriting business process in life insurance works and the parts that allow it to be developed. In specific, understood the underwriting in PT. XYZ is a very important thing.

3.2 Dataset

Dataset taken from the other life insurance company which have the same variables with PT. XYZ to run underwriting. Dataset for training consists of 59.381 records and dataset for testing consists of 19.765 records.

3.3 Pre-Processing

This step have three phases as described in figure 4, data exploration and adjustments will be made as needed. This stage used to avoid missing data, avoiding data errors (noisy data), and avoiding inconsistent data in each dataset.

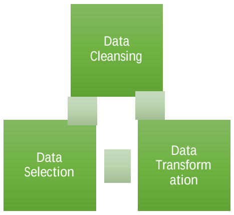

Figure 4 : Pre-processing phase

3.4 Design Experiment

At this stage several steps are carried out repeatedly to build a model.To build the model, the input used is training data and input parameters for the algorithm. The same input will be run against the Support vector Machine (SVM) algorithm using the kernel, Naive Bayes or Random Forest

3.5 Building Model

After the parameters and values have been determined, it will be done building the model using tools that will help to produce the model. The construction of this model is done by using the training data that has been provided. The construction of this model will produce training results and models.

3.6 Model Evaluation

Model testing will be carried out on all models built according to the specified algorithm. Both Support Vector Machine (SVM) models using Kernel, Naive Bayes or Random Forest will be tested with the same dataset. The model built will be validated using accuracy, precision, and recall.

3.7 Deployment

The deployment stage is the process for combining models with systems that run in production so that the resulting model can generate added value to the business. After the best model has been found through testing and comparison of the three selected models, the model to be selected will be used in the Underwriting system.

4. RESULT AND DISCUSSION

4.1 Business Understanding

To avoid risks to the company, PT XYZ identifies and selects risks by processing the personal data of prospective customers provided called the Underwriting process. PT XYZ need The system to give the number 1 for received and 0 for rejected.

4.2 Data Understanding

Variables and their description explained in table 2. Most of the available data is normalized by converting variable values to numeric values. While in plain view, the Product_Info_2 column is a variable that contains data that has not been normalized. For the training data classification algorithm, it must use numeric data, so it needs to be normalized.

Table 2 : Dataset attributes DetailAttribute data type identification

Id	Key ID identification of proposals
Product_Info_1-7	Collection of variables that have been normalized and related to the insurance product chosen by the customer
Ins_Age	Age of prospective customers who have been normalized
Ht	Normalized height of prospective customers
Wt	Normalized weight of prospective customers
BMI	BMI (Body mass Index) of normalized customer candidates
Employment_Info_1-6	Collection of variables that have been normalized and related to the work history of prospective customers
InsuredInfo_1-6	A collection of variables about normalized customer information
Insurance_History_1-9	Collection of variables about the prospective customer's insurance history that has been normalized
Family_Hist_1-5	Collection of variables about the customer's family history of disease that has been normalized

Medical_History_1-41	Collection of variables about the customer's medical history that has been normalized
Medical_Keyword_1-4	Collection of dummy variables about medical keywords and dealing with potential customers who have been normalized
Response	Target variable about the risk level of potential customers (between 1-8)

The data type and quantity are explained in table 3 below:

Table 3 : Attribute data type

Tipe Data	Number of column
Int64	108
Float64	18
Object	1

4.3 Empty Value of Attribute Identification

Found several fields that have empty values for several rows of data as shown below. In the classification algorithm blank data is not allowed and must have a value, so it is necessary to normalize the data.

Found 10 attributes with an empty data capacity of more than 30% of the total available data as in the table 4 below.

Table 2 : List of 30% Empty Data value on dataset

Attributes Name	Empty Value (%)
Medical_History_10	99.06
Medical_History_32	98.14
Medical_History_24	93.60
Medical_History_15	75.10
Family_Hist_5	70.41
Family_Hist_3	57.66
Family_Hist_2	48.26
Insurance_History_5	42.77
Family_Hist_4	32.31

4.4 Variable Response

Variable response is the target of the expected data classification. There are eight response as described in figure 5.They are 1,2,3,4,5,6,7,8 which will be used as a risk level for potential customers. For training data, the distribution of response data that has been provided is as shown below:

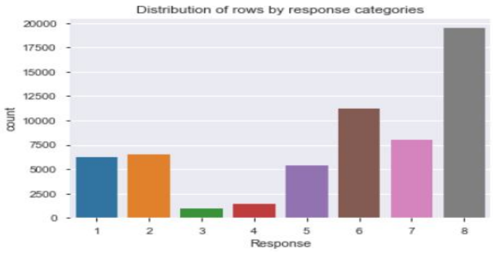

Figure 5 : Variable response Value Distribution

4.5 Pre-processing

Data transformation. Transform Product_Info_2 Attribute to numericProduct_Info_2 attribute has an object data type consisting of letters and numbers so it needs to be transformed variable from object to numerical.

This research will help PT. XYZ to screen prospective customer proposals faster and more accurately so that it will increase company productivity.In this study, the target variable will be changed to binary. Since the majority of the data is with response 8, this research assumes that the proposal is accepted. While the remaining 1-7 are rejected customers. Variable response will be transformed as described in figure 6.

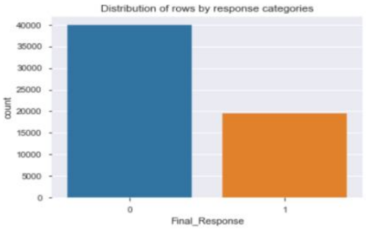

Figure 6 : Variable response Value Distribution in Binary version.

Data Cleansing. In this stage, data cleaning will be done. There are several data cleaning techniques that are performed according to the needs of model development. As described in figure 7 we need to remove outliers on the BMI, H, Wt, Ins_Age attributes of the response. Outliers will be removed as described in figure 8.

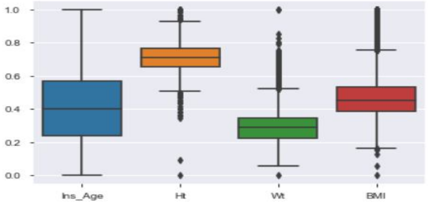

Figure 7 : BMI, H,Wt, Ins_Age Attributes outliers

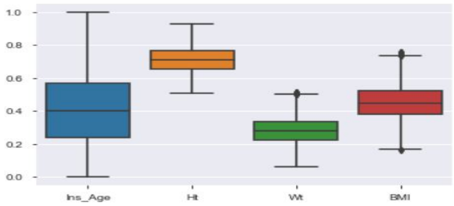

Figure 8 : BMI, H,Wt, Ins_Age Attributes after remove outliers

Feature Selection. To support the development of the model, several types of attribute selection are made according to development needs. Remove attributes that have empty data values greater than 30%.Combining all medical_keyword attributes into Total_Medkwrds. Combining all medical_history attributes Total_MedHis.

4.6 Modeling

In this study, the data will be separated using the 70:30 method. The target variable that will be predicted is binary (1 and 0).This modeling is done with the help of jupyter notebook with notebook server version 6.0.0 and server running with python 3.7.4.

- SVM Using Kernel

Modeling with this algorithm is done with three variations. Where variation is done by using the kernel from SVM. Confussion matrix SVM using linear kernel described in table 5, kernel Polynomual described in table 6, kernel gausian described in table 7.

Table 5 : Confussion matrix Model SVM (kernel = Linear)

		Prediciton (received/rejected)	
		0	1
Actual	0	9727	1625
	1	1754	4099

Table 6 : Confussion matrix Model SVM (kernel = Polynomial)

		Prediciton (received/rejected)	
		0	1
Actual	0	11157	195
	1	5223	630

Table 7 : Confussion matrix Model SVM (kernel = gausian)

		Prediciton (received/rejected)	
		0	1
Actual	0	11290	62
	1	5726	127

- Random Forest

There are two experiments used to build models.first experiment (default parameter of sklearn.ensamble). First experiment confussion matrix described in table 8, and the second experiment described in table 9.

Table 8 : Random forest First experiment Confussion Matrix.

		Prediciton (received/rejected)	
		0	1
Actual	0	9937	1415
	1	1728	4125

Second experiment (first experiment model tuning). In this second experiment, the results of modeling in the first experiment are tuned. Model tuning is done by correcting the parameter values as follows:

```
n_estimators = np.arange(100,1200,200)
max_depth = np.arange(10,120,10)
min_samples_split = [20,30,50]
min_samples_leaf = [10,20,30,40]
```

With the new parameter range, the best parameters will be obtained by using the RandomForestClassifier library. So that the best parameters obtained are:

```
'n_estimators': 1100,
'min_samples_split': 50,
'min_samples_leaf': 10,
'max_features': 'auto',
'max_depth': 10,
'bootstrap': True
```

Table 9 : Random forest Second experiment Confussion Mattrix.

		Prediciton (received/rejected)	
		0	1
actual	0	9901	1451
	1	1714	4139

- Naive Bayes

In experiments using Naive Bayes, the development of the model uses the default parameters ({‘priors’: None, ‘var_smoothing’: 1e-09}) of sklearn.naive_bayes produces a model with the following results. Confussion matrix for Naive Bayes described in table 10.

Table 10. Naive bayes experiment Confussion Mattrix.

		Prediciton (received/rejected)	
		0	1
actual	0	6069	5283
	1	854	4999

4.7 Model Evaluation

In this section, the best modeling that can be used to predict whether proposals are submitted to the insurance company PT. XYZ will be accepted or rejected (requires re-analysis).

In this paper, accuration, precision, recall are the need to be campared for every models. Table 11 described model evaluation using SVM, table 12 model evaluation using random forest and table 13 model evaluation Naive Bayes. The summary model evaluation result for all models described in table 14 and figure 9 as graph version.

Table 11 : SVM using kernel model evaluation

	Kernel Linear	Kernel Polynomial	Radial Basis Function (RBF)
Accuration	0.80	0.69	0.66
Precision	0.72	0.76	0.67
Recall	0.72	0.11	0.02
Training Accuration	0.809	0.687	0.67

Table 12 : Random Forest model evaluation

	First experiment	Second experiment
Accuration	0.82	0.82
Precision	0.74	0.74
Recall	0.70	0.72
Training Accuration	1.0	0.876

Table 13 : Naive Bayes model evaluation

	Prior (None)
Accuration	0.64
Precision	0.49
Recall	0.53
Training Accuration	0.647

Table 14 : Model Evaluation Summary Results

	Naive Bayes	SVM using Kernel (Linear)	Random Forest
Precision	0.49	0.72	0.76
Recall	0.53	0.72	0.72
Accuration	0.64	0.8	0.82

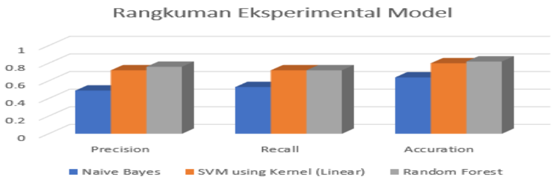

Figure 9 : Model Evaluation Summary Results Graph.

As the business goals of PT. XYZ in terms of auto underwriting with machine learning is to get clean proposals that will certainly be accepted and given policy services quickly. So, to determine the best model to adjust business objectives is to see the highest precision of the model developed. Precision describes the value of the presentation of the ability of the model in a positive data prediction (in this case the proposal is received).

In this study the model using the Random Forest algorithm gives the highest precision value of 0.85 followed by SVM using a linear kernel of 0.72 and Naive Bayes of 0.49. Thus, the algorithm that will be used in accordance with the business objectives of PT. XYZ is Random Forest.

4.8 Deployment

Deployment after this research will be done as figure 10.

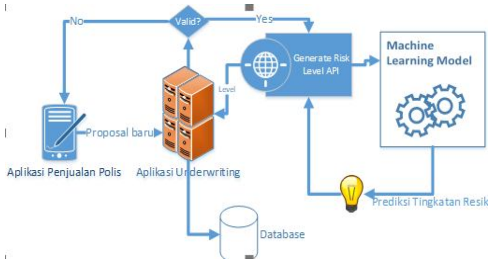

Figure 10 : Deployment of machine learning on the target system.

The built model will be deployed on a different server. API will be a link between applications that use machine learning.Implementation of machine learning in helping PT. XYZ to determine the level of risk prospective customers will shorten business processes, but will change some of the systems that are already running.

In addition to system changes and business processes, PT. XYZ will shorten the underwriting process time. After the data is complete and the Underwriting process begins, a risk assessment will be obtained after the machine learning process is complete. Thus TAT can be shortened and does not require 40 days as it is today. In addition, PT. XYZ does not need to provide human resources of more than 50 people to carry out the underwriting process. Manual work has been replaced with the system with the help of machine learning.

5. CONCLUSION

In this paper we can conclude that:Underwriting process in PT. XYZ can be assisted by machine learning to provide more profit for the company. Machine learning can help to speed up the risk assessment of potential customers. The current process requiremaximum 40 days (TAT Underwriting at PT. XYZ)while system will provide the result as soon as data already completed.

PT. XYZ can reduce the human resources needed as an underwriter. Human resources needed by PT. r can be replaced by machine learning.

This study found that the algorithm using the Random Forest have highest precision of 0.85. Model with the Support Vector Machine (SVM) algorithm using a linear kernel precision is 0.72 and precision using Naive Bayes algorithm is0.49.

REFERENCES

1. P. S. Nurfadilah, **"ekonomi.kompas.com,"** Kompas.com, 6 11 2018. [Online]. Available: https://ekonomi.kompas.com/read/2018/11/06/192907626/aaji-potensi-pasar-asuransi-jiwa-di-indonesia-masih-934-persen. [Accessed 5 8 2019].
2. A. Bhalla, **"Enhancement in Predictive Model for Insurance Underwriting, "**International Journal of Computer Science & Engineering Technology (IJCSET), vol. 3, 2012.
3. I. Poola**, "How Artificial Intelligence in Impacting Real Life Every day,"** International Journal of Advance Research and Development. , vol. 2, no. 10, pp. 96-100, 2017.
4. M. A. a. P. Deb, **"Machine learning: the new 'big thing' for competitive advantage,"**Int. J. Knowledge Engineering and Data Mining, vol. 5, no. 4, 2015.
5. A. D. A. P. N. R. Sumit Das, "**Applications of Artificial Intelligence in Machine Learning: Review and Prospect, "**International Journal of Computer Applications, vol. 115, no. 9, pp. 31-41, 2015.
   https://doi.org/10.5120/20182-2402
6. A. S. B. F. Tuga Mauritsius**, "Bank Marketing Data Mining using CRISP-DM Approach, "**International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no. 5, 2019.
7. N. A. A. K. M. O. N. Mohammad Ashraf Ottom, **"Classification of Mushroom Fungi Using Machine Learning Techniques,"** International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no. 5, 2019.
   https://doi.org/10.30534/ijatcse/2019/78852019
8. M. B. K. E. B. M. B. Mohssen Mohammed, **"Machine Learning Algorithms and Applications,"** Taylor & Francis Group, 2017.
9. Y. S. X. L. Yingjie Tian, **"recent advances on support vector machines research,"**technological and economic development of economy , pp. 5-33, 2012.
10. A. B. W. D. H. Anto Satriyo Nugroho, **"Support Vector Machine, Teori dan Aplikasinya dalam Bioinformatika,"** in Indonesian Scientific Meeting in Central Japan, Gifu, 2003.
11. K. M. K. M. R. A. A. A. F. Rabia Emhamed Al, **"Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity,"** in Journal of Electrical and Electronic Engineering and Information Technology, 2018.
12. G. V. A. G. V. M. B. ̈. T. O. G. M. B. P. P. R. W. V. D. J. V. A. P. D. C. M. B. P. Fabian Pedregosa, **"Scikit-learn: Machine Learning in Python,"**Journal of Machine Learning Research , 2011.
13. M. G. Mehmet Sait Vural, **"Criminal prediction using Naive Bayes theory,"** Neural Computing and Applications, 2016.
14. M. M. SONA **TAHERI, "learning the naive bayes classifier with optimization models,** "International Journal of Applied Mathematics and Computer Science, vol. 23, no. 4, p. 787–795, 2013.
    https://doi.org/10.2478/amcs-2013-0059
15. C. J. C. D. A. K. Santra, **"Genetic Algorithm and Confusion Matrix for Document Clustering,"** IJCSI International Journal of Computer Science Issues, vol. 9, no. 1, 2012.
16. D. POWERS, **"evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation,"**Journal of Machine Learning Technologies , vol. 2, no. 1, 2011.