ISSN: 2508-7894 © 2021 KODISA & KAIA.
KJAI website: http://www.kjai.or.kr
doi: http://dx.doi.org/10.24225/kjai.2021.9.1.21

A customer credit Prediction Researched to Improve Credit Stability based on Artificial Intelligence

Ji-Hui MUN[1], Sang Woo JUNG[2]

Received: February 15, 2021. Revised: May 01, 2021. Accepted: June 05, 2021

Abstract

In this Paper, Since the 1990s, Korea's credit card industry has steadily developed. As a result, various problems have arisen, such as careless customer information management and loans to low-credit customers. This, in turn, had a high delinquency rate across the card industry and a negative impact on the economy. Therefore, in this paper, based on Azure, we analyze and predict the delinquency and delinquency periods of credit loans according to gender, own car, property, number of children, education level, marital status, and employment status through linear regression analysis and enhanced decision tree algorithm. These predictions can consequently reduce the likelihood of reckless credit lending and issuance of credit cards, reducing the number of bad creditors and reducing the risk of banks. In addition, after classifying and dividing the customer base based on the predicted result, it can be used as a basis for reducing the risk of credit loans by developing a credit product suitable for each customer. The predicted result through Azure showed that when predicting with Linear Regression and Boosted Decision Tree algorithm, the Boosted Decision Tree algorithm made more accurate prediction. In addition, we intend to increase the accuracy of the analysis by assigning a number to each data in the future and predicting again.

keywords : Machine Learning, Lnear Regression, Boosted Decision Tree Regression, Credit loans, Credit Card, Delinquency

Mayor Classification Code : Artificial Intelligence

1. Introduction

Korea's credit card industry has been steadily developing since the 1990s, and the number of credit card users in Korea is increasing. The introduction of credit cards in Korea provided many conveniences to customers and companies and contributed to the development of the national economy, but the neglect of credit card companies and the provision of credit cards to low-credit consumers led to high delinquency rates in the credit card industry (Lee & Park, 2015). In addition, measuring and evaluating the risk for various forms of credit risk assets is a critical decision-making issue for today's financial institutions, and since the 1997 and 2008 global financial crises, more careful management of financial assets has become important in the financial markets. Banks develop their own demonstration models and improvement of risk management in financial institutions can also help the economic growth of the country in which the institution belongs, which is why the industry is increasingly

[1] First Author, Student, Department of Medical IT, Eulji University, Korea, Email: mghmun@naver.com
[2] Corresponding Author, General Manager, ALLFORLAND Co.Ltd, Korea, Email: cki723@all4land.com

interested in data-based credit risk management models(Kim & Ahn, 2016).

Therefore, in this paper, based on prior research on the domestic credit card industry and prediction of corporate credit ratings, to create a model that predicts an individual's credit rating, Kaggle merges the personal information data submitted by the credit card applicant and the personal behavior information data. With Azure from Microsoft, we want to predict customer credit by analyzing information from Microsoft by gender, car, property, number of children, education, marital status, and employment status. The positive effect that can occur when these predictions are made is that you can predict the likelihood of a credit loan delinquency based on customer information, reducing indiscriminate credit card issuance and increasing bank credit. In addition, by developing various credit products according to the expected credit rating, you can reduce the risk of bad credit for individuals and reduce damage to the bank, contributing to strengthening the competitiveness of the bank. In this paper, we analyzed predictions using Linear Regression and Boosted Decision Tree algorithms and compared their performance.

2. Literature Review

2.1. Machine Learning and Credit rating models

Machine learning is a kind of artificial intelligence that programs a set of rules into a computer. Rather, it allows computers to derive rules and perform specific tasks by self-learning given data. Machine learning is as if humans can perceive certain situations through various experiences. It is the process of discovering rules or patterns that lead to such results by learning many data sets composed of values. According to the paper “A Study on Methods to Prevent the Spread of COVID-19 Based on Machine Learning written by Youngsang KWAK, Min Soo KANG, Using program Azure ML. The content of the study was conducted as a K program to find a self-diagnosis method to prevent the spread of COVID-19 based on machine learning(Kwak & Kang, 2020).

"Applying Artificial Intelligence for Diagnostic Classification" by Eun Soo Choi, Hee Jeong Yoo, Min Soo Kang, and Soon Ae Kim of Korean Autism Spectrum Disorder", Azure ML using a multiclass decision forest algorithm was applied, and the diagnostic algorithm score value of 1,269 Korean ADI-R test data was used for prediction. In the second experiment, we used 539 Korean ADI-R case data (over 48 months with verbal language) to apply mutual information to rank items used in the ADI diagnostic algorithm(Choi et al, 2020).

According to Hwang-hyun Kwon, data-based analysis and derivation of business insights based on it are considered more important than ever in the financial sector, and machine learning is an indispensable element in this process. Discussed how to apply machine learning to the evaluation model(Kwon, 2020).

2.2. Artificial Intelligent Program

・Azure Machine Learning studio

Azure is Microsoft's cloud computing platform in service since 2010. Azure Machine Learning Studio can solve various problems that are inherent problems of existing machine learning libraries and tools. In Azure, data input, output, and visualization are natively supported, and representative machine learning algorithms used by data scientists are prepared. Azure can also use the platform to easily model and build web services to apply to a wide variety of devices. Unlike existing cloud platforms, it provides an easy-to-access GUI environment in consideration of user convenience. In the form of dragging and dropping blocks, scripts written in R and Python languages can be inserted in the form of blocks, and the results can be checked through visualization(Kang et al., 2018).

2.3. Linear Regression

Linear regression, which started in statistics, is being effectively used as a method for modeling and inference through learning about the distribution of data along with advances in artificial intelligence and machine learning technologies. Linear regression is a method of finding a linear correlation between one or more independent variables through data analysis and modeling. This method models the relationship with the dependent variable when there are several independent variables, and through this, the value of the dependent variable that needs to be predicted is inferred from the modeling results through learning from the existing data. The linear regression method is widely used as an analysis method to predict the result when the result to be predicted can be quantified and modeled in relation to several input variables. In modeling for predicting outcomes, a regression model is generally created using the least squares method and the predicted results are analyzed(Lim, 2018).

2.4. Boosted Decision Tree Regression

The decision tree promoted in Azure Machine Learning uses an efficient implementation of the Mart gradient elevation algorithm. Build each regression tree step by step by measuring the error in each step using a predefined loss function and correcting the error in the next step. In a regression problem, Boost builds a series of trees step by step and then selects the optimal tree with a random indistinguishable loss function.

Maximum number of leaves per tree in Boosted Decision Tree Regression displays the maximum number of terminal nodes (leaves) that can be created in any tree. The minimum number of samples per leaf node indicates the minimum number of cases required to create a terminal node (leaf) in the tree. The learning rate enters a number from 0 to 1 that defines the step size during learning. The learning rate determines the rate or rate of convergence of learners for the optimal solution. The number of trees generated shows the total number of decision trees to be created in the ensemble. By creating more decision trees, you can potentially get better tests, but it will increase your learning time. For the initial random number, enter an optional non-negative integer to use as an initial random value.

3. Related Research

3.1. Credit evaluation model

Today, there are many cases that require a lot of insightful results or interpretations. In addition, studies on the credit rating model continue using statistical methods.
Currently, what is commonly understood as a credit score. It is a process in which information about credit applicants and credit accounts is quantified and combined to calculate points. This score indicates credit power by using the customer's credit card issued and using the score(Wi, 2006).

3.2. Prediction and analysis of factors affecting credit loans.

An Analysis of Factors Affecting Janghee Lee and Hoyeong Jung's Household Credit Loan Opening, according to the commercial education research, the purpose of contributing to the system was to analyze the factors affecting the credit rating ability of household loans, which can predict and measure credit risk, and discover new credit rating items. As a result of the analysis, it was found that those with a credit card usage rate of 20-29% showed the least amount of delinquency in loans. Second, in terms of preference and satisfaction with the loan execution criteria, those who preferred or satisfied unguaranteed loans showed the least amount of delinquency. Third, in terms of sensitivity to loan interest rates, those who respond sensitively to loan interest rates and additional interest rates also showed the least amount of loan delinquency. Fourth, people with foreign travel experience were rather experienced in foreign travel. It was found that there is a relationship between foreign travel and loan delinquency(Lee & Jung, 2006).

3.3. Suggestion of credit evaluation criteria

According to a study on derivation of credit for the prediction of delinquency of mutual savings banks by Jun Heo and Seung-gon Ko, general mutual savings banks are an independent and flexible evaluation system according to the rapidly changing financial environment due to the small size of the scale and the lack of information on market changes. Without operating a credit rating, most of these are judged overdue based on personal rating information of general credit rating agencies, and there are many problems in the customer management system. Therefore, we compare the evaluation criteria of general credit rating agencies with the evaluation criteria based on the delinquency prediction model using data mining techniques to diagnose the feasibility and as a basis, we present the criteria for the credit rating system of the mutual savings bank(Heo & Ko, 2008).

3.4. Comparative study of credit rating prediction model

According to Comparative study of prediction models for Hyeong-kwon Park and 3 others the corporate bond rating, the research on the corporate bond credit rating prediction model was conducted based on a number of financial characteristic variables that are expected to be used by credit rating agencies for corporate bond credit rating evaluation. support vector machine, and random forest. In addition, by integrating corporate bond credit rating data from 2013 to 2017 and financial indicators used in previous studies, the previously published forecasting models were applied to the same data and the forecasting performance was compared(Park et al., 2018).

After favoring the currency crisis, financial institutions introduced personal credit rating systems for lending to customers. The personal credit rating system predicts future credit risk and determines whether to approve a loan based on demographic information and banking performance of the customer. It is a decision system. Regarding the credit rating system raised earlier, we developed a model that can determine whether to approve a loan by predicting the credit risk for customers applying for a loan(Lee, 2012).

3.5. Prediction with Azure Machine Learning

According to Development of Prediction Model for

Diabetes Using Machine Learning by Duck-Jin kim and Zhixuan Quan, they used Azure machine learning provided by Microsoft to predict and study diabetes risk factors according to each data. The algorithms used for prediction are Two-class Logistic Regression and Multiclass Logistic Regression. The analysis results were used to predict high-risk and low-risk patients for diabetes and contribute to public health(Kim & Quan, 2018).

In “A study on Comparison of Lung Cancer Prediction Using Ensemble Machine Learning” by Yu-jin Nam and Won-ji Shin, the probability of lung cancer was predicted using Azure machine learning from Microsoft. They made predictions through kaggle's data SVM, Two-class Support Decision Jungle, and Multiclass Decision Jungle algorithm were used. Of the three algorithms, SVM's algorithm showed the most ideal accuracy. Through this, the probability of developing lung cancer was predicted. The lack of results was to cooperate with institutions in the future to secure and supplement lung cancer big data suitable for the domestic situation to improve accuracy(Nam & Shin, 2019).

In “A Study on Methods to Prevent Pima Indians Diabetes using SVM” by Sanghyuck You and Minsoo Kang, used Azure to find the major inclusions of Pima Indians based on machine learning. Using SVM (Support vector machine) as an analysis technique. Through a decision tree and correlation analysis, we found three important factors predicting Pima Indie diabetes at 70%. Diabetes diagnosis is validated by analyzing and preventing the cause of diabetes in advance by analyzing and preventing the cause of diabetes in advance(You & Kang, 2020).

4. Data Set

In this paper, we attempt to predict the possibility of default and borrowing of credit cards using personal information and data submitted by credit card applicants. So, to predict customer creditworthiness through machine learning, we conducted a study by combining two sets of data, one containing the personal information of the bank customer and the customer's behavior patterns. The data set for the study was taken from kaggle and the contents of the data set are shown in Table 1 below.

Table 1: Application_record.csv data description

Variable name	Variable content
ID	Customer number
CODE_GENDER	Gender
FLAG_OWN_CAR	Own Car
FLAG_OWN_REALTY	Property
CNT_CHILDREN	Number of children
AMT_INCOME_TOTAL	Annual income
NAME_INCOME_TYPE	Income category
NAME_EDUCATION_TYPE	Education level
NAME_FAMLIY_STATUS	Marital Status
NAME_HOUSING_TYPE	Housing Type
DAYS_BIRTH	Birth
DAYS_EMPLOYED	Start date of employment
FLAG_MOBIL	Do you have a cell phone
FLAG_WORK_PHONE	Mobile phone at work
FLAG_PHONE	Phone or not
FLAG_EMAIL	Email or not
OCCUPATION_TYPE	Occupation Type
CNT_FAM_MEMBERS	Family Size

Application_record.csv is a data set containing customer's personal information and consists of a total of 18 data including ID, gender, car, number of children, income, education level, family status, housing type, birth, employed, phone, email, family size. There are a total of 18 columns and 438,557 rows.

Table 2: Credit_record.csv data description

Variable name	Variable content
ID	Customer number
MONTH_BALANCE	Overdue month record
STATUS	Status

Note: Status Mean (0: Delinquency 1-29 days 1: Overdue 30-59 days 2: Overdue 60-89 days 3: Overdue 90-119 days 4: Overdue 120-149 days 5: Overdue or bad debt, overdue amortization 150 days C: Applicable Monthly repayment X: no monthly loan)

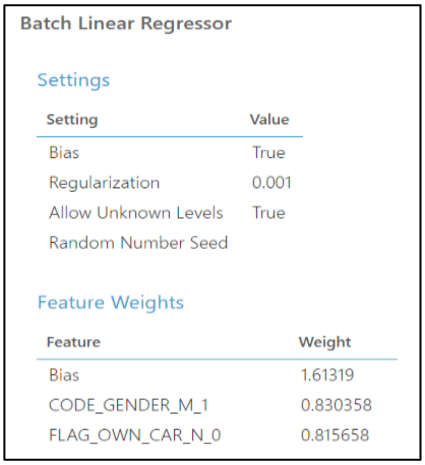
Credit_record.csv is a record of credit card user behavior. Each data is composed of ID, MONTH_BALANCE, and STATUS. MONTH_BALANCE proceeds in reverse order with the extracted month as the starting point. 0 is the current month and -1 is the previous month. It has a total of 3 columns and 1048575 rows.

5. Experimental model and Results

5.1. Experimental process model

In order to predict using bank customer information and credit loan delinquency data, preprocessing was performed with Clean Missing Data in application_record.csv and credit_record.csv, respectively. Also, to combine different data, it is combined through Join Data. The ID was set as the join key in order to combine the two data with an inner join and extract only the data with an intersection.

Next, the data after the inner join was pre-processed with Clean Missing Data once more, and then the necessary data was selected from the entire data using Select Columns in Dataset. In this paper, the goal is to predict the delinquency and delinquency period according to gender, own car, property, number of children, education level, marital status, and employment, so only necessary data were selected.



Batch Linear Regressor

Settings

Setting	Value
Bias	True
Regularization	0.001
Allow Unknown Levels	True
Random Number Seed	

Feature Weights

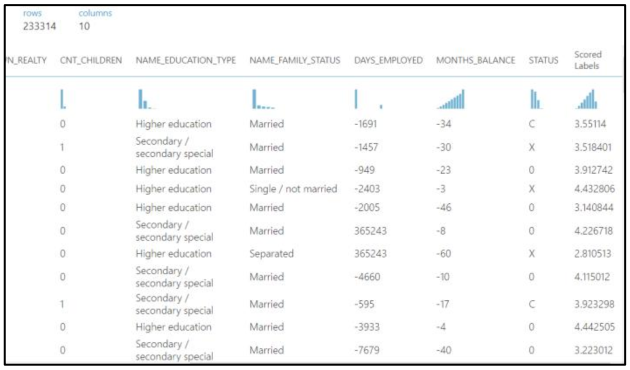
Feature	Weight
Bias	1.61319
CODE_GENDER_M_1	0.830358
FLAG_OWN_CAR_N_0	0.815658

Figure 1: Linear Regression Test Data Result

Then, the data was separated after setting the training data to 0.7 and the test data to 0.3 using Split Data. The linear regression algorithm and the enhanced decision tree regression algorithm were executed using STATUS (meaning delinquency) as a label in the Train Model for the linear regression algorithm and the enhanced decision tree regression algorithm. Least squares method was used as the linear regression algorithm, and the L2 regularization weight was set to 0.001. In addition, a single parameter was used as an enhanced decision tree regression algorithm, and all were set as default values. In order, it refers to the composition of the maximum number of leaves per tree (maximum terminal node), minimum number of samples, learning rate, and number of trees. The maximum number of leaves per tree is proportional to overfitting and training time but increasing the size of the tree and increasing the accuracy. You can increase it. Also, if the learning rate step size is too large, the optimal answer may not be obtained.

In addition, the more the number of trees is made, the wider range can be obtained, but the training time increases. Figure 1 is the result of visualizing the Train Model for the Linear Regression algorithm. Here, the degree to which the feature affects the label is called weight. When the feature has a value of 0, the default value of this label is called the bias. In Figure 1, when the feature values are all 0, it can be seen that the overdue value has a value of about 1.61319.

Figure 2 shows the Scored Model predicting the state of the evaluation data set for regression analysis among the two algorithms. Here, the Scored Labels area is a model that shows how to predict credit card delinquency when the same data as training data is provided.



rows 233314 columns 10

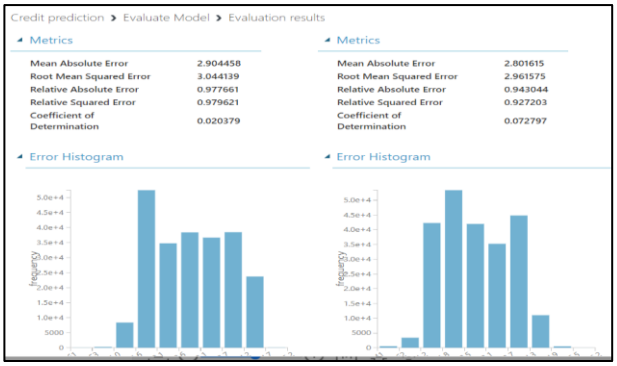
N_REALTY	CNT_CHILDREN	NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	DAYS_EMPLOYED	MONTHS_BALANCE	STATUS	Scored Labels
	0	Higher education	Married	-1691	-34	C	3.55114
	1	Secondary / secondary special	Married	-1457	-30	X	3.518401
	0	Higher education	Married	-949	-23	0	3.912742
	0	Higher education	Single / not married	-2403	-3	X	4.432806
	0	Higher education	Married	-2005	-46	0	3.140844
	0	Secondary / secondary special	Married	365243	-8	0	4.226718
	0	Higher education	Separated	365243	-60	X	2.810513
	0	Secondary / secondary special	Married	-4660	-10	0	4.115012
	1	Secondary / secondary special	Married	-595	-17	C	3.923298
	0	Higher education	Married	-3933	-4	0	4.442505
	0	Secondary / secondary special	Married	-7679	-40	0	3.223012

Figure 2: Score Model of Linear Regression

The score model is the process of showing how the model provides results individually as we enter each data we have. In other words, it can be said to be a similar principle to scoring, but the Score model allows us to compare whether the original data and the predicted data are the same while comparing what we have one by one. However, when there is a lot of data, it is difficult to grasp briefly how accurate the model is overall. The Evaluate Model, which is used to overcome this, serves to check the overall performance of the model.

5.2 Experimental model results

Figure 3 shows the Evaluate Model for Linear Regression and Boosted Decision Tree Regression, respectively. Mean Absolute Error of Linear Regression was 2.904458, and Boosted Decision Tree Regression was 2.801615. Coefficient of Determination is shown as a value between 0 and 1, Linear Regression is 0.020379 and Boosted Tree Regression is 0.072797



Credit prediction › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	2.904458
Root Mean Squared Error	3.044139
Relative Absolute Error	0.977661
Relative Squared Error	0.979621
Coefficient of Determination	0.020379

Metrics

Mean Absolute Error	2.801615
Root Mean Squared Error	2.961575
Relative Absolute Error	0.943044
Relative Squared Error	0.927203
Coefficient of Determination	0.072797

Figure 3: the Evaluate Model for Linear Regression and Boosted Decision Tree Regression

Since MAE measures how close the prediction is to the actual result as the mean absolute error, the lower the score is, the better, and the Coefficient of Determination, the coefficient of determination, means that the closer the value is to 1, the better it is. In both algorithms, the value of the Boosted Decision Tree is fine, but it shows a better value than Linear Regression.

6. Conclusion

In this paper, we predicted the outcome of loan delinquency based on customer information based on bank customer information and behavioral data provided by Kaggle. In our country, since the financial crisis, the recognition of each individual and company's credit rating and the degree of credit rating prediction has become increasingly important. As a result, methods and algorithms for predicting current credit ratings with various predictions and criteria for predicting credit are also increasing. The research conducted in this paper shows the better results of the Boosted Decision Tree among the two algorithms that credit personal information using linear regression and Boosted Decision Tree based on Microsoft's Azure.

As a result, through this prediction, the bank predicts the delinquency of the customer in case of a credit loan, suggests whether the loan is approved or the amount available for loan, etc., and can be calculated according to customer information. It can also be used to categorize customer segments to develop a variety of credit products, and it can prevent excessive customer loan situations, contributing to the number of bad creditors in maintaining the number of bad credits in the country.

While conducting this research, it was unfortunate that Korea's customer information data could not be used to create a predictive model as it was conducted on the kaggle site. In addition, it was difficult to know which data to select and predict based on going through the process of combining two data into one. Although the result of checking the actual evaluation model is not good, it will be possible to make a better model if the data is specifically processed for the next prediction and then re-prediction is performed. In the next study, it will be good to consider a model that proceeds with prediction by assigning numbers to each data for more accurate prediction.

Acknowledgment

This work was supported by the research grant of the KODISA Scholarship Foundation in 2021.

References

Lee, I. H., & Park, M. S. (2015). The History and current Status of the Domestic Credit Card Industry. *Korea Economic Forum*. 8(3), 12-21.

Kim, S. J., & Ahn, H. C. (2016). Application of Random Forests to Corporate Credit Rating Prediction. *The Journal of Business and Economics, 32*(1), 187-211.

Kwak, Y. S., & Kang, M. S. (2020). A Study on Methods to Prevent the Spread of COVID-19 Based on Machine Learning. *Korean Journal of Artificial Intelligence, 8*(1), 7-9. doi: http://dx.doi.org/10.24225/kjai.2020.vol8.no1.7

Choi, E. S., Yoo, H. J., Kang, M. S., and Kim, S. A. (2019). Applying Artificial Intelligence for Diagnostic Classification of Korean Autism Spectrum Disorder. *Original Article, 17*(11), 1090-1095.
https://doi.org/10.30773/pi.2020.0211

Kwon, H. H. (2020). Machine Learning and Finance: Machine Learning-Based Credit Rating Model. Korea Economic Research Institute of KDB Industrial Bank, Seoul, Korea. https://eiec.kdi.re.kr/policy/domesticView.do?ac=0000151746&issus=S&pp=20&datecount=&pg=

Kang, M. S., Kang, H. J., Yoo, K. B., Ihm, C. H., & Choi, E. S. (2018). *Getting started with Machine Learning using Azure Machine Learning studio*. Seoul, Korea: Hanti media.

Lim, H. I. (2018). Similarity Analysis of Programs through Linear Regression of Code Distribution. *Journal of Digital Contents Society*. *19*(7), 1357-1363.

Microsoft. (2020). Retrieved October 2020 from https://docs.microsoft.com/ko-kr/azure/machine-larning/algorithm-module-reference/boosted-decision-tree-regression

Lee, J. H., & Jung, H. Y. (2006). Factors Analysis on the Credit Scoring System in the Credit Loan of Household from Bank. *The Journal of Business Education*, *12*, 201-220.

Wi, S. S. (2006). How to select the optimal model for predicting personal credit score. A Case of Chonnam City. Domestic Master's Thesis, Chonnam National University Graduate School, Chonnam.

Heo, J., & Ko S. G. (2008). A Study on the Deriving Credit Score for Estimating Loan Delinquency in a Mutual Saving Bank. *Journal of The Korean Data Analysis Society*, *10*(5), 2795-2809.

Park, H. W., Kang, J. Y., Heo, S. W., & Yu, D. H. Comparative study of prediction models for corporate bond rating. *The Korean Journal of Applied Statistics, 31*(3), 367-382.

Lee, S. M. (2012). A study on development of credit scoring model by using classification methods. Domestic Master's Thesis University of Ulsan Graduate School, Ulsan

Kim, D. J., & Quan, Z. (2018). Development of Prediction Model for Diabetes Using Machine Learning. *Korean Journal of Artificial Intelligence*, *6*(1), 16-20

Nam, Y. J., & Shin, W. J. (2019). A Study on Comparison of Lung Cancer Prediction Using Ensemble Machine Learning. *Korean Journal of Artificial Intelligence*, 7(2), 19-24.

You, S. H., & Kang, M.S. (2020). A Study on Methods to Prevent Pima Indians Diabetes using SVM. *Korean Journal of Artificial Intelligence, 8*(2), 7-10.