ICIC Express Letters
Volume **15**, Number **12**, December **2021**
 ISSN 1881-803X
pp. **1329–1336**

CREDIT SCORING USING MACHINE LEARNING: A CASE STUDY FROM A LEASING COMPANY

Febri Dwiandriani* and Tuga Mauritsius

Information Systems Management Department, BINUS Graduate Program – Master of Information Systems Management
Bina Nusantara University
Jl. Kebon Jeruk Raya No. 27, Kebon Jeruk, Jakarta Barat 11530, Indonesia
*Corresponding author: febri.dwiandriani@binus.ac.id; tuga.mauritsius@binus.ac.id

Received February 2021; accepted May 2021

Abstract. *This study aims to find out how machine learning algorithms help and determine the level of risk of motor vehicle loan debtors. The data is obtained from PT XYZ, which is a leasing company providing vehicle loans based in Jakarta, Indonesia. In the modeling process, the researchers choose random forest algorithm, Naïve Bayes and support vector machine with kernel: rbf to be compared. Selection of the right algorithm depends on the results of modeling evaluation using the metrics precision, accuracy, and recall. The results of this study show that the model based on SVM with kernel: rbf algorithm demonstrates the highest performance with a precision value of 0.95 and an accuracy value of 0.81. With this level of accuracy, this study reveals that machine learning approaches coupled with the company's historical data may significantly contribute to assessing the risk of granting a loan.*
Keywords: Machine learning, Random forest, SVM with kernel: rbf, Naïve Bayes, Evaluation model

1. **Introduction.** Financing companies are "business entities that carry out goods and/or service financing activities" according to the Financial Services Authority Regulation Number 35/POJK.05/2018 Subsection 1 [1]. Multipurpose financing itself is a financing company for goods and/or services required by a debtor for use/consumption and not for business purposes or productive activities within the agreed period following the regulations of the Financial Services Authority Number 35/POJK.05/2018 Subsection 1 [1]. The growth of multipurpose finance companies is in line with human needs for fulfilling consumption in life, including consumption credit. According to a survey of Bank Indonesia in the fourth quarter of 2019 it showed that there was an increase in new credit from the consumption credit classification, one of which was motor vehicle credit by 24.4% compared to the previous quarter [2].

The growth of a business cannot be separated from the risks, namely the quality of credit receivables of each finance company. The quality of financing receivables can be seen from the Non-Performing Financing (NPF) of financing receivables. OJK recorded a net NPF per 2018 at 0.69%, a decrease from 2017 at 0.93%. Net NPF is the calculation of financing that is included in the substandard, doubtful, and loss-making category [3].

According to Bank Indonesia (BI) Regulation No. 7/2/PBI/2005 the risk of default or bad credit is recorded in the collectability which has substandard, doubtful, and loss conditions. Bad credit occurs when the debtor is in arrears of payment obligations for more than 90 days. Therefore, screening prospective debtors requires careful credit analysis and careful analysis. The standard principals of finance services are 5C (Character, Capacity, Capital, Collateral and Condition) and 7P (Personality, Party, Purpose, Prospect, Payment, Profitability and Protection) [4].

DOI: 10.24507/icicel.15.12.1329

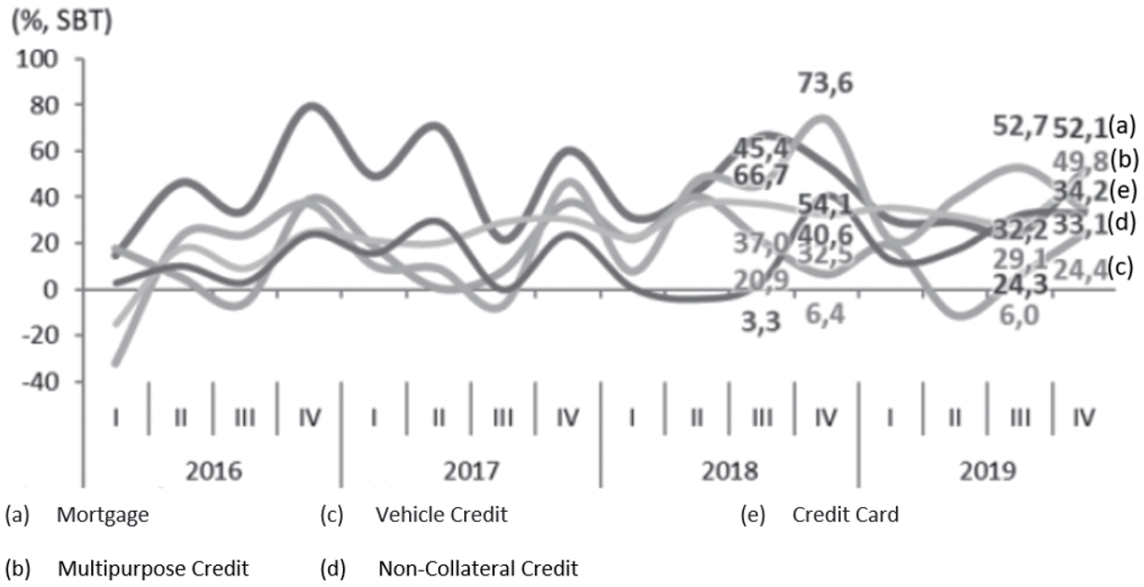


FIGURE 1. New credit growth by credit class [2]

By using the 5C and 7P, companies can find out the history of potential borrowers related to loan transactions and their background. This can be used by companies to innovate with the latest technological developments, Artificial Intelligence (AI). Quoted from the last report of the World Economic Forum and the French Prudential Authority (ACPR), the importance of developing artificial intelligence in the banking and insurance sector meets the minimum criteria for governance and control as well as transparency of credit risk [5].

From previous researches of machine learning in credit to determine good credit or bad credit, it can be concluded that machine learning can help crediting companies in making decisions to provide credit or credit interest to prospective customers. The proper application of the algorithm and evaluation matrix is able to show good data processing performance. In general, the selection of both classification and regression method algorithms can support data processing at least on a scale of thousands. For evaluation models such as accuracy, Receiver Operating Characteristic (ROC) curve, true positive rate, and recall, the selected algorithm is used to determine the algorithm's ability in accuracy, sensitivity, and rate value to data predictions [7,8].

Therefore, this study discusses the technique of applying machine learning algorithms by using three different algorithms, with a case study of an independent leasing company in Jakarta which has branch offices throughout Indonesia, and the company's business lines, namely providing vehicle loans and business funds for individuals or business entities. The aim of this research is to compare the results that are nearly accurate to predict the likelihood profile of the debtor with the profile of bad credit. The implementation of this research adopts the Cross-Industry Standard Process for Data Mining (CRIS-DM) framework with a good understanding of the business and needs of data mining projects.

2. Literature Review.

2.1. **Leasing company.** In general, leasing companies are classified as multipurpose finance companies. In the regulation of the Financial Services Authority Number 35/PO JK.05/2018 Subsection 1, the definition of a financing company is "a business entity that carries out financing activities for goods and/or services". The growth of finance companies in Indonesia until June 2019 was 182 companies, and the composition of financing receivables was still dominated by multipurpose financing of IDR 263.82 trillion or 60%

of total receivables [1]. Multipurpose Financing can be done in several ways as follows [1]:

1) Finance lease;

2) Purchases by installment payment;

3) Fund facilities, namely the purchase of goods and/or services distributed to debtors for consumption purposes and not for business purposes or productive activities within the agreed period.

2.2. **Machine learning.** Machine learning is an interesting part of artificial intelligent technology [6]. Machine learning is concerned with methods for increasing knowledge or performance, as improvements often involve analyzing data from the environment and making predictions about unknown quantities, and over many years aspects of data analysis [9]. The most important thing about machine learning, first is digging data to find patterns and build models. Second, use the results of data mining itself. The use of data mining results must inform the data mining process [10].

Machine learning techniques for this study use supervised learning because the data has specific target variables. Supervised techniques are used specifically for classification or predicting a target. Classification, regression, and causal modeling are generally solved by supervised methods [10]. The researcher decided to use three algorithms for modeling, namely: SVM using the kernel, Naïve Bayes and random forest. Based on previous research, the following is a comparison between the three algorithms [19,20].

Table 1. Comparing of algorithm capability

Comparison criteria	**SVM using kernel**	**Naïve Bayes**	**Random forest**
Accuracy in general	*	*	**
Speed of classification	****	****	*****
Tolerance to redundant data	***	*	**
Tolerance to missing value	**	****	***
Tolerance to irrelevant attributes	****	**	***
Tolerance to highly independent attributes	***	*	**
Tolerance to noise	**	***	**
Time for learning	***	*	***
Memory size	***	**	*****
Time for predicting	*****	**	*****

Support Vector Machine (SVM) is a machine learning approach, which is specifically used for classification and regression problems [11]. The basic SVM classification can produce a linear classifier that can be used to separate data that can be separated linearly. SVM can be extended to draw a non-linear decision boundary by converting the input from its original space to a high-dimensional space [12]. "Naïve Bayes classifier is an algorithm in data mining techniques that applies the Bayes theory in classification. The Bayes decision theorem is a fundamental statistical approach to pattern recognition. Naïve Bayes is based on the simplifying assumption that attribute values are conditionally independent when given an output value [13]".

The random forest model is an ensemble learning method that builds a series of decision trees at the time of training and produces a class which is a class mode (classification) or average prediction (regression) of each tree. The minimum number of samples required to split a node was set to two, and the minimum sample per leaf was set to one [14]. The selected algorithm model is then executed using the CRIS-DM framework guidelines.

3. **Proposed Methodology.** This research refers to the CRIS-DM (Cross-Industry Standard Process for Data Mining) framework. The CRIS-DM framework has a proposed process consisting of six steps starting with a good understanding of the business and the needs of a data mining project and ending with implementing solutions that meet specific business needs [9].

3.1. **Business understanding.** This research takes a case study of a financial company engaged in the provision of motor vehicle loans in Jakarta, Indonesia (called PT XYZ). In the existing condition, the provision of credit is carried out semi-automatically, meaning that there is still employee interference in the decision to grant credit. This has an impact on the length of data validation and decision making on the profile of prospective debtors even though it is assisted by the use of existing survey and reporting applications. The longest average time required by the head office to make decisions is 11 days to the head of the branch office by holding an internal meeting.

The decision on credit capability is based on the old credit history and profile of the prospective customer if the prospective customer is old meanwhile a new customer will be screening by its profile. That matter is done by the credit analysis department which coordinates with surveyor teams.

3.2. **Data understanding.** The data source from internal database which is taken from transaction data in 2019 and debtor profiles. The following is the information contained in it.

Table 2. Detail of variables

Id	Description	Data type
Gender	A collection of variables regarding the gender of the prospective debtor and has been normalized	Object
Marital_Sts	A collection of variables about the history of marital status	Object
Occupation	A collection of variables related to the job of a prospective debtor	Object
Education	Collection of variables related to the educational history of the prospective debtor	Object
Age	Age of normalized debtor candidates	Int64
Income	Collection of variables related to the monthly salary of prospective debtors	Int64
Dependency	A collection of variables about the number of members of a prospective customer except himself	Object
Credit_Sts	Target variable regarding the credit risk level of prospective debtors (1-0)	Int64

3.3. **Data preparation.** This stage is done to avoid missing data, avoid data errors (noisy data), and avoid inconsistent data in the dataset. Each data record and variable that is owned will be explored one by one so that we get a clean and quality dataset to develop a model. The data taken from the company database has gone through a normalization process but for the Age, Education, Marital_Sts, and Occupation variables, it is necessary to return to normalization through the next preprocessing process.

3.4. **Model building.** At this stage, there are repeated steps to build a model. To build a model, the input used is the training data and input parameters for the algorithm. The development uses a software that the same input will be tested on the Support Vector Machine (SVM) algorithm using the kernel, Naïve Bayes or random forest.

3.5. **Testing and evaluation.** The evaluation models will be validated using accuracy, precision, and recall. The accuracy value can be a reference for the performance of the algorithm used in the study. The precision value in this study answers the question "What percentage of prospective debtors are truly bad of all prospective debtors who are predicted to be bad?" while the recall value answers the research question "What percentage of prospective debtors are predicted to be bad compared to all prospective debtors who are actually bad?" [21].

3.6. **Deployment.** The implementation stage is made after the research gets the best model through testing and comparison of the three selected models. The development design starts after all the data is valid, and the system will automatically process it using the API to process the data carried out by the machine learning model that is built. The prospective debtor data will be managed by a machine so that it automatically generates a prediction of the risk level of potential debtors. The results will then be stored in a database that can be used to follow the company's business processes.

4. **Result and Discussion.** This research will provide new knowledge in the form of a system that helps the credit analysis process which can produce a faster and more accurate level of risk for potential borrowers. The research data was taken from a multipurpose credit provider company in Jakarta, PT XYZ. The data used is credit history in 2019. Data for potential debtors is needed by the system to issue a level of risk by displaying values 1 and 0, where the value 1 is current credit status with 249,670 records while the value 0 is bad credit status with 100,205 records.

The variables on Table 2 will be cleansing and transformation phases according to the research needs. First, checking the input value is blank on each research data variable. For null input on Marital_sts, it is filled with "Married" as the standard input. Occupation is filled with "Others" and in education it is filled with "Undefined".

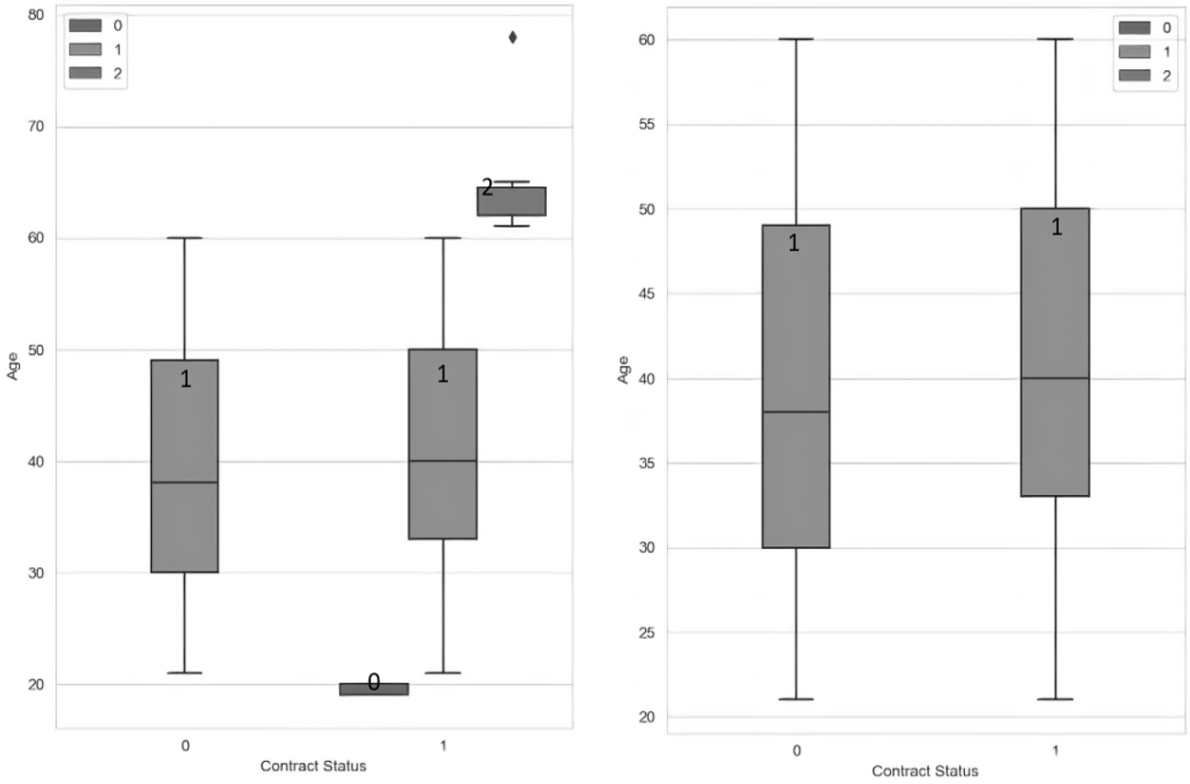
Before continuing to cleansing phase, a normalization is needed in the age variable. The age variable is discrete, and needs normalization to continuous. The normalization process will create the age_bins value with the grouping data into 3 classes, class 0 ($<$ 21 years), class 1 (21-60 years), and class 2 ($>$ 60 years). Data cleaning is adjusted in this study to check the age variable because PT XYZ has standards for the age group as one of the requirements for applying for a motor vehicle loan is someone aged 21-60 years (according to OJK regulations).

Figure 2(a) shows that there are prospective debtors who show that there are outliers aged over 60 years and less than 21 years old in the good credit class (1). Meanwhile, in the bad credit class (0) the age range is under the standard with an average below 40. Good credit class (1) has an average age range of 40 years.

So it is necessary to remove outliers that does not exist within the range of 21 years to 60 years. The change of class 0 value is 100,205 and class 1 value is 249,636, to produce a diagram like Figure 2(b).

After the cleansing and transformation phases, the data tested for modeling was divided into 70% training data and 30% testing data. The modeling process is carried out using Jupyter Lap version 2.1.5 and Python 3.8.3. The setting parameters of algorithms are shown as Table 3.

After the modeling process for each algorithm, an evaluation result is obtained in the form of a classification report containing the precision, accuracy, and recall values. This study prioritizes the algorithm with the best precision value because it wants to get potential predictive results, where the value reflects the presentation value of the modeling ability to predict data positively. The results of classification reports on each algorithm can be seen in Table 4.



(a) (b)

FIGURE 2. Before (a) and after (b) the cleansing phase of the outlier in age variable

TABLE 3. Setting parameters of algorithm

	Random forest	**Naïve Bayes**	**SVM with kernel: rbf**
Parameter setting	n_estimators = 100 max_depth = 3 bootstrap = False min_samples_leaf = 1 min_samples_split = 2 criterion = entropy	prior = None var_smoothing = 1e-09	c = 10 gamma = 10 cache_size = 200

TABLE 4. Result of model evaluation

	Random forest	**Naïve Bayes**	**SVM with kernel: rbf**
Test accuracy	0.78	0.71	**0.81**
Precision	0.98	0.92	**0.95**
Recall	0.77	0.73	**0.82**

Based on the results of the classification report evaluation on each algorithm, it was found that the SVM with kernel: rbf algorithm was the best. Precision value of SVM with kernel: rbf is 0.95 with an accuracy of 0.81 and a recall of 0.82. The precision value of 0.95 is generated by the SVM with kernel: rbf algorithm towards approximately 349,000 records answering questions as much as 0.95% indicating the bad credit populations truth. While the accuracy value shows that the performance of the support vector machine algorithm with kernel: rbf is considered good enough to manage large data.

So in research with data from PT XYZ the algorithm that can be used is SVM with kernel: rbf. Through the random forest tree diagram (Figure 3) the flow of the profile

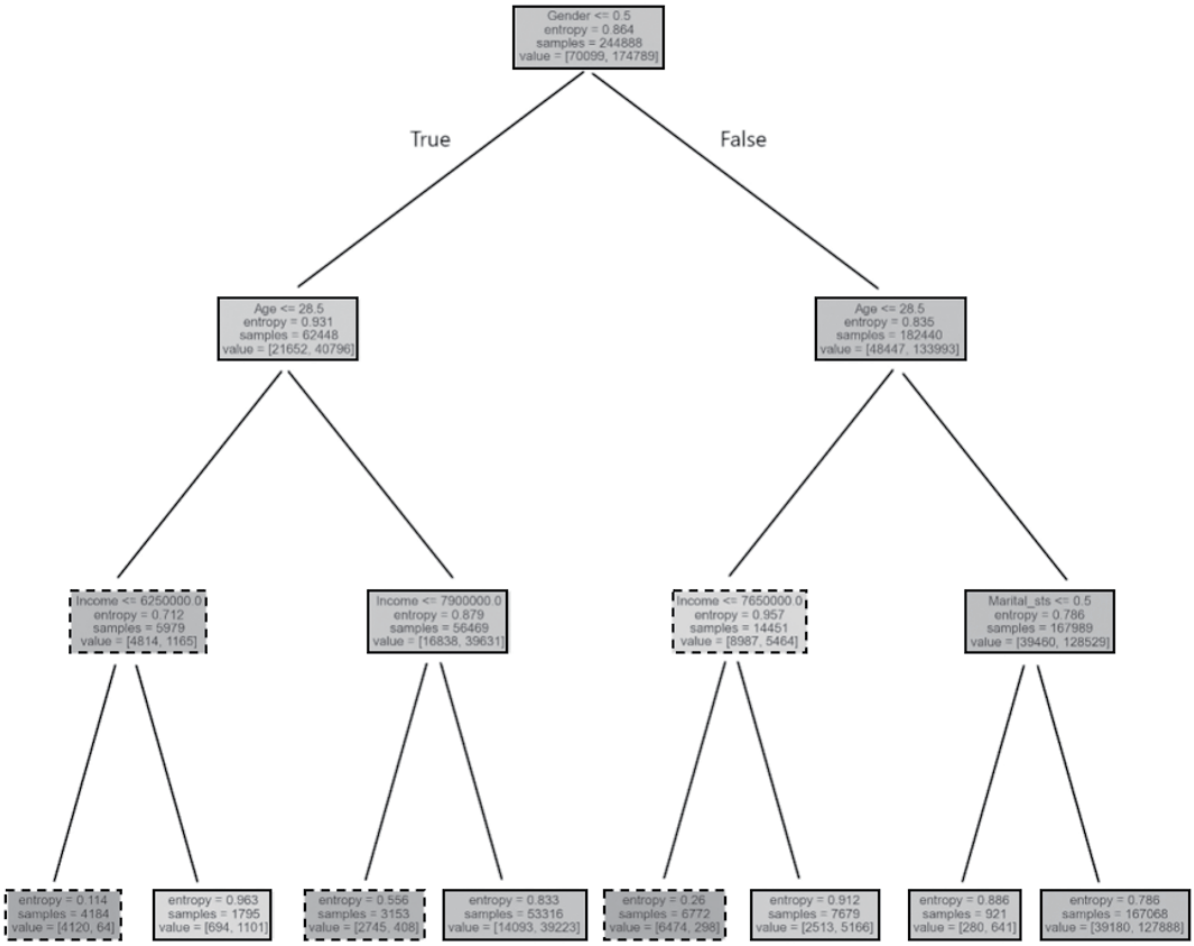


FIGURE 3. Tree diagram

decisions of potential debtors with bad credit can be seen. The solid box shows class 0 (bad credit) and the dash box shows class 1 (current credit).

The conclusion from the tree diagram shows that one of the potential categories of potential debtors including bad credit is someone with the gender of male (where gender > 0.5) and has an age of more than 28.5 years with marital status of divorce, where the total value is 127,888 of sample test, is 167,068 and the entropy value is 0.786 where the smaller the entropy value, the better it is used in extracting a class.

5. **Conclusion.** Based on results of this study, machine learning algorithms can help a vehicle credit company such as the PT XYZ case study to measure the level of risk of both good and bad debtors. Precision value of each algorithm reflects the presentation value of the modeling ability to predict data positively. In conclusion, the algorithm implementation for the PT XYZ case study can use the SVM with kernel: rbf algorithm. The model built will be implemented on a different server and the API works as a link between machine learning applications. Suggestions in further research are that researchers can develop other classification algorithms and data adjustments and supporting devices such as adequate hardware.

REFERENCES

[1] Financial Services Authority, *Risk Assessment of Money Laundering and Terrorism Financing Crimes in the Financial Services Sector in 2019*, Jakarta, 2019.
[2] Financial Services Authority, *Financial Institution Statistics 2018*, Jakarta, 2018.
[3] Bank Central of Indonesia, *Quarter IV-2019 Banking Survey*, Jakarta, 2019.

[4] Sudjana, Credit policies written off or written off by state-owned banks in the perspective of legal certainty, *Legal Policy Scientific Journal*, vol.12, no.3, pp.331-348, 2018.
[5] H. Phaure and E. Robin, *Artificial Intelligence for Credit Risk Management*, Deloitte SAS – Member of Deloitte Touche Tohmatsu Limited, Paris, 2020.
[6] S. Das, A. Dey, A. Pal and N. Roy, Applications of artificial intelligence in machine learning: Review and prospect, *International Journal of Computer Applications*, vol.115, no.9, pp.31-41, 2015.
[7] I. Menarianti, Classification of data mining in determining lending to cooperative customers, *Scientific Journal of Teknosains*, vol.1, no.1, pp.36-45, 2015.
[8] C. A. Hargreaves, Machine learning application to identify good credit customers, *International Journal of Advanced Engineering and Technology*, vol.3, no.3, pp.31-35, 2019.
[9] E. Turban, R. Sharda, D. Delen and D. King, *Business Intelligence: A Managerial Approach*, Pearson Education, Inc., New Jersey, 2011.
[10] F. Provost and T. Fawcett, *Data Science for Business*, O'Reilly, CA, 2013.
[11] H. Kaur and V. Kumari, Predictive modelling and analytics for diabetes using a machine learning approach, *Applied Computing and Informatics*, 2018.
[12] D. Kancherla, J. D. Bodapati and V. N, Effect of different kernels on the performance of an SVM based classification, *International Journal of Recent Technology and Engineering (IJRTE)*, vol.7, no.5S4, pp.1-6, 2019.
[13] R. M, S. H and S. M, Application of data mining for evaluating student academic performance using the Naive Bayes classifier algorithm, *EECCIS*, vol.7, no.1, pp.59-64, 2013.
[14] R. E. Al, K. M. Kwayu, M. R. Alkasisbeh and A. A. Frefer, Comparison of machine learning algorithms for predicting traffic accident severity, *Journal of Electrical and Electronic Engineering and Information Technology*, 2018.
[15] C. Albon, *Machine Learning with Python Cookbook*, O'Reilly, US, 2018.
[16] J. P. Mueller and L. Massaron, *Machine Learning for Dummies*, John Wiley & Sons, Inc., New Jersey, 2016.
[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher and M. Perro, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, vol.12, pp.2825-2830, 2011.
[18] B. J. Hutagaol and T. Mauritsius, Risk level prediction of life insurance applicant using machine learning, *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, vol.9, no.2, pp.2213-2220, 2020.
[19] F. Y. Osisanwo, J. E. T. Akinsola, O. Awodele, J. O. Hinmikaiye, O. Olakanmi and J. Akinjobi, Supervised machine learning algorithms: Classification and comparison, *International Journal of Computer Trends and Technology (IJCTT)*, vol.48, no.3, pp.128-138, 2017.
[20] V. Jatana, *Researchgate/Machine Learning Algorithms*, https://www.researchgate.net/publication/332902498, Accessed in August, 2020.
[21] D. M. W. Powers, Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation, *Journal of Machine Learning Technologies*, pp.37-63, 2011.