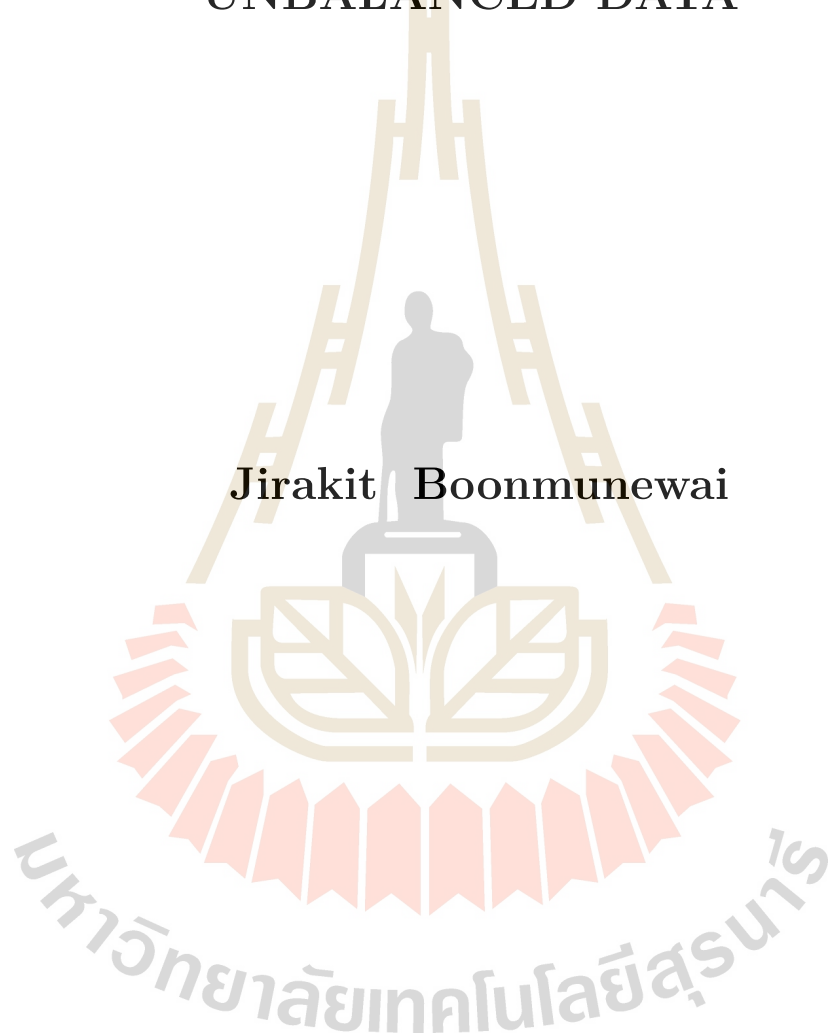


**A COMPARATIVE STUDY OF MACHINE
LEARNING TECHNIQUES TO DEAL WITH
UNBALANCED DATA**

Jirakit Boonmunewai



**A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Applied Mathematics**

Suranaree University of Technology

Academic Year 2019

การศึกษาเปรียบเทียบเทคนิคการเรียนรู้ของเครื่อง
สำหรับข้อมูลที่มีปัญหาข้อมูลไม่สมดุล



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาคณิตศาสตร์ประยุกต์

มหาวิทยาลัยเทคโนโลยีสุรนารี

ปีการศึกษา 2562

A COMPARATIVE STUDY OF MACHINE LEARNING TECHNIQUES TO DEAL WITH UNBALANCED DATA

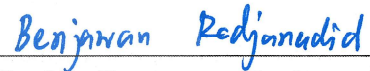
Suranaree University of Technology has approved this thesis submitted in partial fulfillment of the requirements for a Master's Degree.

Thesis Examining Committee



(Assoc.Prof.Dr.Eckart Schulz)

Chairperson



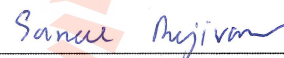
(Asst.Prof.Dr.Benjawan Rodjanadid)

Member (Thesis Advisor)



(Asst.Prof.Dr.Jessada Tanthanuch)

Member (Thesis Co-advisor)



(Assoc.Prof.Dr.Sanae Rujivan)

Member



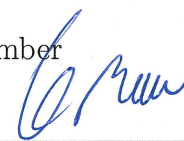
(Dr.Panu Yimmaung)

Member



(Dr.Amornrat Suriyawichitseranee)

Member



(Assoc.Prof.Flt. Lt. Dr.Kontorn Chamniprasart) (Assoc.Prof.Dr.Worawat Meevasana)

Vice Rector for Academic Affairs
and Internationalization

Dean of Institute of Science

จิรกฤต บุญหมื่น ไวย : การศึกษาเปรียบเทียบเทคนิคการเรียนรู้ของเครื่อง
สำหรับข้อมูลที่มีปัญหาข้อมูลไม่สมดุล (A COMPARATIVE STUDY OF MACHINE
LEARNING TECHNIQUES TO DEAL WITH UNBALANCED DATA).
อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร.เบญจวรรณ โรจนดิษฐ์, 64 หน้า.

การพยากรณ์การขอยกเลิกใช้บริการ/ปัญหาความไม่สมดุล/ต้นไม้การตัดสินใจ/
ตัวแบบนาอูฟเบย์/ซัพพอร์ตเวกเตอร์แมชชีน

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาปัญหาความไม่สมดุลที่ส่งผลต่อการขอยกเลิกใช้บริการ
ของลูกค้าธนาคาร โดยใช้เทคนิคการสุ่มตัวอย่างก่อนที่จะนำไปสร้างตัวแบบเพื่อใช้ในการพยากรณ์
แนวโน้มของลูกค้าที่จะขอยกเลิกใช้บริการ ในการศึกษาครั้งนี้พิจารณาการชักตัวอย่างแบบวิธี
สังเคราะห์ข้อมูลใหม่ และการชักตัวอย่างลดย่างสุ่ม รวมถึงศึกษาการสร้างตัวแบบด้วย วิธีต้นไม้
ตัดสินใจ วิธีจำแนกประเภทแบบนาอูฟเบย์ และวิธีซัพพอร์ตเวกเตอร์แมชชีน โปรแกรมหลักที่ใช้
ในการดำเนินการวิจัยได้แก่ RapidMiner Studio รุ่น 9.6

จากผลการศึกษาพบว่าการใช้การสุ่มตัวอย่างแบบวิธีสังเคราะห์ข้อมูลใหม่ร่วมกับการสร้าง
ตัวแบบด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน ให้ประสิทธิภาพในการทำนายสูงที่สุด โดยมีค่าความ
แม่นยำร้อยละ 92.99 ค่าวัดประสิทธิภาพร้อยละ 91.37 ค่า AUC ร้อยละ 96.4 และ อัตราลบเท็จร้อยละ
ละ 7.01

สาขาวิชาคณิตศาสตร์

ปีการศึกษา 2562

ลายมือชื่อนักศึกษา

ลายมือชื่ออาจารย์ที่ปรึกษา

ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

จิรกฤต

เบญจวรรณ

J. Tanthana

JIRAKIT BOONMUNEWAI : A COMPARATIVE STUDY OF
MACHINE LEARNING TECHNIQUES TO DEAL WITH
UNBALANCED DATA. THESIS ADVISOR : ASST. PROF.
BENJAWAN RODJANADID, Ph.D. 64 PP.

CHURN PREDICTION/IMBALANCED PROBLEM/
DECISION TREE/NAÏVE BAYES/SUPPORT VECTOR MACHINE

The purpose of this research was to study an imbalance problem which affects to the churn prediction for bank customers. In this study, we considered two sampling techniques, synthetic minority over-sampling technique (SMOTE) and random under-sampling, and three prediction modelings, i.e. decision tree classifier, Naïve Bayes classifier and support vector machine classifier. All calculations in the thesis were done by Rapid Miner Studio software version 9.6.

The study showed that the support vector machine classification model with SMOTE sampling technique performed most efficiently, with recall 92.99%, F-score 91.37%, area under curve (AUC) 96.4% and false negative rate 7.01%.

School of Mathematics

Academic Year 2019

Student's Signature Jirakit Boonmuneai

Advisor's Signature Benjawan Rodjanadid

Co-Advisor's Signature J. Tanthand

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my thesis advisor Asst. Prof. Dr. Benjawan Rodjanadid and co-advisor Asst. Prof. Dr. Jessada Tanthanuch for giving me the opportunity to do this research. They were kind in supervising and guiding me every time when I encountered difficulties in doing this thesis. In addition, they also assisted me with the SUT-thesis format in \LaTeX . Further more, I also would like to give many thanks for the professional support received from Dr. Panu Yimmuang, Assoc. Prof. Dr. Eckart Schulz and all professors in School of Mathematics, Institute of Science, Suranaree University of Technology (SUT). Also many thanks to Mr. Mangkon Damnet, Mr. Pannawit Kongmuangpak, Mr. Wipoosit Choongklang and all of friends in SUT for taking parts in supporting my work.

My gratitude extends to all people of my family for their kindness giving me a good encouragement.

Jirakit Boonmunewai

มหาวิทยาลัยเทคโนโลยีสุรนารี

CONTENTS

| | Page |
|--|-------------|
| ABSTRACT IN THAI | I |
| ABSTRACT IN ENGLISH | II |
| ACKNOWLEDGEMENTS | III |
| CONTENTS | IV |
| LIST OF TABLES | VII |
| LIST OF FIGURES | VIII |
| CHAPTER | |
| I INTRODUCTION | 1 |
| 1.1 Research Objectives | 2 |
| 1.2 Scope and Limitations | 2 |
| 1.3 Research Procedure | 3 |
| 1.4 Expected Results | 3 |
| II LITERATURE REVIEW | 4 |
| 2.1 Artificial Intelligence | 4 |
| 2.2 Machine Learning | 4 |
| 2.2.1 Supervised Machine Learning | 5 |
| 2.3 Classification | 6 |
| 2.3.1 Decision tree | 8 |
| 2.3.2 Bayes classification methods | 10 |
| 2.3.3 Support vector machine | 11 |
| 2.4 Imbalance Data Technique | 13 |

CONTENTS (Continued)

| | Page |
|--|-------------|
| 2.4.1 Oversampling | 14 |
| 2.4.2 Undersampling | 14 |
| 2.4.3 Synthetic Minority Oversampling Technique (SMOTE) | 14 |
| 2.5 Model Evaluation | 15 |
| 2.5.1 Validation | 15 |
| 2.5.2 Confusion matrix | 16 |
| 2.5.3 Receiver operation characteristic curve (ROC)- Area under the ROC curve | 17 |
| 2.6 Related Researches | 18 |
| III RESEARCH METHODOLOGY | 21 |
| 3.1 Tools | 21 |
| 3.2 Data Collection | 21 |
| 3.3 Data Selection | 22 |
| 3.4 Data Preparation | 22 |
| 3.4.1 Normalizing data | 22 |
| 3.4.2 Encoding | 23 |
| 3.4.3 Data sampling | 23 |
| 3.5 Model Selection | 23 |
| 3.6 Data Splits | 24 |
| 3.7 Model Evaluation | 24 |
| IV RESULTS AND DISCUSSION | 25 |
| 4.1 Dataset | 25 |
| 4.2 Data Preparation | 32 |

CONTENTS (Continued)

| | Page |
|--|-------------|
| 4.2.1 Normalization data | 32 |
| 4.2.2 Encoding | 33 |
| 4.2.3 Sampling | 33 |
| 4.2.4 Data splitting | 35 |
| 4.2.5 Feature selection | 35 |
| 4.3 Modeling | 35 |
| 4.3.1 Decision tree | 35 |
| 4.3.2 Naïve Bayes | 37 |
| 4.3.3 Support vector machine | 38 |
| 4.4 Performance Evaluation | 40 |
| V CONCLUSION AND RECOMMENDATION | 46 |
| REFERENCES | 49 |
| APPENDICES | |
| APPENDIX A FEATURE SELECTION OF THE MODEL | 54 |
| A.1 The weight of attribute after used feature selection | 55 |
| APPENDIX B CONFUSION MATRIX PERFORMANCE | 59 |
| B.1 the confusion matrix for each model | 60 |
| CURRICULUM VITAE | 64 |

LIST OF TABLES

| Table | | Page |
|-------|--|------|
| 4.1 | Variable of customer data. | 26 |
| 4.2 | Descriptive statistics of continuous data. | 32 |
| 4.3 | Description of 5 variable. | 33 |
| 4.4 | Result of Decision tree model for a balanced dataset. | 36 |
| 4.5 | Result of Decision tree model for imbalance dataset. | 36 |
| 4.6 | Result of Naïve Bayes model for balanced dataset. | 37 |
| 4.7 | Result of Naïve Bayes model for imbalanced dataset. | 38 |
| 4.8 | Result of Support vector machine model for balanced dataset. | 39 |
| 4.9 | Result of Support vector machine model for imbalance dataset. | 39 |
| 4.10 | Evaluation results for training 70% of dataset, test 30% of dataset. | 40 |
| 4.11 | Evaluation results for training 80% of dataset, test 20% of dataset. | 41 |
| 4.12 | Evaluation results for 5-fold cross validation. | 41 |

LIST OF FIGURES

| Figure | | Page |
|--------|---|------|
| 2.1 | Process(a). | 7 |
| 2.2 | Process(b). | 7 |
| 2.3 | Decision tree model. | 8 |
| 2.4 | The Margin of hyperplane. | 12 |
| 2.5 | k -fold cross validation. | 16 |
| 2.6 | Confusion Matrix. | 17 |
| 2.7 | AUC and ROC curve. | 18 |
| 4.1 | Gender variable distribution. | 27 |
| 4.2 | Geography variable distribution. | 27 |
| 4.3 | HasCrCard variable distribution. | 28 |
| 4.4 | IsActiveMember variable distribution. | 28 |
| 4.5 | NumOfProducts variable distribution. | 29 |
| 4.6 | Tenure variable distribution. | 29 |
| 4.7 | Exited variable distribution. | 30 |
| 4.8 | Age variable Histogram. | 30 |
| 4.9 | Balance variable Histogram. | 31 |
| 4.10 | CreditScore variable Histogram. | 31 |
| 4.11 | EstimatedSalary variable Histogram. | 32 |
| 4.12 | Exited variable distribution after used sampling. | 34 |
| 4.13 | ROC curves of model 5-fold cross validation. | 42 |
| 4.14 | ROC curves of model split validation 70%. | 43 |

LIST OF FIGURES (Continued)

| Figure | | Page |
|---------------|---|-------------|
| 4.15 | ROC curves of model split validation 80%. | 44 |
| A.1 | Weight value for split validation 70%, 30% after using the optimize weight function in Rapidminer. | 56 |
| A.2 | Weight value for split validation 80%, 20% after using the optimize weight function in Rapidminer. | 57 |
| A.3 | Weight value for cross validation after using the optimize weight function in Rapidminer. | 58 |
| B.1 | Confusion matrix in cross validation 70%, 30% technique by using SVM, Decision tree and Naive Bayes classifier | 61 |
| B.2 | Confusion matrix in split validation 80%, 20% technique by using SVM, Decision tree and Naive Bayes classifier | 62 |
| B.3 | Confusion matrix in cross validation technique choose $k = 5$ by using SVM, Decision tree and Naive Bayes classifier | 63 |

CHAPTER I

INTRODUCTION

Many businesses, such as telephone service companies, credit card business, insurance company and retail groups, need to retain the current customers. It is worth to predict the characteristics of customers who will be discontinued or tend to cancel the service, which is called *churn prediction*. Churn prediction can help businesses determine when customer's behavior changes in advance, focusing on customers who are likely to stop the service. The prediction helps companies in creating a campaign for drawing the customers not to leave the service. It reduces the risk of the customer service termination.

Customer data plays a major role in the churn prediction. It is important to analyze how one can implement existent customer data in a prediction model. Moreover, the performance of the obtained model must be considered. Nowadays, artificial intelligence (AI) is the most popular approach to be applied for analysis and developing a churn prediction model. Decision tree, Naïve Bayes classification and support vector machine are branches of AI which can be used for modelling many prediction models. The mentioned techniques were deployed in this thesis for modelling churn prediction models. However, the number of data points used to create a churn prediction model is the major factor influencing the performance of the model. Normally, we need almost the same more enough number of data in each group for AI training. Here, the data used in the churn prediction model training is classified into 2 groups, retain and close. A balanced number of both groups makes the model creation efficient. Indeed, data in real life is unbalanced.

There are also some techniques to deal with unbalanced data. In this thesis, we considered synthetic minority over-sampling technique (SMOTE) and random under-sampling. The proposed techniques help in transforming unbalanced data to balanced data.

This research is to study an imbalanced data classification problem. Many techniques dealing with the problem will be reviewed. The class imbalance problem for the churn prediction model will be explored. An example related to the problem will be shown and compared. RapidMiner is the major program used in this thesis for implement the normalization of data, the re-sampling of data, balancing data and modelling the churn prediction model.

1.1 Research Objectives

1. To evaluate some techniques of resampling for imbalance data in churn prediction for bank customer.
2. To study some techniques of machine learning for solving the classification of churn problem.

1.2 Scope and Limitations

1. The techniques for resampling in this study consist of the random under-sampling method and SMOTE method.
2. The techniques for solving the classification problem in this study consist of the decision tree classifier, the Naïve Bayes classifier and SVM.

1.3 Research Procedure

The research work proceeded as follows:

1. Study the techniques for resampling, namely the random undersampling method and SMOTE.
2. Study classification algorithm in data mining, namely decision tree classifier, Naïve Bayes classifier and SVM.
3. Study the program Rapid Miner Studio Version 9.6.
4. Understand and prepare the bank customer data(data from <https://www.kaggle.com/shrutimechlearn/churn-modelling>).
5. Construct the model for customer churn prediction in bank customer.
6. Analyze the performances with each model that we get from(5).

1.4 Expected Results

1. Would be able to obtain the model for churn modeling imbalance data set.
2. Compare affect dataset balance and unbalance.
3. Compare the performance sampling technique.

CHAPTER II

LITERATURE REVIEW

In this chapter, the knowledge of the fundamental mathematics regarding classification problem, decision tree model, naïve Bayes model, svm, imbalance data techniques and related research work are reviewed.

2.1 Artificial Intelligence

Artificial Intelligence (AI) is one of the most famous fields in science and engineering. It currently encompasses a huge variety of subfields, ranging from the general (learning and perception) and relevant to any intellectual task. (Russell and Norvig, 2002) However, nowadays, the tool to solve most problems is machine learning, and deep learning is one of many machine learning which has been rapidly achieved.

2.2 Machine Learning

Machine learning is an application of AI that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. It focuses on the development of computer programs that can access data and use it to learn for themselves. The process of learning begins with observations or data to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers to learn automatically without human intervention or assistance and

adjust actions accordingly (Expert System Team 2017). Also machine learning algorithms are often categorized as the following:

1. Supervised Machine Learning
2. Unsupervised Machine Learning
3. Reinforcement Machine Learning

2.2.1 Supervised Machine Learning

Supervised Machine learning is the computational task of learning correlation between variables in training data set and then utilising this information for creating the predictive model capable of inferring annotations for new data (Fabris, Magalhaes, and Freitas, 2017). In supervised machine learning, we have an input variable X and an output variable Y and we use an algorithm to learn the mapping from the input to the output.

$$Y = f(X)$$

The goal is to approximate function so well that when the new input data X is introduced the model predicts the output variable Y for that data. The learning is called as supervised learning when instances are given with known labels. The feature can be continuous, categorical or binary (Kotsiantis, Kanellopoulos, and Pintelas, 2006). The supervised learning problem can be further grouped into regression and classification problem.

1. **Classification:** when the output variable is a categorical, such as "yes" or "no" and "churn" or "no churn" then it is considered as Classification problems.

2. **Regression:** when the output variable is a real value, then such problems are considered as Regression problem.

2.3 Classification

Classification is a function in data mining that assigns data in a collection to target classes. The objective of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks. Data classification is a two-step process, including a learning step and a classification step. In the first step, a classifier is built, which explains a predetermined set of data classes or concepts. This is the learning step (or training phase). The classification algorithm builds the classifier by analyzing a training set made up of a database and their associated class labels. Let X represent an n -dimensional attribute vector, $X = (x_1, x_2, \dots, x_n)$, depicting n measurements from n database attributes respectively, A_1, A_2, \dots, A_n . Each X is assumed to belong to a predefined class, which is determined by another database attribute called the class label attribute. The class label attribute is discrete-valued and unordered. This first step of the classification process can also be viewed as the learning of a mapping or function, $y = f(X)$, that can predict the associated class label y of a given X . In this learning step, we wish to develop a mapping or function that separates the data classes. This mapping is represented in the form of classification rules, decision trees, or mathematical formulae. In our example, the mapping is represented as classification rules that identifies loan applications as being either safe or risky (Figure 2.1). The rules can be used to categorize future data, as well as provide deeper insight into the data contents. They also provide a compressed data representation. In the second step (Figure 2.2), the model is used for classification

(Mitchell, 1997), (Han, Kamber, and Pei, 2012).

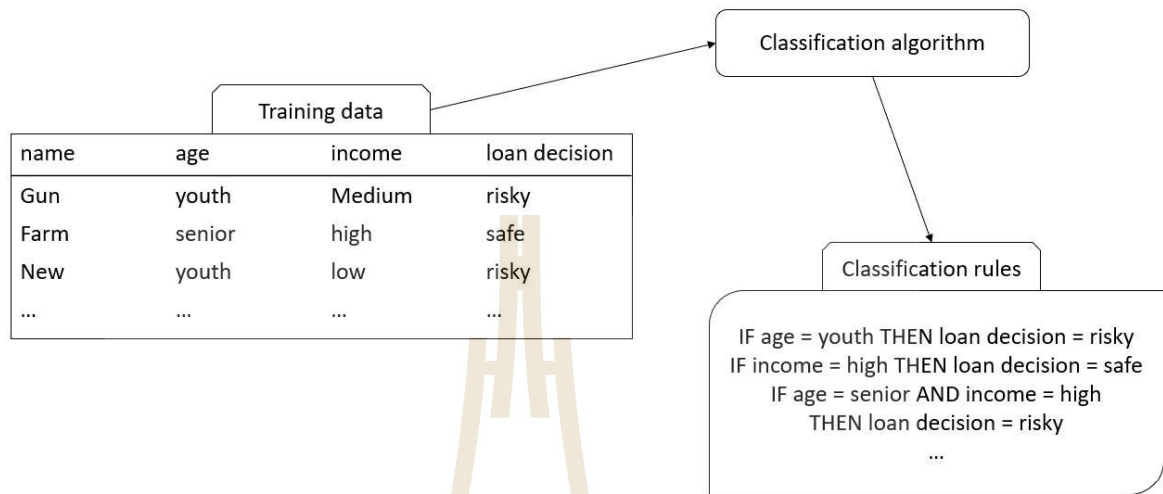


Figure 2.1 Process(a).

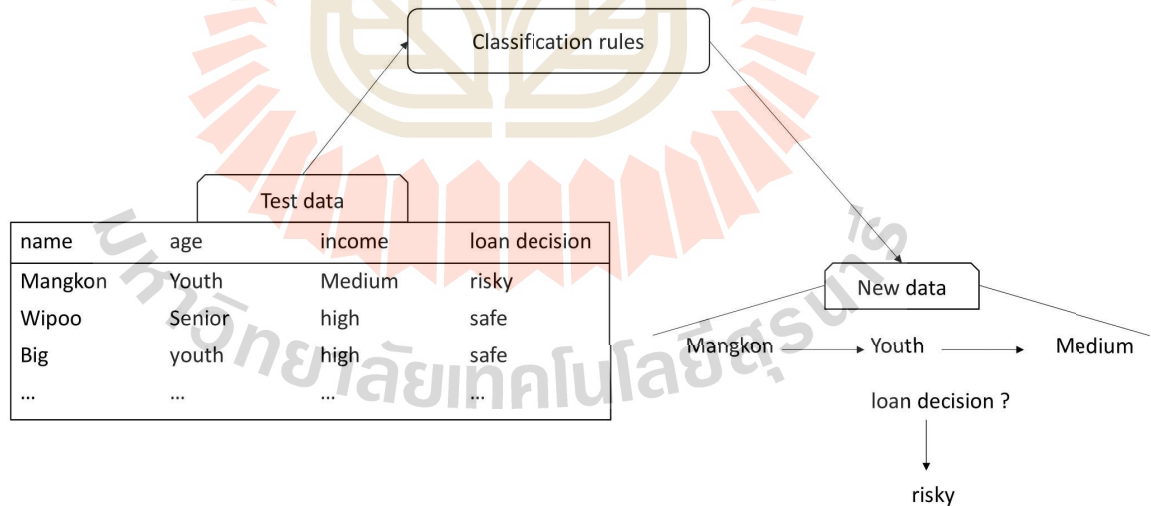


Figure 2.2 Process(b).

2.3.1 Decision tree

A decision tree is a flowchart-like tree structure (Han, Kamber, and Pei, 2012), where each decision node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. The top node in a tree is the root node. A typical decision tree is shown in Figure 2.3.

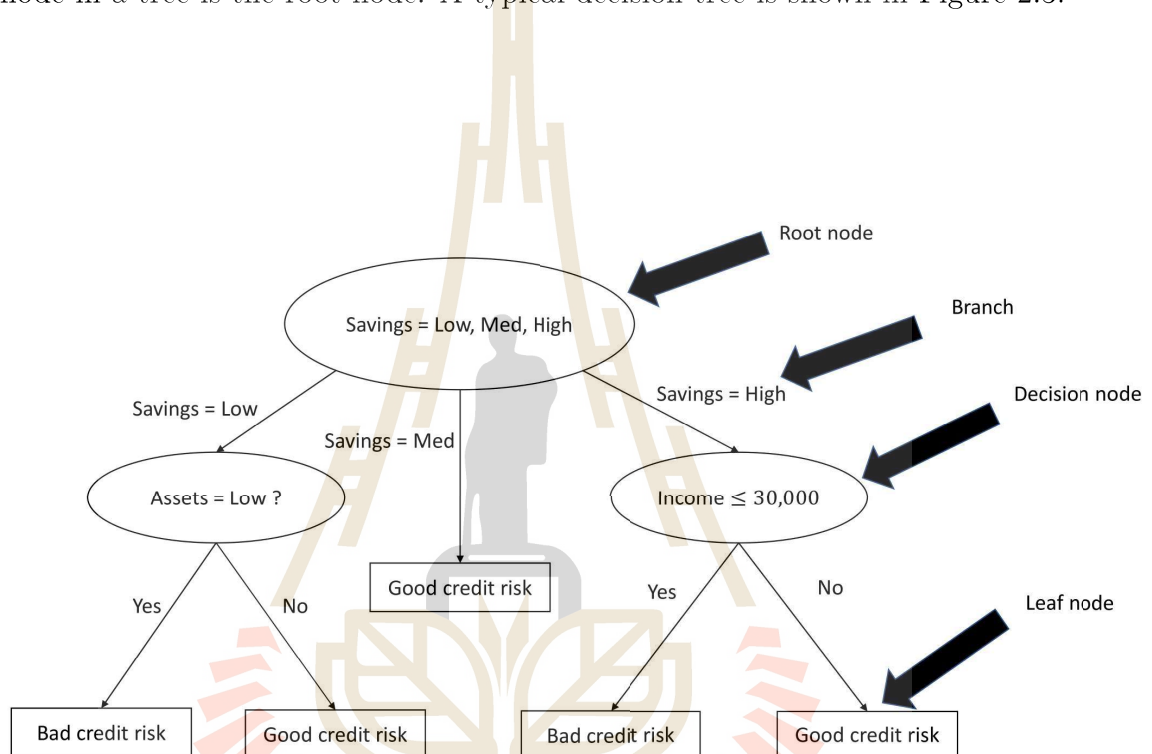


Figure 2.3 Decision tree model.

Given a data set X , for which the associated class label is unknown, the attribute values of the X are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that X . Decision trees can easily be converted to classification rules. In the late 1970s and early 1980s, J. Ross Quinlan, a researcher in machine learning, developed a decision tree algorithm known as ID3 (Iterative Dichotomiser). This work expanded the earlier

work on concept of learning systems, described by E. B. Hunt, J. Marin, and P. T. Stone. Quinlan, later presented C4.5 (a successor of ID3), which became a benchmark to which the newer supervised learning algorithms are often compared. In 1984, a group of statisticians (L. Breiman, J. Friedman, R. Olshen, and C. Stone) published the book Classification and Regression Trees (CART), which described the generation of binary decision trees. In ID3 we need to calculate entropy for each attribute in the data set and test the attribute from the value of entropy, then choose the attribute with high entropy gain to chooses decision node. From information theory, the size of entropy gain effects the attribute selection. So we use the attribute with highest entropy gain to be the decision node (Ying, 2017). If a set S has s samples inside, it will decide classification attributes with n values C_i , $i = (1, 2, 3, \dots, n)$, s_i is the number of samples in C_i . For a sample data set, the total Entropy is

$$I(s_1, s_2, s_3, \dots, s_n) = - \sum_{i=1}^n p_i \log_2(p_i)$$

p_i is the possibility of any sample belonging to C_i . For example, if a sample S has an attribute A , and A has w different values $a_1, a_2, a_3, \dots, a_w$, then sample S is split by attribute A into w subsets $S_1, S_2, S_3, \dots, S_w$. S has some samples in S_j , and it has value a_j in attribute A , and these subsets are new branches split by some value of attribute A . If s_{ij} is the number of samples in S_j whose class is C_i , the Entropy of A is

$$E(A) = \sum_{j=1}^w \frac{s_{1j} + s_{2j} + s_{3j} + \dots + s_{nj}}{s} I(s_{1j}, s_{2j}, s_{3j}, \dots, s_{nj}),$$

$$I(s_{1j}, s_{2j}, s_{3j}, \dots, s_{nj}) = - \sum_{i=1}^n p_{ij} \log_2 p_{ij}$$

$p_{ij} = \frac{|S_{ij}|}{S_j}$ is the possibility of samples in S_j belonging to class c_i . So when we use

attribute A to split S , the Entropy gain is

$$Gain(A) = I(s_1, s_2, s_3, \dots, s_n) - E(A)$$

C4.5 is the technique which we are interested in.

C4.5 In the C4.5 algorithm we need to calculate the entropy for each attribute in the data set and to find the value of gain ratio from attribute, then choose the attribute with higher gain ratio to split the node. The gain ratio is the value of gain divided by split information (J. Han, M. Kamber, and J. Pei, 2012). For example, if sample S has a feature A and A has w different values $a_1, a_2, a_3, \dots, a_w$, according to values of A we can divide sample S into w subsets $\{S_1, S_2, S_3, \dots, S_w\}$

$$GainRatio(A) = \frac{Gain(A)}{SplitI(A)}, SplitI(A) = - \sum_{j=1}^w p_j \log_2 p_j$$

2.3.2 Bayes classification methods

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities. Bayesian classification is based on Bayes' theorem. Bayes' theorem, named after 18th-century British mathematician Thomas Bayes, is a mathematical formula for determining conditional probability. Conditional probability is the probability of one event occurring with some relationship to one or more other events. The conditional probability of an event B is the probability that the event will occur given the knowledge that an event A has already occurred, this probability is written $P(B|A)$. We can calculate that by using Bayes' theorem as

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(B) \cdot P(A|B)}{P(A)}$$

We will use it in Naïve Bayes classifier.

Naïve Bayes classifier Naïve Bayes is a conditional probability model, given by a problem instance to be classified, represented by a vector $X = (x_1, x_2, x_3, \dots, x_n)$ representing some n features, it assigns to this instance probabilities $P(C_k|x_1, x_2, x_3, \dots, x_n)$, for each of K possible classes C_k . Using Bayes' theorem, the conditional probability can be decomposed as

$$P(C_k|X) = \frac{P(C_k) \cdot P(X|C_k)}{P(X)}$$

In Naïve Bayes classifier, the assumption of class conditional independence is made, then we get

$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i)$$

We can easily estimate the probabilities $P(x_1|C_i), P(x_2|C_i), P(x_3|C_i), \dots, P(x_n|C_i)$ from the training dataset.

2.3.3 Support vector machine

Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane, a method for the classification of both linear and nonlinear data. Its algorithms use a set of mathematical functions that are defined as the kernel. The function of the kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions. These functions can be different types. For example it uses a nonlinear mapping to transform the original training data into high dimension. To explain SVM, let's first look at the simplest case of two class problems where the classes are linearly separable. Let the data set D be given as $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_l, y_l)$, where $x_i \in R^n$ is the vector attribute of training dataset with associated class labels, y_i for $i = 1, 2, 3, \dots, l$. Each y_i can take one of two values, either 1 or -1 . If all the examples in D can be separated exactly by the hyperplane $w \cdot x + b = 0$ and

the distance from the nearest sample point to the hyperplane is the maximum, we state that the data samples can be separated by the optimal hyperplane, which is also called the maximum margin hyperplane as shown in (Figure 2.4) (He and Ma, 2013).

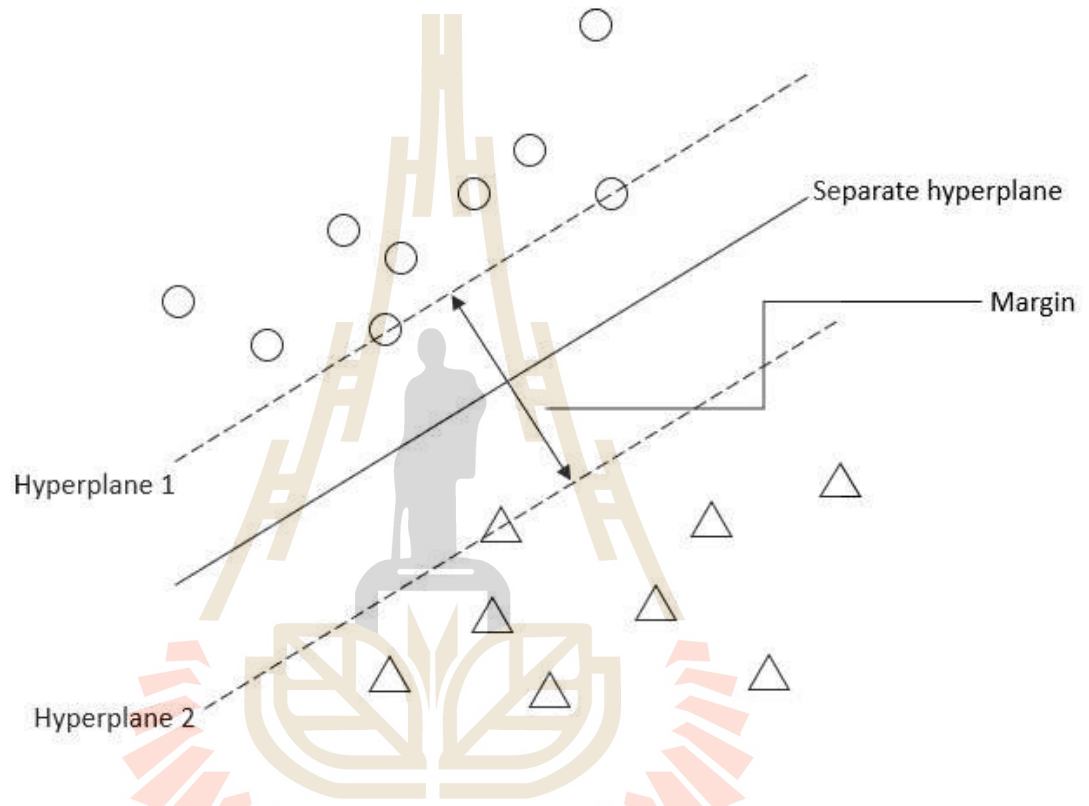


Figure 2.4 The Margin of hyperplane.

The problem of the optimal classification hyperplane is transformed into the following optimization problem by

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i, \\ & y_i(w \cdot x + b) - 1 + \xi_i \geq 0, \\ & \xi_i \geq 0, i = 1, 2, 3, \dots, l \end{aligned}$$

where $\xi = (\xi_1, \xi_2, \xi_3, \dots, \xi_l)^T$, C is the penalty parameter and $C > 0$, which controls the degree of penalty for misclassification samples. In addition, the greater

the value of C , the greater the penalty for error. The corresponding Lagrangian function is

$$L(w, b, \xi, \alpha) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i - \sum_{i=1}^l \alpha_i (y_i (w \cdot \varphi(x_i) + b) - 1 + \xi_i) - \sum_{i=1}^l \beta_i \xi_i$$

where α_i, β_i are Lagrangian multipliers and $\alpha_i > 0, \beta_i > 0$. We can obtain the following dual problem by

$$\begin{aligned} \min \quad & \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j k(x_i, x_j) - \sum_{i=1}^l \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^l \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, i = 1, 2, \dots, l, w \in R^n, b \in R \end{aligned}$$

where $k(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$ is the kernel function (Xie, W., Liang, G., Dong, Z., Tan, B., and Zhang, B. (2019)).

Definition (Kernel function) A function $k(x, y)$ is called a kernel on $R^n \times R^n$ if there exist a map φ from space R^n to Hilbert space H , $\varphi : R^n \rightarrow H$ such that $k(x, y) = \varphi(x) \cdot \varphi(y)$ where (\cdot) denotes the inner product of space H

2.4 Imbalance Data Technique

Imbalanced data typically refers to a problem with classification problems where the classes are not represented equally. There are problems where a class imbalance is not just common, it is expected. In binary classification, we call the class of more data the majority class and the other the minority class. For this problem, we are interested in random undersampling and synthetic minority oversampling technique (SMOTE).

2.4.1 Oversampling

Oversampling is a sampling technique which balances the data set by replicating the examples of the minority class. It is also called upsampling. Oversampling is also divided into two types: Random Oversampling and Informative Oversampling (Sonak and Patankar, 2015). Random Oversampling is the method which balances the class distribution by replicating the randomly chosen minority class examples. Informative Oversampling method synthetically generates minority class examples based on a pre-specified criterion (Ramyachitra and Manikandan, 2014).

2.4.2 Undersampling

Undersampling is an efficient method for balancing data. This method uses a subset of the majority class to train the classifier. In undersampling, we delete some examples of the majority class. Undersampling methods are divided into Random Undersampling and Informative Undersampling. Random Undersampling is simple, it randomly eliminates samples from the majority class till the data set gets balanced. Informative Undersampling method selects only the required majority class examples based on a pre-specified selection criterion to make the data set balanced (Sonak and Patankar, 2015).

2.4.3 Synthetic Minority Oversampling Technique (SMOTE)

Oversampling is a sampling technique which balances the data set by replicating the examples of the minority class, it is also called upsampling. Oversampling is also divided into two types: Random Oversampling and Informative

Oversampling (Sonak and Patankar, 2015). Random Oversampling is the method which balances the class distribution by replicating the randomly chosen minority class examples. Informative Oversampling method synthetically generates minority class examples based on a pre-specified criterion (Ramyachitra and Manikandan, 2014). This technique has the procedure as following: First, use k nearest neighbor to find k point, near the data point in minority class for point M , M is any point in the minority class. After that pick a random point two considered points, the picked point is called a random synthetic point. In 2-dimension we can define point $m_i(c_1, c_2, c_3, \dots, c_l)$, m_i is new point by

$$c_1 = a_1 + (b_1 - a_1) \times rand('UNIFORM'),$$

$$c_2 = a_2 + (b_2 - a_2) \times rand('UNIFORM'),$$

$$c_3 = a_3 + (b_3 - a_3) \times rand('UNIFORM'),$$

$$c_l = a_l + (b_l - a_l) \times rand('UNIFORM'),$$

where m_1 is point between point $M(a_1, a_2, a_3, \dots, a_l)$ and point $M_1(b_1, b_2, b_3, \dots, b_l)$, $a_1, a_2, a_3, \dots, a_l$ and $b_1, b_2, b_3, \dots, b_l$ are the features on the point and l is number of feature (Kesornsit, Lorchirachoonkul, and Jitthavech, 2018).

2.5 Model Evaluation

2.5.1 Validation

In the k -fold cross validation technique, the labeled information is separated into k equivalent fragments. One of the k sections is utilized for testing, and the

remaining $(k - 1)$ portions are utilized for training. This procedure is iterated k times by utilizing every one of the k fragments as the test set (Pantazi, X.-E., Moshou, and Bochtis, 2019).

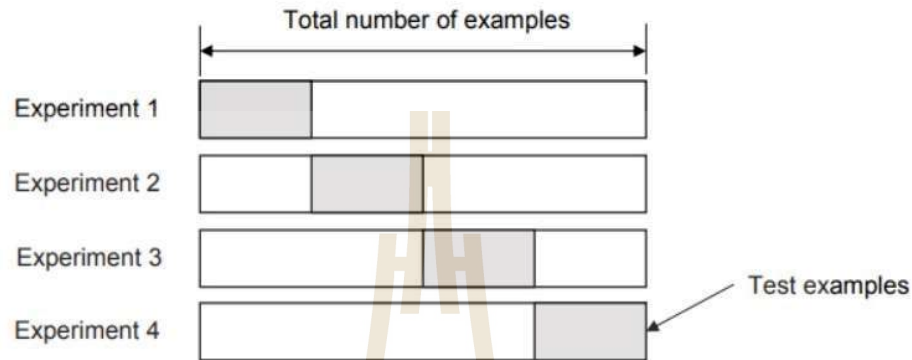


Figure 2.5 k -fold cross validation.

The advantage is that the entire data is used for training and testing. The error rate of the model is the average of the error rate of each iteration. This technique can also be called a form of the repeated hold-out method. The error rate could be improved by using the stratification technique.

2.5.2 Confusion matrix

Confusion matrix is a performance measurement for machine learning classification problem where the output can be two or more classes. It is a kind of table which helps know the performance of the classification model on a set of test data for which the true values are known (Xie Liang, Dong, Tan, and Zhang, 2019).

Where True positive(TP) is the case in which we predicted yes and the actual output was also yes, True negative(TN) is the cases in which we predicted no and the actual output was no and False positive(FP) is the cases in which we predicted yes and the actual output was no, False negative(FN) is the cases in

| | | | |
|------------------|----------|---------------------|---------------------|
| | | Actual Values | |
| | | Positive | Negative |
| Predicted Values | Positive | TP : True Positive | FP : False Positive |
| | Negative | FN : False Negative | TN : True Negative |

Figure 2.6 Confusion Matrix.

which we predicted no and the actual output was yes. Accuracy, Classification precision, Recall rate and F-measure for the matrix can be calculated as

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

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

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

$$F\text{-measure} = \frac{2(Precision \times Recall)}{Precision + Recall}$$

$$False\ Positive\ rate = \frac{FP}{FP + TN}$$

$$False\ Negative\ rate = \frac{FN}{FN + TP}$$

2.5.3 Receiver operation characteristic curve (ROC)- Area under the ROC curve

Receiver operation characteristic curve(ROC) is a probability curve and area under the ROC curve(AUC) represents degree or measure of separability. It shows the performance of the model's ability to distinguish between classes (Narkhede, 2016).The ROC curve is plotted with TP rate against the FP rate where TP rate is on the Y-axis and FP rate is on the X-axis. The AUC score is

calculated from ROC curve and its value lies between 0 and 1. The larger the value of AUC, the better the model performance is, while a model with 0.5 or below AUC value is accepted as not performing well. ROC curves are typically used in binary classification to study the output of a classifier. To extend ROC curve and ROC area for multi-label classification, it is necessary to binarize the output. One ROC curve can be drawn per label, but one can also draw a ROC curve by considering each element of the label indicator matrix as a binary prediction (micro-averaging).

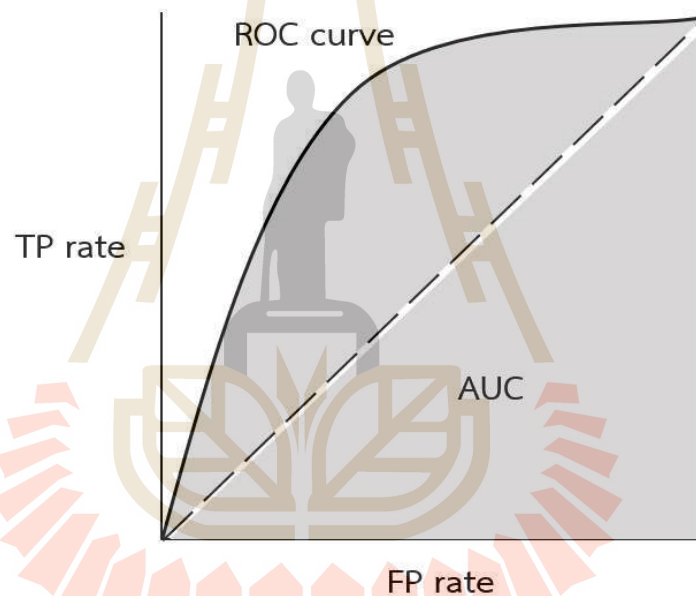


Figure 2.7 AUC and ROC curve.

2.6 Related Researches

Burez, J., and Van den Poel, D. (2009) described an empirical study of both sampling (random and advanced under-sampling) and using more appropriate evaluation metrics (AUC, lift). they investigated the increase in performance of sampling and two specific modeling techniques (gradient boosting and weighted

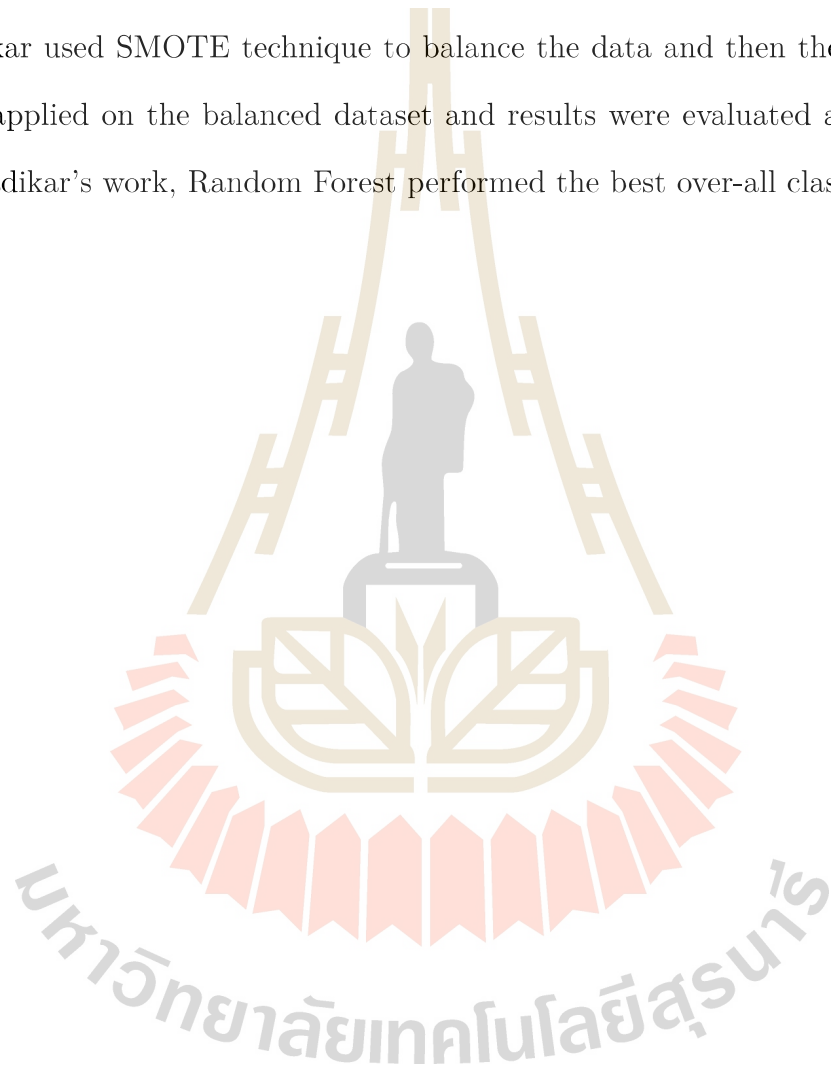
random forest) compare to some standard techniques (logistic regression, random forests). The result showed that under-sampling can lead to improved prediction accuracy, especially when evaluated with AUC. Unlike Ling and Li (1998). And the advanced sampling technique CUBE does not increase predictive performance. Another advanced sampling technique (e.g. SMOTE for over-sampling) might perform better.

Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., Hawalah, A., and Hussain, A. (2016) investigated six well-known sampling techniques and compared the performance of these techniques, i.e., mega-trend diffusion function (MTDF), synthetic minority oversampling technique (SMOTE), adaptive synthetic sampling approach, majority weighted minority oversampling technique, immune centroids oversampling technique and couples top-n reverse k-nearest neighbor—on four publicly available datasets for the telecommunications sector. The empirical results demonstrate that the overall predictive performance of MTDF and rules-generation based on genetic algorithm performed the best as compared with the rest of the evaluated oversampling methods and rules-generation algorithms.

Xie, Liang, Dong, Tan, and Zhang (2019) improved the oversampling algorithm based on the samples' selection strategy for the imbalanced data classification is proposed. On the basis of the Random-SMOTE algorithm, the support vectors are extracted and are treated as the parent samples to synthesize the new examples for the minority class in order to realize the balance of the data. Lastly, the imbalanced data sets are classified with the SVM classification algorithm. F-measure value, G-mean value, ROC curve, and AUC value are selected as the performance evaluation indexes. Experimental results showed that this improved algorithm demonstrates a good classification performance for the imbalanced data

sets.

Wadikar, D. (2020) compared several technics of supervised machine learning methods use for creating a churn prediction model, i.e. logistic regression, random forest, support vector machine(svm) and neural network. First, Wadikar applied all models yo the imbalance dataset and results were evaluated. Next, Wadikar used SMOTE technique to balance the data and then the same models were applied on the balanced dataset and results were evaluated and compared. In Wadikar's work, Random Forest performed the best over-all classifier.



CHAPTER III

RESEARCH METHODOLOGY

This chapter presents the process used in this research. The process consists of 5 parts as follows:

1. Data collection
2. Data selection
3. Data preparation
4. Model selection
5. Model evaluation

3.1 Tools

The empirical part of this study is done by using RapidMiner studio version 9.6 (education license). The RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation machine learning, deep learning, text mining and predictive analytics (wikipedia.org).

3.2 Data Collection

In this research we used data from the Kaggle dataset store, The churn Modeling classification dataset is available on <https://www.kaggle.com/shrutimechlearn/churn-modelling>

3.3 Data Selection

The working data in this research is extracted from the Kaggle dataset. This dataset contains details of the customers in a financial institution and the target variable is a binary variable reflecting the fact whether a customer has left the bank (closed his account : churn customer) or continues to be a customer. The data has 8,166 instances (203 churn and 7,963 non churn data) and 13 attributes. In this research, we extracted 11 attributes for building the required and targeted feature:

1. Age
2. Balance
3. Credit score
4. Estimated salary
5. Gender
6. Geography
7. Having credit card
8. Still active member
9. Number of products
10. Tenure
11. Exited

3.4 Data Preparation

3.4.1 Normalizing data

Normalization is a technique applied as a part of data pre-processing for the building machine learning method. The goal of normalization is to change the values to a common scale (Wadikar, 2020). In this research, the continuous variables were normalized, and we use Min-Max normalization. Min-Max normalization works by seeing how much greater a field values is, than the minimum

values $\min(X)$, and then scaling this difference by the range (Larose and Larose, 2014).

That is

$$X^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

where

X refers to our original field values,

X^* refer to the normalized field value.

3.4.2 Encoding

The dataset includes continuous and categorical variable. But SVM algorithm accepts only numeric data, so the categorical data is converted into 0 and 1 by using the label encoding. In this dataset, we found 5 variables were which categorical variables. These values were transformed to numerical type.

3.4.3 Data sampling

The dataset had 8,166 instances includes 203 churned customers and 7,963 non-churned customers. It was imbalanced, the proportion of churners was only 2.48% when compared to the majority class. In this research , the random under sampling and SMOTE techniques were used to handle class imbalance problem.

3.5 Model Selection

In this step, the pre-processed data obtained was used to build the machine learning model for predicting the bank customer churn. The objective of the research was to build supervised machine learning model to do the comparative

study and to find the best model for prediction. Decision Tree, Naïve Bayes and SVM were used to predict the bank customer churn.

3.6 Data Splits

For trustworthy and unbiased performance, a model needs to be evaluated on the unseen data which will determine how well a model has learned. The evaluation results will also arbitrate the model generalization and performance so that it is not over-fitting (Tanveer, 2019). In our study, for the unbiased validation of machine learning algorithm performance, the training and testing steps were performed on separated subsets of the data. In this research we used (1) training data 70% and testing data 30%, (2) training data 80% and testing data 20%, moreover we used k -fold cross validation with $k=5$.

3.7 Model Evaluation

Model evaluation is an integral part of data mining. For the classification problem, where data is highly imbalanced, confusion matrix, precision, recall and AUC are preferred evaluation metrics (Tanveer, 2019). In this research we used value of precision, recall, accuracy, false positive rate, false negative rate and F-score to evaluate the model.

CHAPTER IV

RESULTS AND DISCUSSION

This chapter presents the results of our experiments for the churn prediction of customers and discusses, the results of different machine learning methodologies, including, decision tree classifier, Naive Bayes classifier and SVM.

4.1 Dataset

The dataset has 8,166 instances, consisting of 203 instances in the churn class and 7,963 instances in the non churn class. We used 11 variables for building the model and the meaning of each variable is shown in 4.1, while the distribution of each variable is shown in figure 4.1-4.11.



มหาวิทยาลัยเทคโนโลยีสุรนารี

Table 4.1 Variable of customer data.

| Variable | Meaning | Type of variable |
|-----------------|---|--|
| Age | Age of the customer | continuous |
| Balance | Bank balance of the customer | continuous |
| CreditScore | Credit score of the customer | continuous |
| EstimatedSalary | Estimated salary of the customer in Dollars | continuous |
| Gender | Gender of the customer | nominal 0=female, 1=male |
| Geography | The country to which the customer belongs | nominal 0=France, 1=Spain, 2=Germany |
| HasCrCard | Binary Flag for the customer holds a credit card or not | nominal 0=no, 1=yes |
| IsActiveMember | Binary Flag for the customer is member or not | nominal 0=no, 1=yes |
| NumOfProducts | Number of bank products the customer is utilising | discrete |
| Tenure | Number of years for which the customer has been | discrete |
| Exited | Binary flag for the customer closed or retained | nominal 0=customer is retained 1=customer closed account |

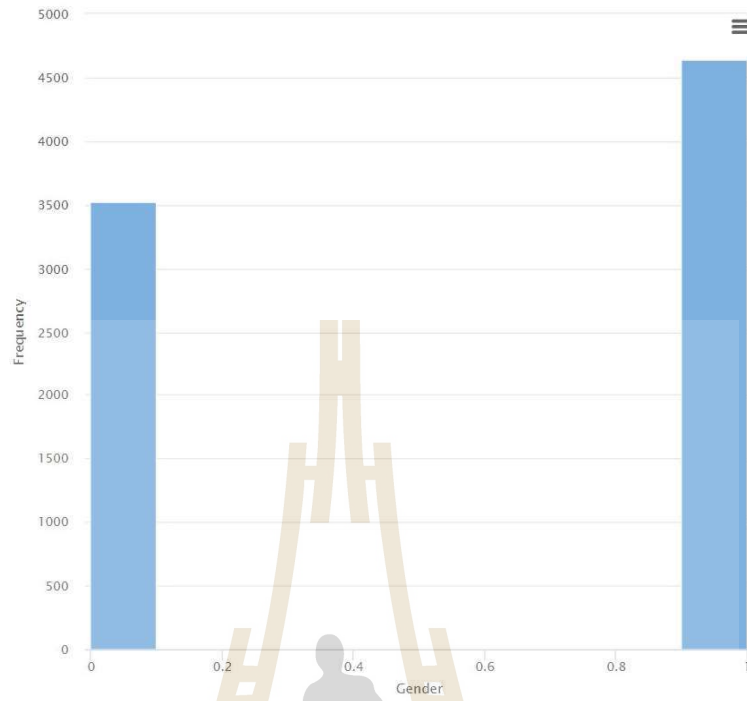


Figure 4.1 Gender variable distribution.

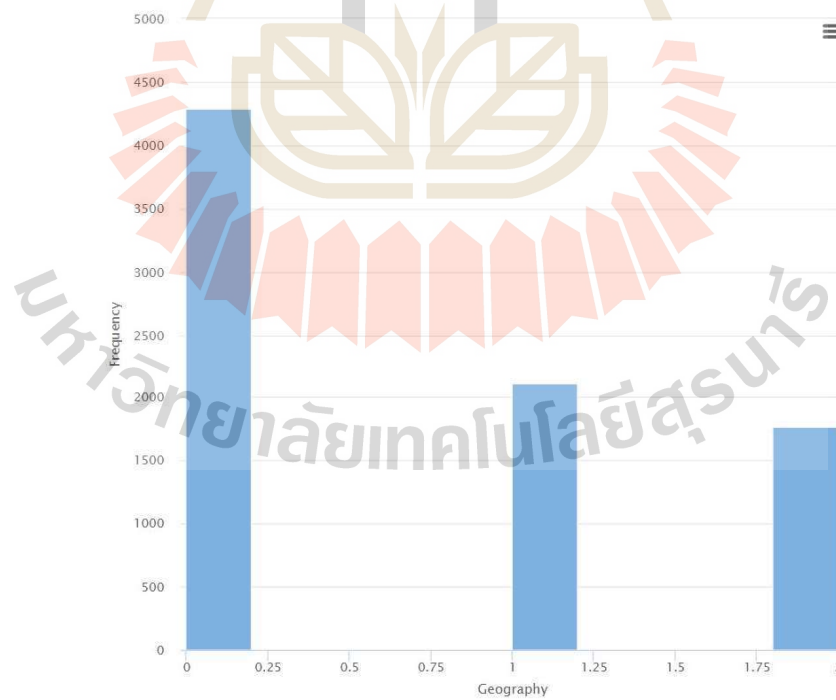


Figure 4.2 Geography variable distribution.

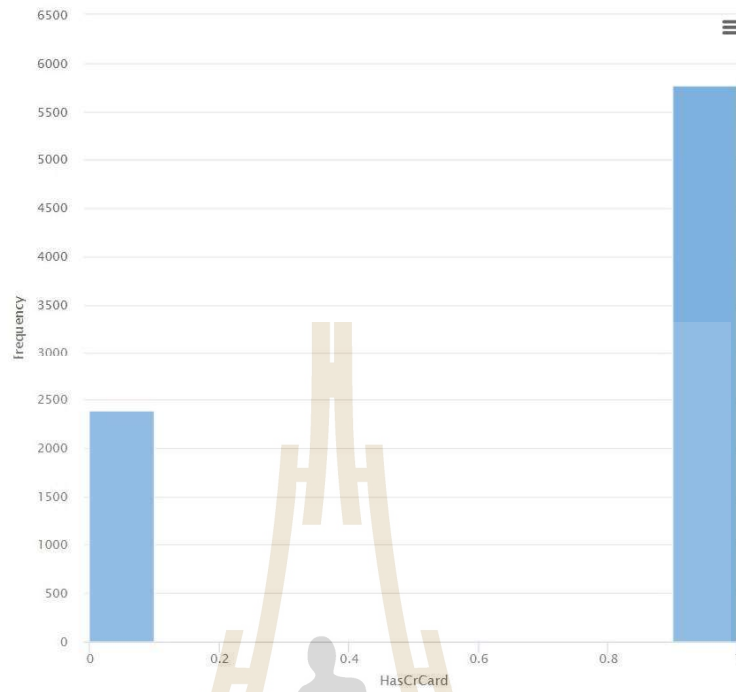


Figure 4.3 HasCrCard variable distribution.

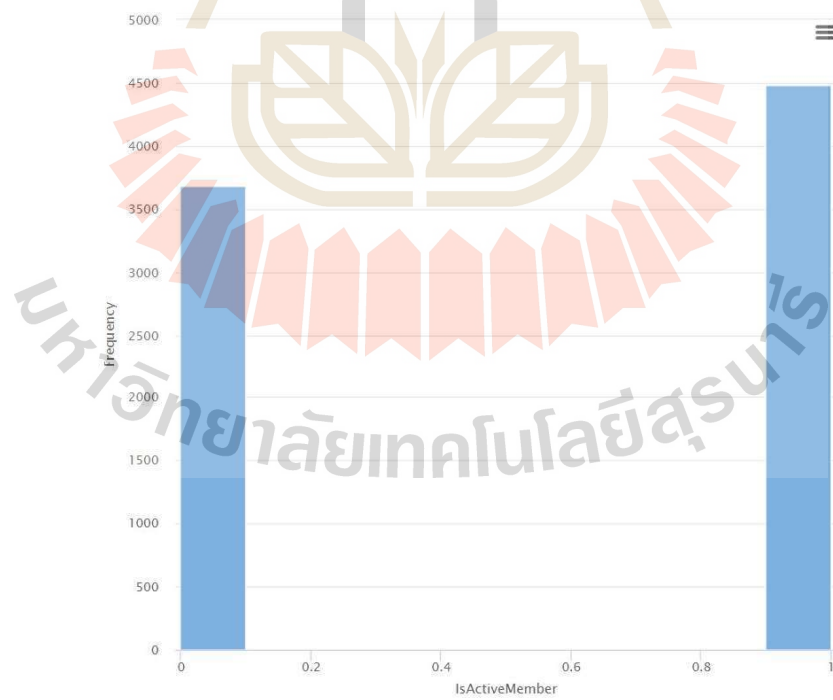


Figure 4.4 IsActiveMember variable distribution.

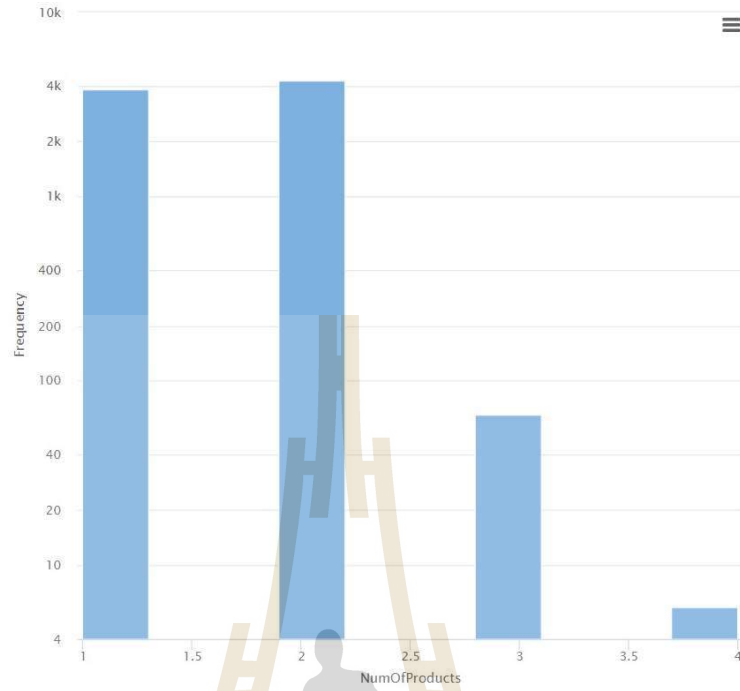


Figure 4.5 NumOfProducts variable distribution.

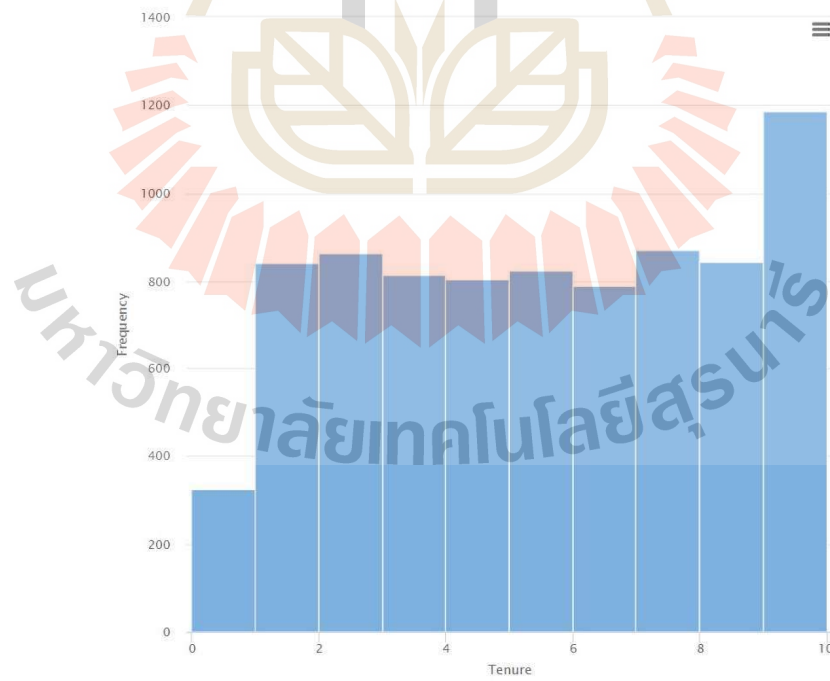


Figure 4.6 Tenure variable distribution.

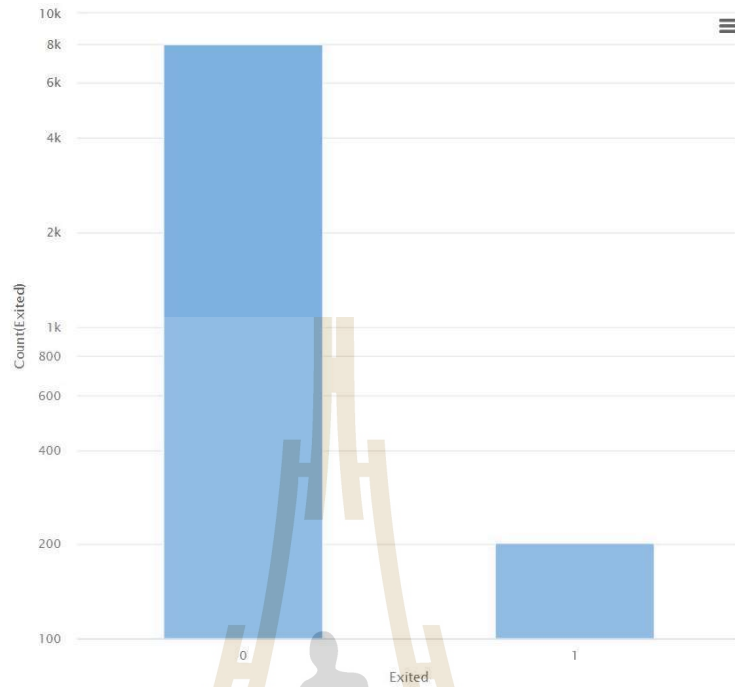


Figure 4.7 Exited variable distribution.

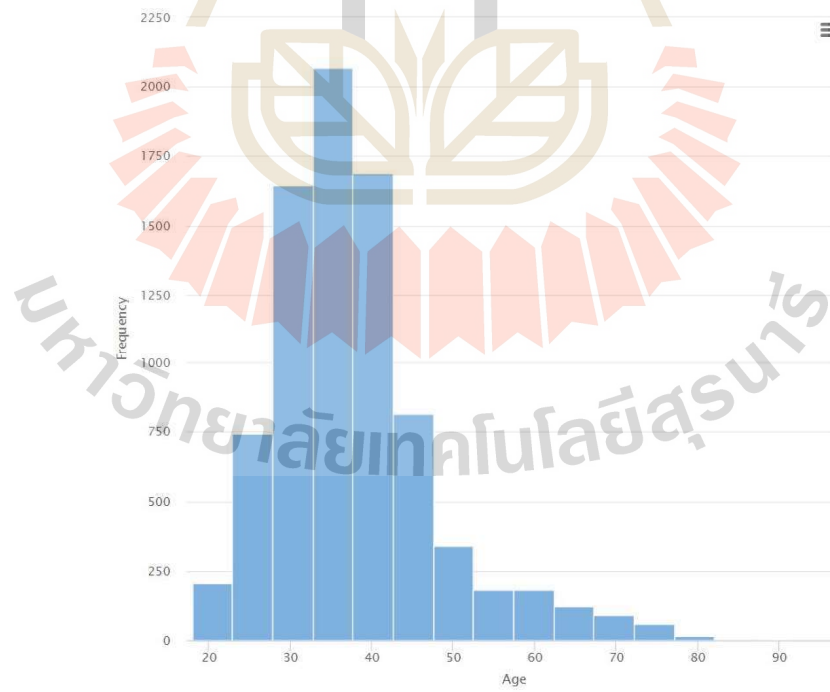


Figure 4.8 Age variable Histogram.

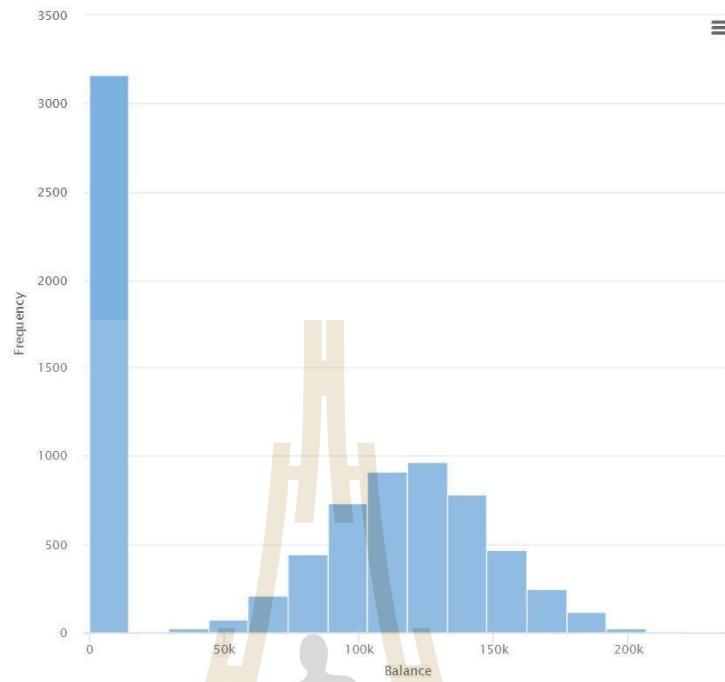


Figure 4.9 Balance variable Histogram.

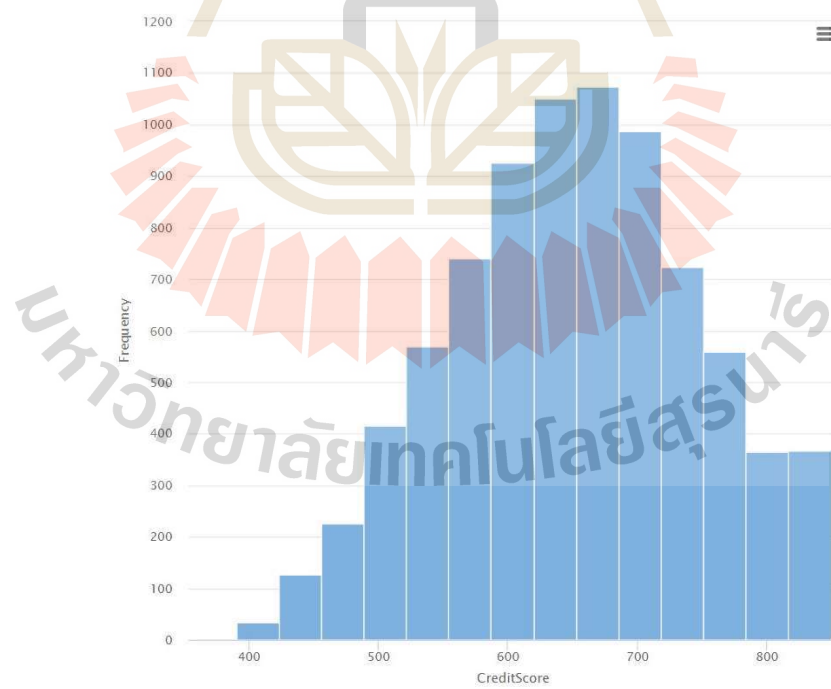


Figure 4.10 CreditScore variable Histogram.

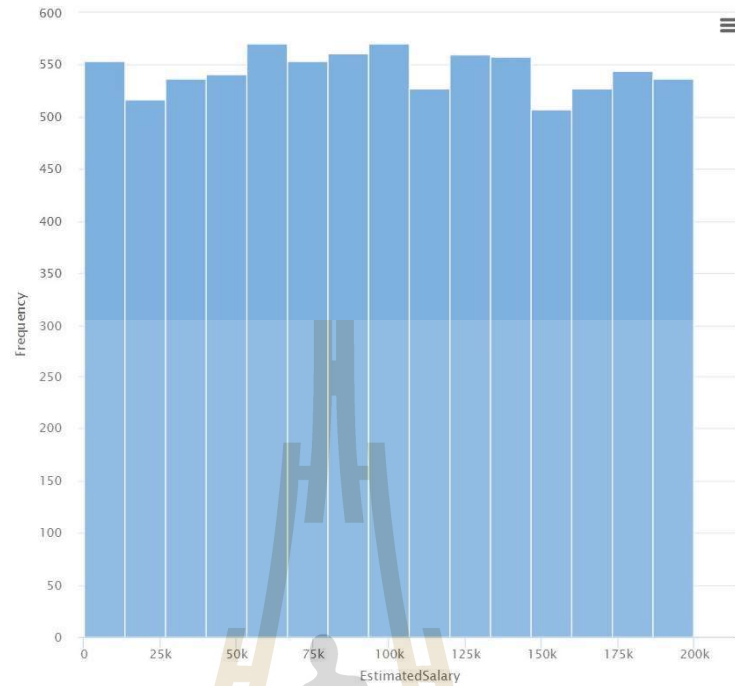


Figure 4.11 EstimatedSalary variable Histogram.

Table 4.2 Descriptive statistics of continuous data.

| Variable | Min | Max | Mean | Variance | Sd |
|-----------------|-------|-----------|-----------|----------|-----------|
| Age | 18 | 92 | 37.61 | 103.82 | 10.19 |
| Balance | 0 | 221,532.8 | 73,276.9 | 3.94E+09 | 62,803.46 |
| CreditScore | 358 | 850 | 651.67 | 9,176.67 | 95.79 |
| EstimatedSalary | 90.07 | 199,992.5 | 99,774.42 | 3.3E+09 | 57,443.71 |

4.2 Data Preparation

4.2.1 Normalization data

In this research, the independent variables namely Age, Balance, CreditScore and EstimatedSalary are continuous variable. Thus, we used min-max

normalization to transform the values of these variables to lie between 0 and 1.

4.2.2 Encoding

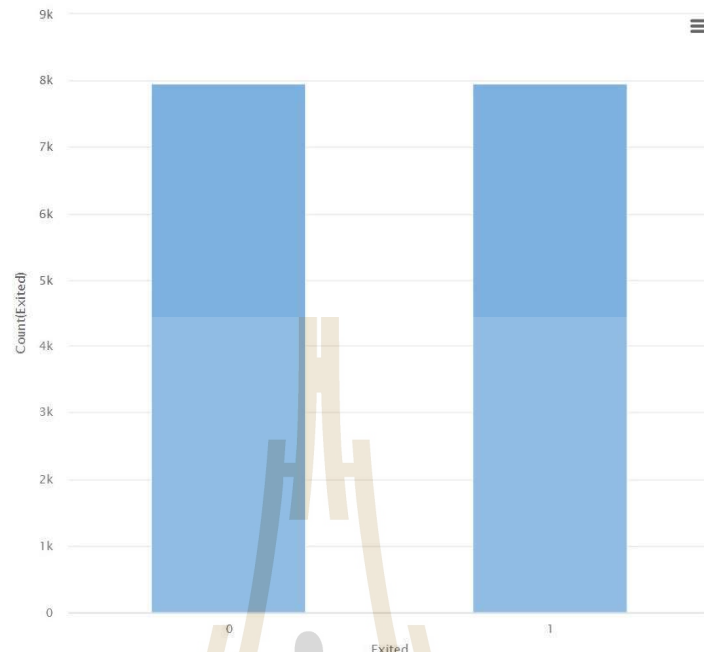
In this research, 5 categorical variable, including, Gender, Geography, HasCrCard, IsActiveMember and Exited, were encoded into numerical form.

Table 4.3 Description of 5 variable.

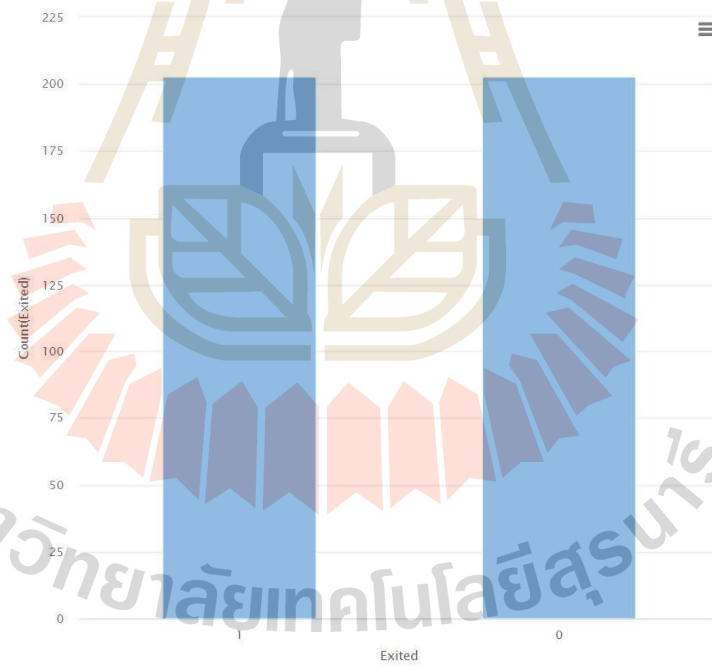
| Variable | Description | Type |
|----------------|---|---------|
| Gender | Gender of the customer | Integer |
| Geography | The country from which the customer belongs | Integer |
| HasCrCard | Binary Flag for the customer holds a credit card or not | Integer |
| IsActiveMember | Binary Flag for the customer is member or not | Integer |
| Exited | Binary flag for the customer closed or retained | Integer |

4.2.3 Sampling

The target variable 'Exited' is the class imbalance and its distribution is shown in figure 4.7.



(a) SMOTE.



(b) Random under sampling.

Figure 4.12 Exited variable distribution after used sampling.

4.2.4 Data splitting

The final dataset obtained after handling and normalizing dataset and data sampling. In this research, we split the final dataset with 2 proportions i.e. {70% training data, 30% test data} and {80% training data, 20% test data}. Moreover, we used k -fold cross validation with $k = 5$.

4.2.5 Feature selection

In this research we used the optimize weight function in Rapidminer Studio for finding the important features to improve the performance of the classification model (see Appendix A).

4.3 Modeling

The three proposed supervised machine learning models were created to test the best predicted bank customer churn.

4.3.1 Decision tree

The decision tree model was built using RapidMiner Studio version 9.6. The following parameters are considered: criterion, confidence, maximum depth of trees and minimal gain.

In this research, the "Gain Ratio" criterion was selected and the value of the other parameters were shown in Table 4.4-4.5.

The following results were obtained from the decision tree.

Table 4.4 Result of Decision tree model for a balanced dataset.

| Technique | Sampling | No. of Variable | Parameter | Precision | Recall | ACC | AUC | FPr | FNr | F. |
|-----------------------|--------------|-----------------|----------------------------|-----------|--------|-------|-------|-------|-------|-------|
| SMOTE | 70%, 30% | 6 | MD=21, MG=0, Confi=0.13 | 82.92 | 83.13 | 83.42 | 0.898 | 16.30 | 16.87 | 83.02 |
| | 80%, 20% | 6 | MD=31, MG=0, Confi=0.19 | 82.44 | 83.76 | 83.33 | 0.902 | 17.09 | 16.24 | 83.09 |
| | 5-fold cross | 7 | MD=19, MG=0, Confi=0.50 | 83.64 | 90.22 | 86.29 | 0.909 | 17.64 | 9.78 | 86.80 |
| Random under sampling | 70%, 30% | 5 | MD=11, MG=0, Confi=0.07 | 73.68 | 86.15 | 76.23 | 0.801 | 35.02 | 13.85 | 79.43 |
| | 80%, 20% | 7 | MD=31, MG=0, Confi=0.19 | 70.45 | 79.49 | 74.07 | 0.777 | 30.95 | 20.51 | 74.70 |
| | 5-fold cross | 5 | MD=11, MG=0, Confi=0.50 | 76.50 | 75.37 | 76.11 | 0.778 | 23.15 | 24.63 | 75.93 |

Remark. FPr is false positive rate, FNr is false negative rate and F is F-measure. Values of Precision, Recall, Accuracy, FPr and FNr are presented in percentage (%). MD is maximum depth, MG is minimum gain and Confi is confidence.

Table 4.5 Result of Decision tree model for imbalance dataset.

| Sampling | No. of Variable | Parameter | Precision | Recall | ACC | AUC | FPr | FNr | F. |
|--------------|-----------------|---------------------------|-----------|--------|-------|-------|------|-------|-------|
| 70%, 30% | 6 | MD=6, MG=0, Confi=0.04 | 57.14 | 6.25 | 97.43 | 0.693 | 0.13 | 93.75 | 11.27 |
| 80%, 20% | 6 | MD=6, MG=0, Confi=0.04 | 66.67 | 4.65 | 97.43 | 0.555 | 0.06 | 95.35 | 8.63 |
| 5-fold cross | 5 | MD=5, MG=0, Confi=0.01 | 60.71 | 8.37 | 97.59 | 0.552 | 0.14 | 91.63 | 14.71 |

Remark. FPr is false positive rate, FNr is false negative rate and F is F-measure. Values of Precision, Recall, Accuracy, FPr and FNr are presented in percentage (%). MD is maximum depth, MG is minimum gain and Confi is confident.

4.3.2 Naïve Bayes

The Naïve Bayes model was built using RapidMiner Studio version 9.6.

The following results were obtained from the Naïve Bayes.

Table 4.6 Result of Naïve Bayes model for balanced dataset.

| Technique | Sampling | No. of Variable | Precision | Recall | ACC | AUC | FPr | FNr | F. |
|-----------------------|--------------|-----------------|-----------|--------|-------|-------|-------|-------|-------|
| SMOTE | 70%, 30% | 9 | 76.40 | 79.36 | 77.98 | 0.847 | 23.33 | 20.64 | 77.85 |
| | 80%, 20% | 9 | 76.37 | 78.63 | 77.65 | 0.847 | 23.29 | 21.37 | 77.48 |
| | 5-fold cross | 9 | 77.10 | 78.14 | 77.46 | 0.848 | 23.21 | 21.86 | 77.62 |
| Random under sampling | 70%, 30% | 6 | 88.46 | 70.77 | 79.51 | 0.866 | 10.53 | 29.23 | 78.63 |
| | 80%, 20% | 7 | 81.58 | 78.63 | 77.65 | 0.847 | 23.29 | 20.51 | 80.52 |
| | 5-fold cross | 7 | 75.65 | 71.92 | 74.39 | 0.802 | 23.15 | 28.08 | 73.74 |

Remark. FPr is false positive rate, FNr is false negative rate and F is F-measure.

Values of Precision, Recall, Accuracy, FPr and FNr are presented in percentage (%).

Table 4.7 Result of Naïve Bayes model for imbalanced dataset.

| Sampling | No. of Variable | Precision | Recall | ACC | AUC | FPr | FNr | F. |
|--------------|-----------------|-----------|--------|-------|-------|-----|-------|-------|
| 70%, 30% | 9 | 100 | 7.81 | 97.59 | 0.843 | 0 | 92.19 | 14.49 |
| 80%, 20% | 9 | 0 | 0 | 97.37 | 0.820 | 0 | 100 | - |
| 5-fold cross | 9 | 100 | 4.43 | 97.62 | 0.800 | 0 | 95.57 | 8.48 |

Remark. FP is false positive rate, FN is false negative rate and F is F-measure. Values of Precision, Recall, Accuracy, FP and FN are presented in percentage (%).

4.3.3 Support vector machine

The SVM model was built using RapidMiner Studio version 9.6. In this research for SVM model 'radial basis function(RBF)' kernel was selected, because the dataset in this study is significant, so only kernel is computationally practical. The following parameters were optimized; kernel gamma(γ), penalty parameter(C) and epsilon(ξ).

The following results were obtained from the SVM.

Table 4.8 Result of Support vector machine model for balanced dataset.

| Technique | Sampling | No. of Variable | Parameter | Precision | Recall | ACC | AUC | FPr | FNr | F. |
|-----------------------|--------------|-----------------|--------------------------|-----------|--------|-------|-------|-------|-------|-------|
| SMOTE | 70%, 30% | 9 | GM=2, C=1, Epsilon=0 | 89.92 | 91.55 | 90.87 | 0.961 | 9.76 | 8.45 | 90.73 |
| | 80%, 20% | 9 | GM=2, C=1.5, Epsilon=0 | 89.55 | 92.36 | 90.99 | 0.964 | 10.33 | 7.64 | 90.93 |
| | 5-fold cross | 9 | GM=2, C=2.3, Epsilon=0 | 89.81 | 92.99 | 91.22 | 0.961 | 10.55 | 7.01 | 91.37 |
| Random under sampling | 70%, 30% | 6 | GM=1, C=0.5, Epsilon=0 | 85.48 | 81.54 | 82.79 | 0.840 | 15.79 | 18.46 | 83.46 |
| | 80%, 20% | 7 | GM=1, C=1, Epsilon=0 | 84.62 | 84.62 | 85.19 | 0.833 | 14.29 | 15.38 | 84.62 |
| | 5-fold cross | 7 | GM=1, C=1.1, Epsilon=0.6 | 81.48 | 75.86 | 79.31 | 0.831 | 17.24 | 24.14 | 78.60 |

Remark. FPr is false positive rate, FNr is false negative rate and F is F-measure. Values of Precision, Recall, Accuracy, FPr and FNr are presented in percentage (%). GM is kernel gamma.

Table 4.9 Result of Support vector machine model for imbalance dataset.

| Sampling | No. of Variable | Parameter | Precision | Recall | ACC | AUC | FPr | FNr | F. |
|--------------|-----------------|------------------------|-----------|--------|-------|-------|------|-------|-------|
| 70%, 30% | 3 | GM=1, C=0.5, Epsilon=0 | 100 | 6.25 | 97.55 | 0.540 | 0 | 93.75 | 11.76 |
| 80%, 20% | 4 | GM=1, C=1.5, Epsilon=0 | 100 | 9.30 | 97.61 | 0.546 | 0 | 90.70 | 17.02 |
| 5-fold cross | 3 | GM=2, C=4.6, Epsilon=0 | 75.65 | 71.92 | 97.81 | 0.668 | 0.08 | 85.22 | 73.74 |

Remark. FPr is false positive rate, FNr is false negative rate and F is F-measure. Values of Precision, Recall, Accuracy, FPr and FNr are presented in percentage (%). GM is kernel gamma.

The confusion matrix was usually used to determine the performance of the model and also to calculate the accuracy, recall, precision, F-measure, false positive rate and false negative rate measure of the model in the classification problem. It was shown in Appendix B.

4.4 Performance Evaluation

This section evaluates the prediction power of the classification models built in 4.3.1-4.3. Table 4.10, 4.11 and 4.12 shows the precision, recall, AUC, FN rate, and F-measure to each models.

Table 4.10 Evaluation results for training 70% of dataset, test 30% of dataset.

| Sampling | Model | Precision | Recall | F-score | AUC | FN rate |
|----------|--------------|-----------|--------------|--------------|--------------|-------------|
| | Original DT | 57.14 | 6.25 | 11.27 | 0.693 | 93.75 |
| | Original NB | 100 | 7.81 | 14.49 | 0.843 | 92.19 |
| | Original SVM | 100 | 6.25 | 11.76 | 0.540 | 93.75 |
| | SMOTE DT | 82.92 | 83.13 | 83.02 | 0.898 | 16.87 |
| 70%, 30% | SMOTE NB | 76.40 | 79.36 | 77.85 | 0.847 | 20.64 |
| | SMOTE SVM | 89.92 | 91.55 | 90.73 | 0.961 | 8.45 |
| | RD under DT | 73.68 | 86.15 | 79.43 | 0.801 | 13.85 |
| | RD under NB | 88.46 | 70.77 | 78.63 | 0.866 | 29.23 |
| | RD under SVM | 85.48 | 81.54 | 83.46 | 0.840 | 18.46 |

Remark. RD under is random under sampling and values of Precision, Recall and FN rate are presented in percentage (%).

Table 4.11 Evaluation results for training 80% of dataset, test 20% of dataset.

| Sampling | Model | Precision | Recall | F-score | AUC | FN rate |
|----------|--------------|-----------|--------------|--------------|--------------|-------------|
| | Original DT | 66.67 | 4.65 | 8.63 | 0.555 | 95.35 |
| | Original NB | 0 | 0 | - | 0.820 | 100 |
| | Original SVM | 100 | 9.30 | 17.02 | 0.546 | 90.70 |
| | SMOTE DT | 82.44 | 83.76 | 83.09 | 0.902 | 16.24 |
| 80%, 20% | SMOTE NB | 76.37 | 78.63 | 77.48 | 0.847 | 21.37 |
| | SMOTE SVM | 89.55 | 92.36 | 90.93 | 0.964 | 7.64 |
| | RD under DT | 70.45 | 79.49 | 74.70 | 0.777 | 20.51 |
| | RD under NB | 81.58 | 78.63 | 80.52 | 0.847 | 20.51 |
| | RD under SVM | 84.62 | 84.62 | 84.62 | 0.833 | 24.14 |

Remark. RD under is random under sampling and values of Precision, Recall and FN rate are presented in percentage (%).

Table 4.12 Evaluation results for 5-fold cross validation.

| Sampling | Model | Precision | Recall | F-score | AUC | FN rate |
|--------------|--------------|-----------|--------------|--------------|--------------|-------------|
| | Original DT | 60.71 | 8.37 | 14.71 | 0.552 | 91.63 |
| | Original NB | 100 | 4.43 | 8.48 | 0.8 | 95.57 |
| | Original SVM | 75.65 | 71.62 | 73.74 | 0.668 | 85.22 |
| | SMOTE DT | 83.64 | 90.22 | 86.80 | 0.909 | 9.78 |
| 5-fold cross | SMOTE NB | 77.10 | 78.14 | 77.62 | 0.848 | 21.86 |
| | SMOTE SVM | 89.81 | 92.99 | 91.37 | 0.961 | 7.01 |
| | RD under DT | 76.50 | 75.37 | 75.93 | 0.778 | 24.63. |
| | RD under NB | 75.65 | 71.92 | 80.52 | 0.847 | 20.51 |
| | RD under SVM | 81.48 | 75.86 | 78.60 | 0.831 | 24.14 |

Remark. RD under is random under sampling and values of Precision, Recall and FN rate are presented in percentage (%).

Figure 4.13 illustrates the ROC curves for the nine prediction models. (in case 5-fold cross validation) , figure 4.14 the ROC curves in case of split validation 70% of training dataset and figure 4.15 the ROC curves in case of split validation 80% of training dataset.

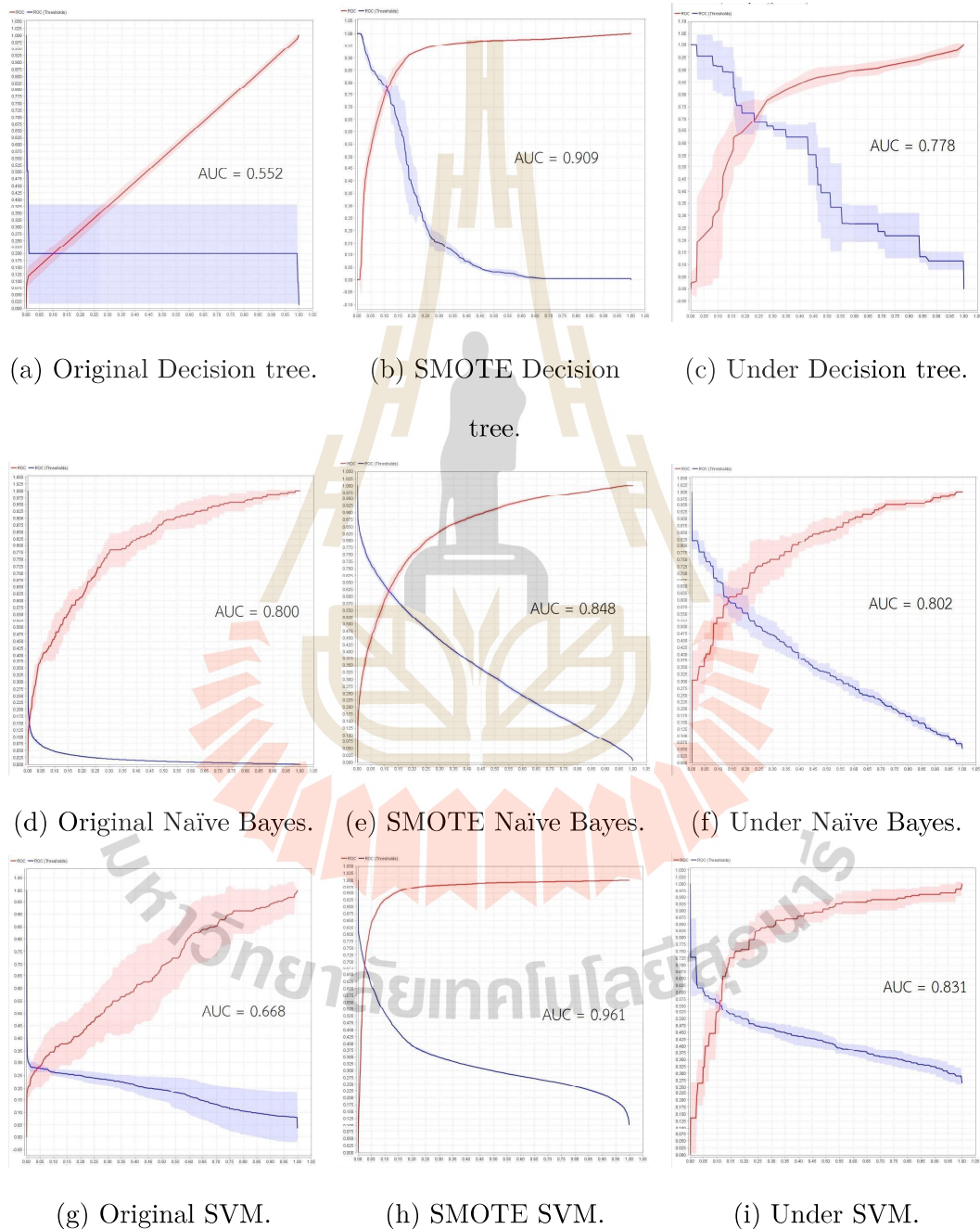
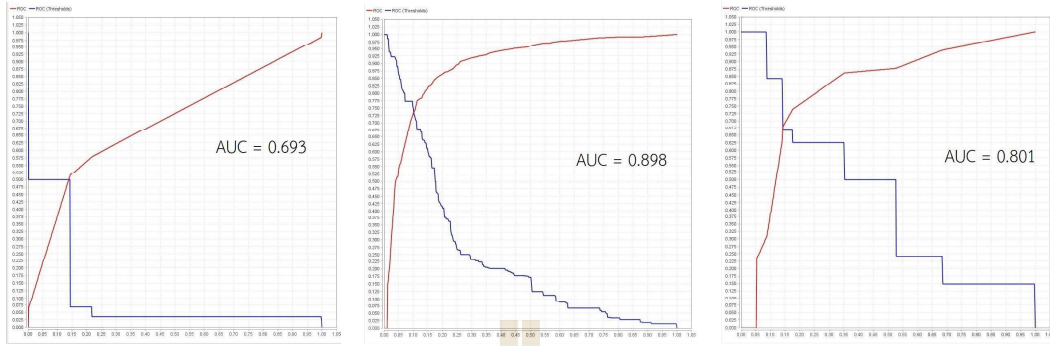
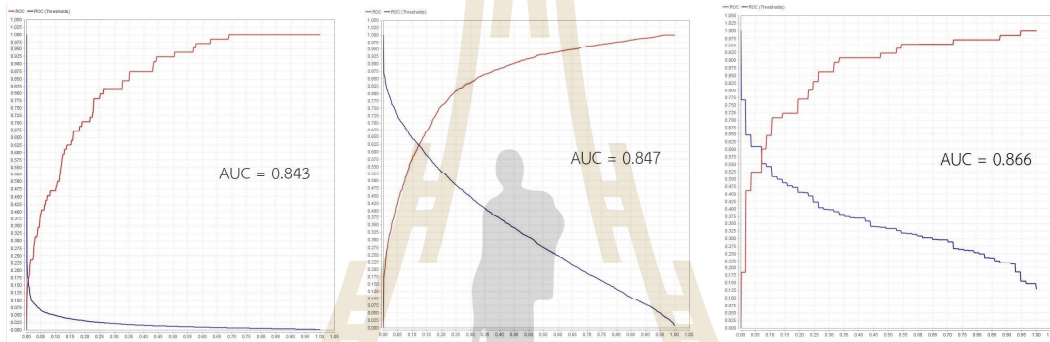


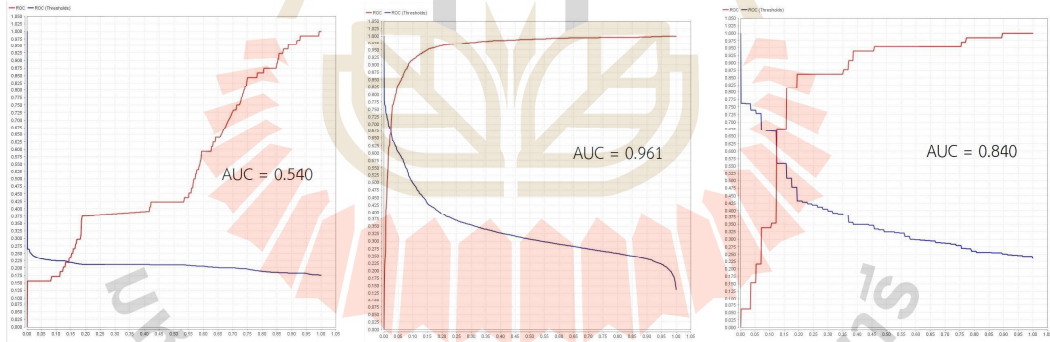
Figure 4.13 ROC curves of model 5-fold cross validation.



(a) Original Decision tree. (b) SMOTE Decision tree. (c) Under Decision tree.



(d) Original Naïve Bayes. (e) SMOTE Naïve Bayes. (f) Under Naïve Bayes.



(g) Original SVM. (h) SMOTE SVM. (i) Under SVM.

Figure 4.14 ROC curves of model split validation 70%.

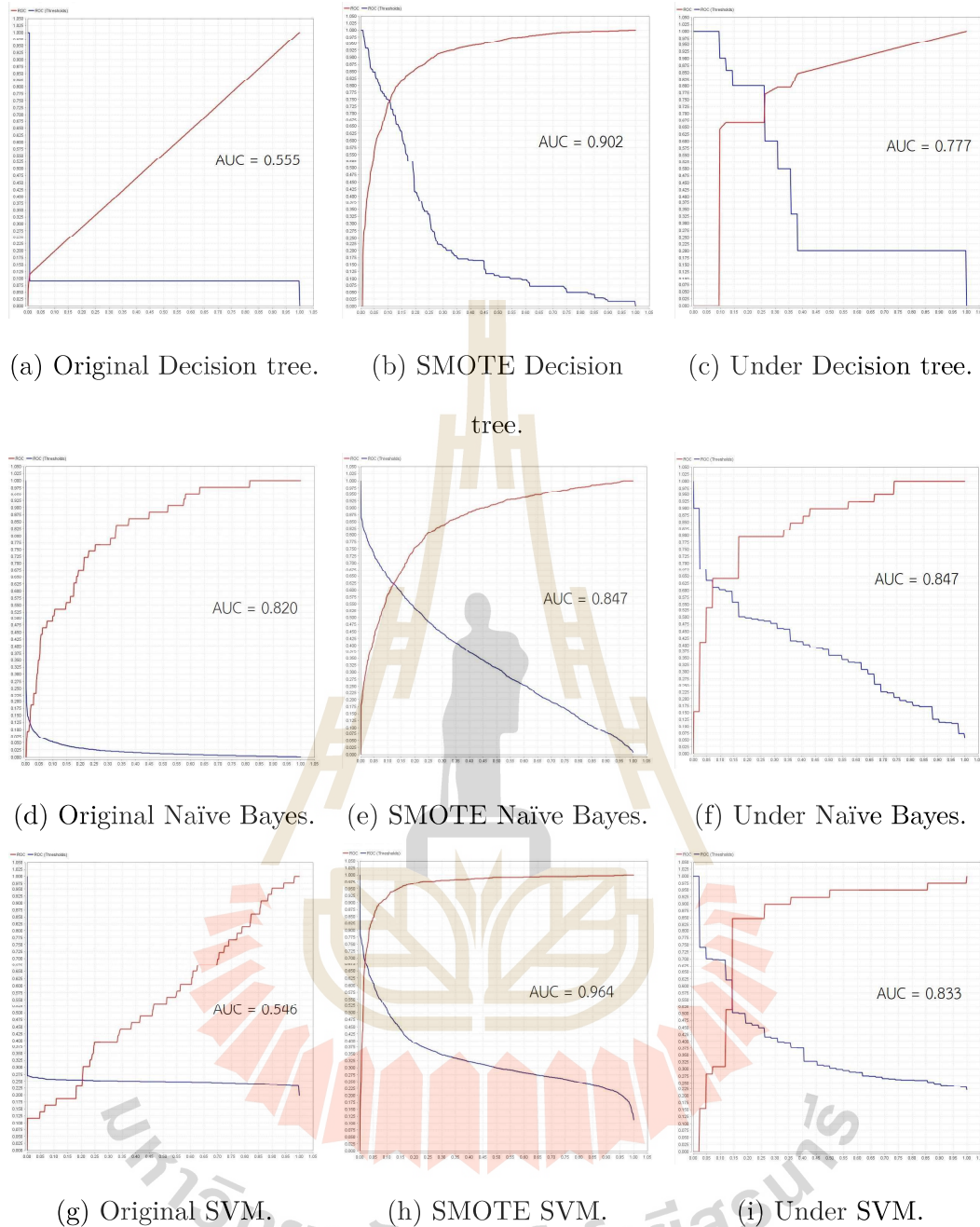


Figure 4.15 ROC curves of model split validation 80%.

All the result show the performance of model, by the splitting of 70% training data: 30% test data, the best model is support vector machine with SMOTE provided recall = 91.55%, AUC = 96.1%, F-score = 90.73% and false negative rate = 8.45%. By the splitting of 80% training data: 20% test data,

the best model is support vector machine with SMOTE provided recall = 92.36%, AUC = 96.4%, F-score = 90.93% and false negative rate = 7.64%. Also, for 5-fold cross validation, the best model is support vector machine with SMOTE provided recall = 92.99%, AUC = 96.1%, F-score = 91.37% and false negative rate = 7.01%. But for all model with original data give the performance less than the model with sampling technique.



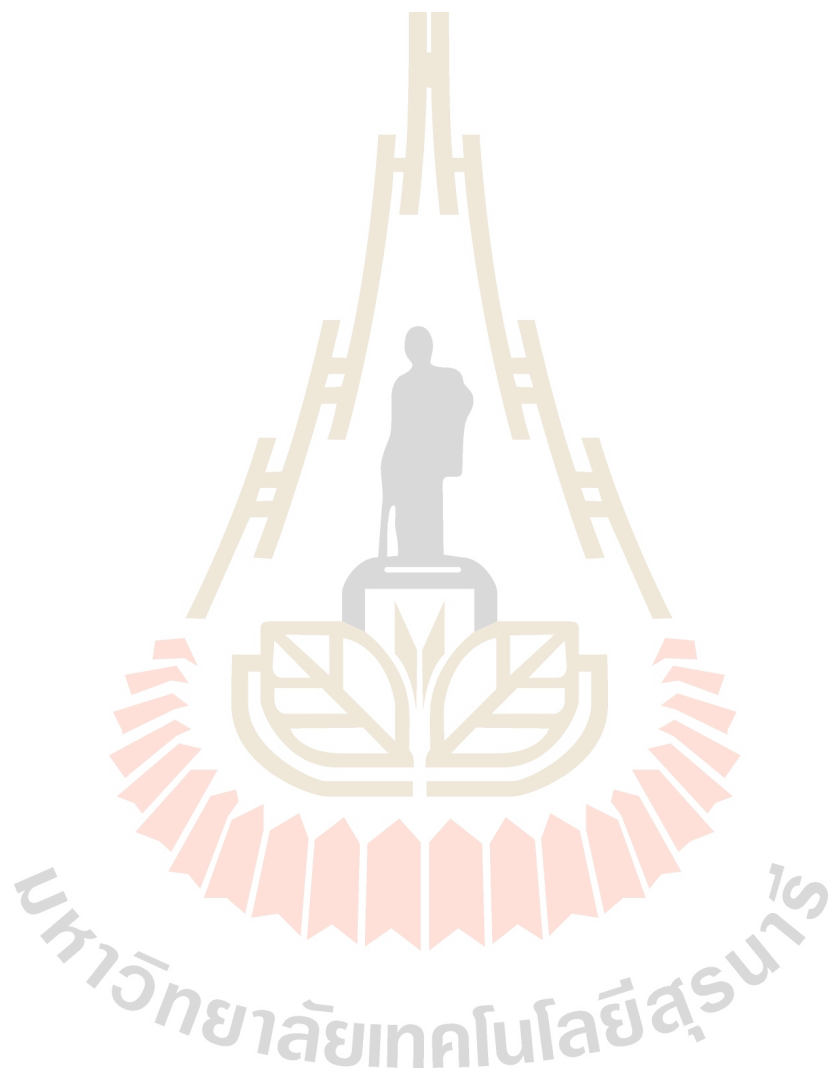
CHAPTER V

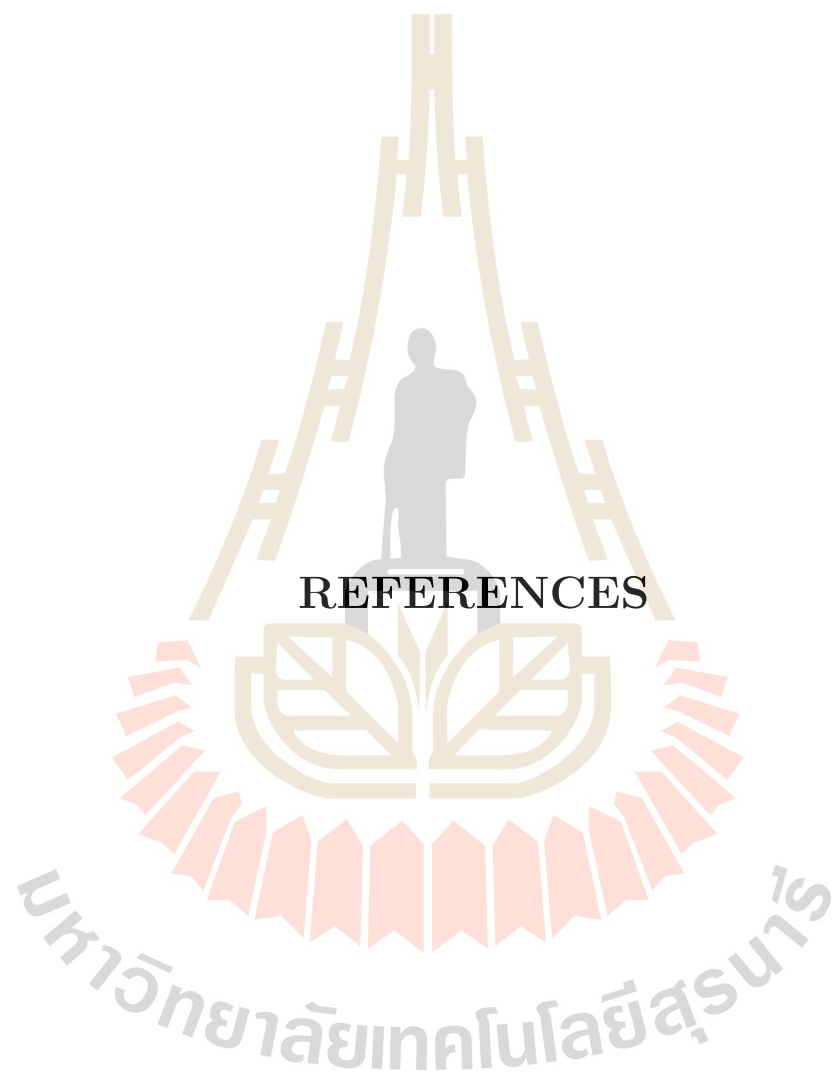
CONCLUSION AND RECOMMENDATION

This thesis studied the churn prediction modelling of bank customers, where the dataset was obtained from Kaggle as mentioned before. Since the ratio of closed status versus retained customer is 203:7963, then the modelling must be concerned about the significance of that different number. The unbalance of the data effects to the validity in modelling even the model has high accuracy. It was found that if one ignores the group of data which has a small size, the recall and AUC rate are low but the false negative rate is high. Therefore, imbalance data techniques are employed in this thesis to improve the efficiency before creating a churn prediction model.

Imbalance data techniques used in this thesis were SMOTE and random under sampling. After that, we applied 3 types of the churn prediction modelling techniques to the modified data, which were decision tree classifier, Naïve Bayes classifier and SVM. The process in all techniques considers how each variable affects to the model significantly and finds the appropriate parameters for each model. The study showed that applying Imbalance data techniques, both SMOTE and random under sampling, provided better results in modelling than one without using those techniques. The accuracy measure was considered as the evaluation metrics to get the best modal. In this research, it was found that the support vector machine technique with balanced dataset using SMOTE technique outperformed the other algorithms in predicting the customer churn for bank customers dataset. For the future work, this research may be extended by developing the sampling

technique for the other imbalance problem and improving the model for the churn prediction problem.





REFERENCES

REFERENCES

- Amin, A., Anwar, S., Adnan, A., Nawaz, M., Howard, N., Qadir, J., Hawalah, A., and Hussain, A. (2016). Comparing oversampling techniques to handle the class imbalance problem: A customer churn prediction case study. **IEEE Access**. 4: 7940-7957.
- Bishop, C. M. (2006), **Pattern Recognition and Machine Learning**. springer.
- Bolton, C. (2010). **Logistic Regression and its Application in Credit Scoring**. University of Pretoria.
- Brown, I., and Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. **Expert Systems with Applications**. 39(3): 3446-3453.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**. 16: 321-357.
- Crone, S. F., and Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. **International Journal of Forecasting**. 28(1): 224-238.
- Expert System Team. (March 2017). What is Machine Learning? A definition. Retrieved from <https://expertsystem.com/machine-learning-definition/>.
- Fabris, F., De Magalhães, J. P., and Freitas, A. A. (2017). A review of supervised machine learning applied to ageing research. **Biogerontology**. 18(2): 171-188.

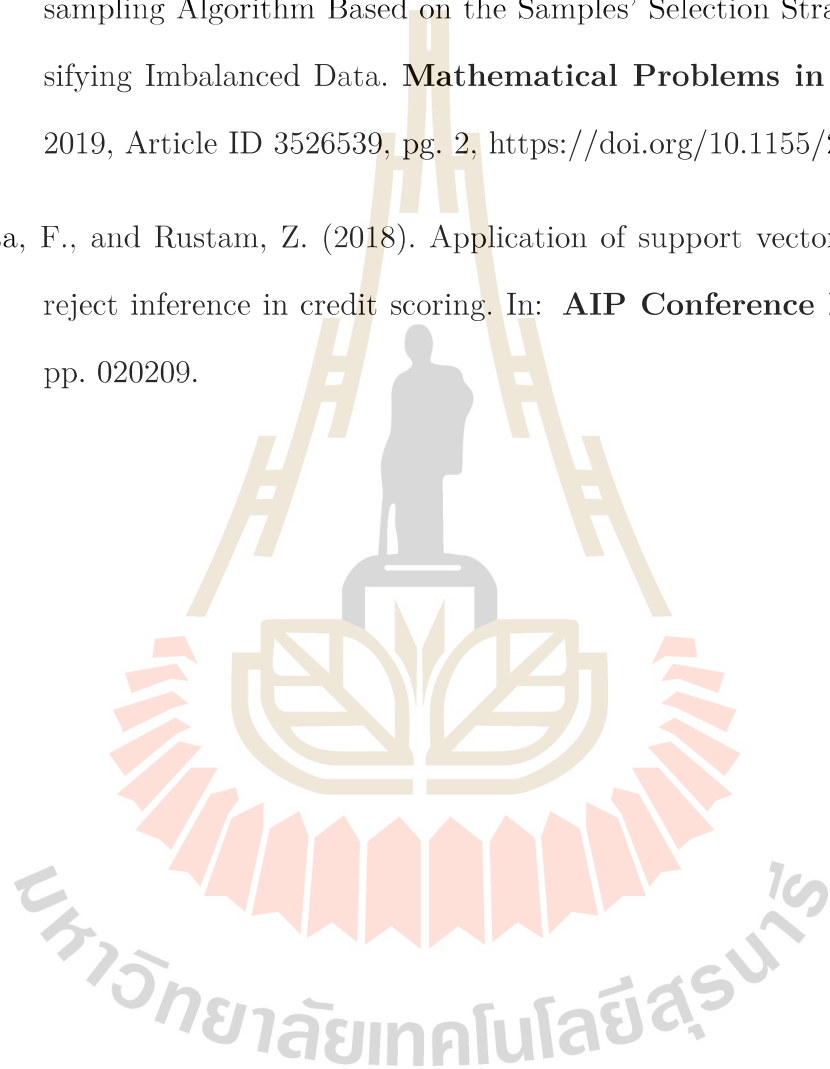
- G. Ying. (2017). Using Decision Tree to Analyze the Turnover of Employees.
- Haltuf, M. (2014). Support vector machines for credit scoring. **University of Economics in Prague Faculty of Finance.**
- Han, J., Pei, J., and Kamber, M. (2011). **Data Mining: Concepts and Techniques.** Third Edition. Elsevier.
- Hanif, A., and Azhar, N. (2017). Resolving class imbalance and feature selection in customer churn dataset. In: **the 2017 International Conference on Frontiers of Information Technology (FIT).** pp. 82-86.
- He, H., and Ma, Y. (2013). **IMBALANCED LEARNING : Foundations, Algorithms, and Applications.** John Wiley & Sons.
- Burez, J., and Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. **Expert Systems with Applications.** 36(3): 4626-4636.
- Kennedy, K. (2013). Credit scoring using machine learning
- Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. (2006). Handling imbalanced datasets: A review. **GESTS International Transactions on Computer Science and Engineering.** 30(1): 25-36.
- Kumar A. (2018). Machine Learning: Validation Techniques. Retrieved from <https://dzone.com/articles/machine-learning-validation-techniques>.
- Larose, D. T., and Larose, C. D. (2014). **Discovering Knowledge in Data: An Introduction to Data Mining.** John Wiley & Sons.
- Mitchell, T. M. (1997). **Machine learning.** McGraw-hill New York.

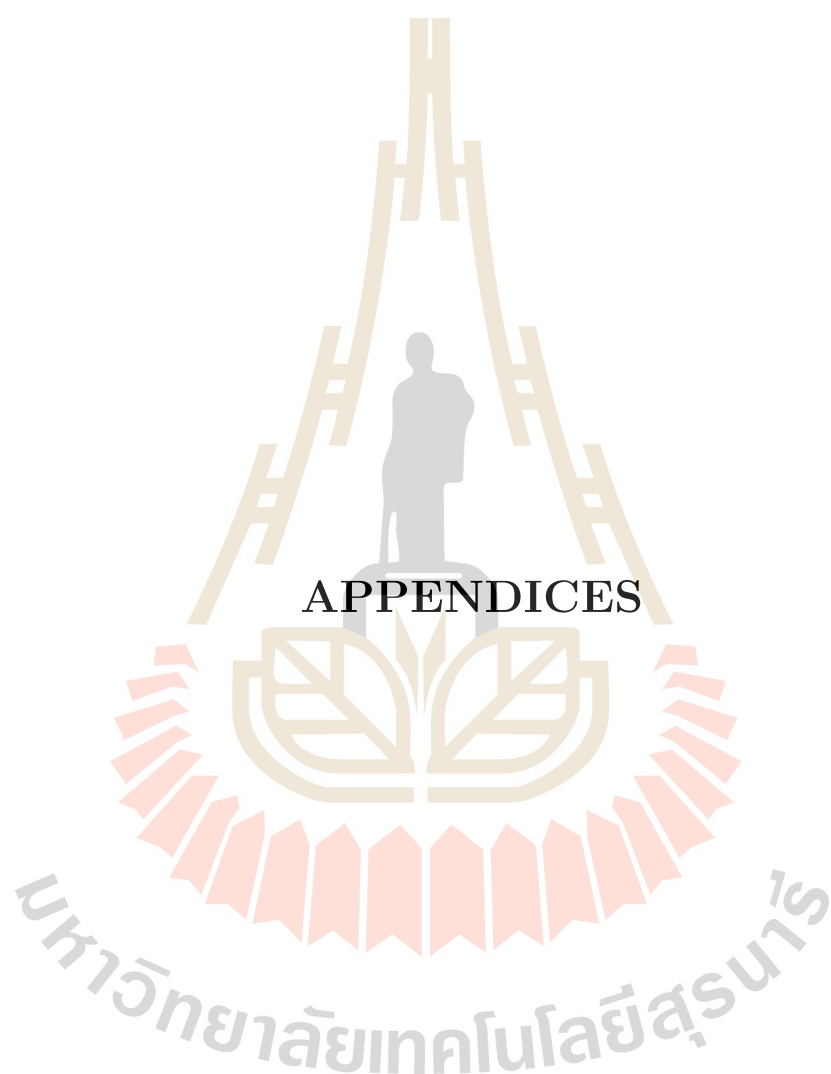
- Narkhede S. (2016). Understanding AUC - ROC Curve. Retrieved from <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>.
- Pantazi, X.-E., Moshou, D., and Bochtis, D. (2019). **Intelligent Data Mining and Fusion Systems in Agriculture**. Academic Press.
- Quinlan, J. R. (1986). Induction of decision trees. **Machine Learning**. 1(1): 81-106.
- Quinlan, J. R. (2014). **C4. 5: Programs for Machine Learning**. Elsevier.
- Ramyachitra, D., and Manikandan, P. (2014). Imbalanced dataset classification and solutions: a review, **International Journal of Computing and Business Research (IJCBR)**. 5(4).
- Russell, S., and Norvig, P. (2002). Artificial intelligence: a modern approach.
- Suksut K. (2016). Imbalance Data Classification Using Data Improvement and Parameter Optimization with Restarting Genetic Algorithm. Ph.D thesis, Suranaree University of Technology, Nakhonratchasima
- Sonak, A., and Patankar, R. (2015). A survey on methods to handle imbalance dataset. **Int. J. Comput. Sci. Mob. Comput.** 4(11): 338-343.
- Tanveer, A. (2019). **Churn Prediction Using Customers' Implicit Behavioral Patterns and Deep Learning**.
- Wadikar, D. (2020). **Customer Churn Prediction(Masters Dissertation)**. Technological University Dublin.

W. Kesornsit, V. Lorchirachoonkul, J. Jitthavech. (2018). Imbalanced Data Problem Solving in Classification of Diabetes Patients. **kku research journal**. 18(3).

Xie, W., Liang, G., Dong, Z., Tan, B., and Zhang, B. (2019). An Improved Oversampling Algorithm Based on the Samples' Selection Strategy for Classifying Imbalanced Data. **Mathematical Problems in Engineering** 2019, Article ID 3526539, pg. 2, <https://doi.org/10.1155/2019/3526539>.

Yaurita, F., and Rustam, Z. (2018). Application of support vector machines for reject inference in credit scoring. In: **AIP Conference Proceedings**. pp. 020209.





APPENDICES

มหาวิทยาลัยเทคโนโลยีสุรนารี

The logo of Sakon Nakhon Rajabhat University is a large, faint watermark in the background. It features a central figure of a person standing on a pedestal, surrounded by a stylized architectural structure with multiple levels of arches. Below the figure is a circular emblem with a book and a sunburst. The entire logo is rendered in a light beige color.

APPENDIX A
FEATURE SELECTION OF THE MODEL

มหาวิทยาลัยเทคโนโลยีสุรนารี

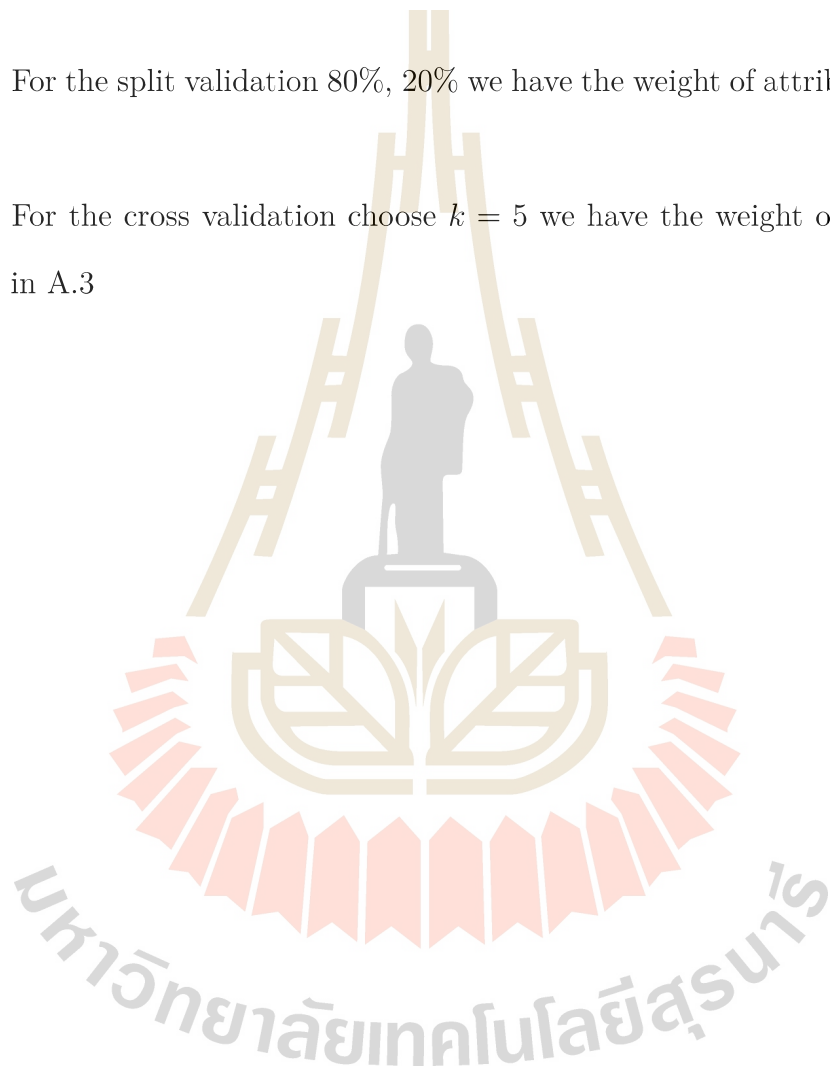
A.1 The weight of attribute after used feature selection

In this chapter we show weight by used optimize weight function in Rapidminer Studio program.

For the split validation 70%, 30% we have the weight of attributes as shown in A.1

For the split validation 80%, 20% we have the weight of attributes as shown in A.2

For the cross validation choose $k = 5$ we have the weight of attributes as shown in A.3



| Attribute | Weight |
|-----------------|--------|
| Age | 0 |
| Balance | 1 |
| CreditScore | 0 |
| EstimatedSalary | 0 |
| Gender | 0 |
| Geography | 1 |
| HasCrCard | 0 |
| IsActiveMember | 0 |
| NumOfProducts | 0.061 |
| Tenure | 0 |

(a) Weight svm original data.

| Attribute | Weight |
|-----------------|--------|
| Age | 1 |
| Balance | 1 |
| CreditScore | 0 |
| EstimatedSalary | 0 |
| Gender | 1 |
| Geography | 1 |
| HasCrCard | 0 |
| IsActiveMember | 1 |
| NumOfProducts | 1 |
| Tenure | 0 |

(b) Weight decision tree original data.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.665 |
| Balance | 0.619 |
| CreditScore | 0.209 |
| EstimatedSalary | 0.367 |
| Gender | 1 |
| Geography | 0 |
| HasCrCard | 0.296 |
| IsActiveMember | 0.602 |
| NumOfProducts | 0.998 |
| Tenure | 0.552 |

(c) Weight naive Bayes original data.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.780 |
| Balance | 0.689 |
| CreditScore | 0.622 |
| EstimatedSalary | 0.866 |
| Gender | 1 |
| Geography | 0.749 |
| HasCrCard | 0.970 |
| IsActiveMember | 0.072 |
| NumOfProducts | 0 |
| Tenure | 0.373 |

(d) Weight svm over sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.763 |
| Balance | 0 |
| CreditScore | 0 |
| EstimatedSalary | 0 |
| Gender | 1 |
| Geography | 1 |
| HasCrCard | 1 |
| IsActiveMember | 1 |
| NumOfProducts | 0.061 |
| Tenure | 0 |

(e) Weight decision tree over sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.794 |
| Balance | 0 |
| CreditScore | 0.646 |
| EstimatedSalary | 0.875 |
| Gender | 1 |
| Geography | 0.765 |
| HasCrCard | 0.972 |
| IsActiveMember | 0.131 |
| NumOfProducts | 0.064 |
| Tenure | 0.508 |

(f) Weight naive Bayes over sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.763 |
| Balance | 0.681 |
| CreditScore | 0 |
| EstimatedSalary | 0 |
| Gender | 1 |
| Geography | 0.735 |
| HasCrCard | 0 |
| IsActiveMember | 0 |
| NumOfProducts | 0.061 |
| Tenure | 1 |

(g) Weight svm under sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 1 |
| Balance | 0 |
| CreditScore | 0 |
| EstimatedSalary | 0 |
| Gender | 1 |
| Geography | 1 |
| HasCrCard | 0 |
| IsActiveMember | 1 |
| NumOfProducts | 0.061 |
| Tenure | 0 |

(h) Weight decision tree under sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.763 |
| Balance | 0 |
| CreditScore | 0 |
| EstimatedSalary | 1 |
| Gender | 0.961 |
| Geography | 0 |
| HasCrCard | 0.934 |
| IsActiveMember | 0 |
| NumOfProducts | 0.061 |
| Tenure | 0 |

(i) Weight naive Bayes under sampling.

Figure A.1 Weight value for split validation 70%, 30% after using the optimize weight function in Rapidminer.

| Attribute | Weight |
|-----------------|--------|
| Age | 0 |
| Balance | 1 |
| CreditScore | 0 |
| EstimatedSalary | 0 |
| Gender | 0 |
| Geography | 0 |
| HasCrCard | 1 |
| IsActiveMember | 0 |
| NumOfProducts | 0.061 |
| Tenure | 0.488 |

(a) Weight svm original data.

| Attribute | Weight |
|-----------------|--------|
| Age | 1 |
| Balance | 0 |
| CreditScore | 0 |
| EstimatedSalary | 0 |
| Gender | 0 |
| Geography | 1 |
| HasCrCard | 1 |
| IsActiveMember | 0.126 |
| NumOfProducts | 1 |
| Tenure | 1 |

(b) Weight decision tree original data.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.780 |
| Balance | 0.689 |
| CreditScore | 0.622 |
| EstimatedSalary | 0.866 |
| Gender | 1 |
| Geography | 0.749 |
| HasCrCard | 0.970 |
| IsActiveMember | 0.072 |
| NumOfProducts | 0 |
| Tenure | 0.474 |

(c) Weight naive Bayes original data.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.780 |
| Balance | 0.689 |
| CreditScore | 0.622 |
| EstimatedSalary | 0.866 |
| Gender | 1 |
| Geography | 0.749 |
| HasCrCard | 0.970 |
| IsActiveMember | 0.072 |
| NumOfProducts | 0 |
| Tenure | 0.474 |

(d) Weight svm over sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.763 |
| Balance | 0 |
| CreditScore | 0 |
| EstimatedSalary | 0 |
| Gender | 1 |
| Geography | 1 |
| HasCrCard | 1 |
| IsActiveMember | 1 |
| NumOfProducts | 0.061 |
| Tenure | 0 |

(e) Weight decision tree over sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.794 |
| Balance | 0 |
| CreditScore | 0.646 |
| EstimatedSalary | 0.875 |
| Gender | 1 |
| Geography | 0.765 |
| HasCrCard | 0.972 |
| IsActiveMember | 0.131 |
| NumOfProducts | 0.064 |
| Tenure | 0.508 |

(f) Weight naive Bayes over sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.763 |
| Balance | 0.681 |
| CreditScore | 0 |
| EstimatedSalary | 0 |
| Gender | 0 |
| Geography | 0.735 |
| HasCrCard | 1 |
| IsActiveMember | 1 |
| NumOfProducts | 0.061 |
| Tenure | 0.488 |

(g) Weight svm under sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 1 |
| Balance | 1 |
| CreditScore | 0 |
| EstimatedSalary | 0 |
| Gender | 1 |
| Geography | 0.735 |
| HasCrCard | 1 |
| IsActiveMember | 0.126 |
| NumOfProducts | 1 |
| Tenure | 0 |

(h) Weight decision tree under sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.763 |
| Balance | 1 |
| CreditScore | 0 |
| EstimatedSalary | 1 |
| Gender | 1 |
| Geography | 0 |
| HasCrCard | 1 |
| IsActiveMember | 0 |
| NumOfProducts | 0.061 |
| Tenure | 1 |

(i) Weight naive Bayes under sampling.

Figure A.2 Weight value for split validation 80%, 20% after using the optimize weight function in Rapidminer.

| Attribute | Weight |
|-----------------|--------|
| Age | 1 |
| Balance | 0 |
| CreditScore | 0 |
| EstimatedSalary | 0 |
| Gender | 0 |
| Geography | 0 |
| HasCrCard | 0 |
| IsActiveMember | 1 |
| NumOfProducts | 0.061 |
| Tenure | 0 |

(a) Weight svm original data.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.763 |
| Balance | 0 |
| CreditScore | 1 |
| EstimatedSalary | 0 |
| Gender | 0.961 |
| Geography | 0 |
| HasCrCard | 0.934 |
| IsActiveMember | 0 |
| NumOfProducts | 0.061 |
| Tenure | 0 |

(b) Weight decision tree original data.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.665 |
| Balance | 0.619 |
| CreditScore | 0.209 |
| EstimatedSalary | 0.367 |
| Gender | 1 |
| Geography | 0 |
| HasCrCard | 0.296 |
| IsActiveMember | 0.602 |
| NumOfProducts | 0.998 |
| Tenure | 0.552 |

(c) Weight naive Bayes original data.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.581 |
| Balance | 0.053 |
| CreditScore | 0.580 |
| EstimatedSalary | 0.431 |
| Gender | 0.909 |
| Geography | 1 |
| HasCrCard | 0.973 |
| IsActiveMember | 0.499 |
| NumOfProducts | 0 |
| Tenure | 0.865 |

(d) Weight svm over sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.763 |
| Balance | 1 |
| CreditScore | 0 |
| EstimatedSalary | 0 |
| Gender | 1 |
| Geography | 0.735 |
| HasCrCard | 0 |
| IsActiveMember | 0.126 |
| NumOfProducts | 0.061 |
| Tenure | 1 |

(e) Weight decision tree over sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.794 |
| Balance | 0 |
| CreditScore | 0.646 |
| EstimatedSalary | 0.875 |
| Gender | 1 |
| Geography | 0.765 |
| HasCrCard | 0.972 |
| IsActiveMember | 0.131 |
| NumOfProducts | 0.064 |
| Tenure | 0.508 |

(f) Weight naive Bayes over sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.763 |
| Balance | 1 |
| CreditScore | 0 |
| EstimatedSalary | 0 |
| Gender | 1 |
| Geography | 0.735 |
| HasCrCard | 0 |
| IsActiveMember | 1 |
| NumOfProducts | 0.061 |
| Tenure | 1 |

(g) Weight svm under sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.763 |
| Balance | 0.681 |
| CreditScore | 0 |
| EstimatedSalary | 0 |
| Gender | 1 |
| Geography | 0.735 |
| HasCrCard | 0 |
| IsActiveMember | 0 |
| NumOfProducts | 0.061 |
| Tenure | 0 |

(h) Weight decision tree under sampling.

| Attribute | Weight |
|-----------------|--------|
| Age | 0.763 |
| Balance | 0 |
| CreditScore | 1 |
| EstimatedSalary | 1 |
| Gender | 1 |
| Geography | 0.735 |
| HasCrCard | 0 |
| IsActiveMember | 1 |
| NumOfProducts | 0.061 |
| Tenure | 0 |

(i) Weight naive Bayes under sampling.

Figure A.3 Weight value for cross validation after using the optimize weight function in Rapidminer.

The logo of Sakon Nakhon Rajabhat University is a large, faint watermark in the background. It features a central figure of a person standing on a pedestal, surrounded by a stylized architectural structure resembling a pagoda or a traditional Thai building. The base of the logo is a semi-circular arrangement of red and orange triangular shapes. The text "APPENDIX B" and "CONFUSION MATRIX PERFORMANCE" is centered over the logo.

APPENDIX B
CONFUSION MATRIX PERFORMANCE

มหาวิทยาลัยเทคโนโลยีสุรนารี

This chapter presents the confusion matrix by model.

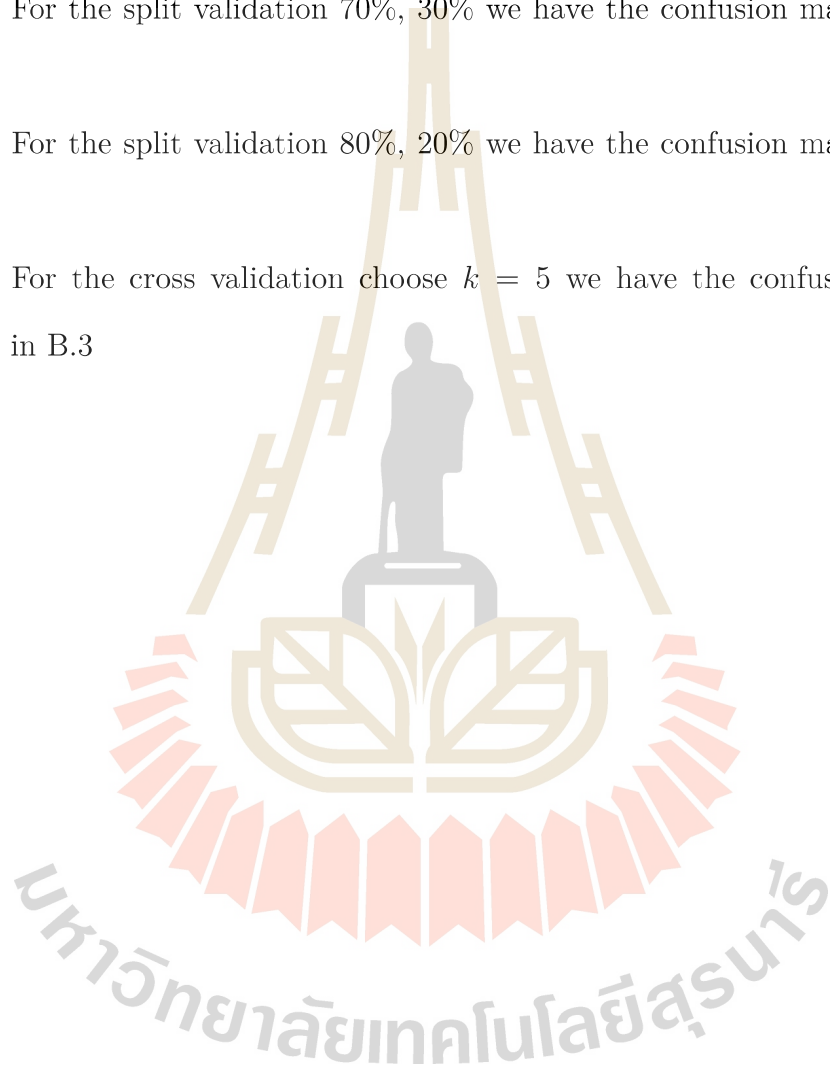
B.1 the confusion matrix for each model

In this chapter we show the confusion matrix.

For the split validation 70%, 30% we have the confusion matrix as shown in B.1

For the split validation 80%, 20% we have the confusion matrix as shown in B.2

For the cross validation choose $k = 5$ we have the confusion matrix as shown in B.3



| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 4 | 0 |
| | Retain | 60 | 2,386 |

(a) SVM original data

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 4 | 3 |
| | Retain | 60 | 2,383 |

(b) Decision tree original data

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 5 | 0 |
| | Retain | 59 | 2,386 |

(c) Naive Bayes original data

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 2,133 | 239 |
| | Retain | 197 | 2,209 |

(d) SVM over sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 1,937 | 399 |
| | Retain | 393 | 2,049 |

(e) Decision tree over sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 1,849 | 571 |
| | Retain | 481 | 1,877 |

(f) Naive Bayes over sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 53 | 9 |
| | Retain | 12 | 48 |

(g) SVM random under sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 56 | 20 |
| | Retain | 9 | 37 |

(h) Decision tree random under sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 46 | 6 |
| | Retain | 19 | 51 |

(i) Naive Bayes random under sampling

Figure B.1 Confusion matrix in cross validation 70%, 30% technique by using SVM, Decision tree and Naive Bayes classifier

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 4 | 0 |
| | Retain | 39 | 1,590 |

(a) SVM original data

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 2 | 1 |
| | Retain | 41 | 1,589 |

(b) Decision tree original data

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 0 | 0 |
| | Retain | 43 | 1,590 |

(c) Naive Bayes original data

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 1,439 | 168 |
| | Retain | 119 | 1,459 |

(d) SVM over sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 1,305 | 278 |
| | Retain | 253 | 1,349 |

(e) Decision tree over sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 1,225 | 379 |
| | Retain | 333 | 1,248 |

(f) Naive Bayes over sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 33 | 6 |
| | Retain | 6 | 36 |

(g) SVM random under sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 31 | 13 |
| | Retain | 8 | 29 |

(h) Decision tree random under sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 31 | 7 |
| | Retain | 8 | 35 |

(i) Naive Bayes random under sampling

Figure B.2 Confusion matrix in split validation 80%, 20% technique by using SVM, Decision tree and Naive Bayes classifier

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 30 | 6 |
| | Retain | 173 | 7,957 |

(a) SVM original data

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 17 | 11 |
| | Retain | 189 | 7,952 |

(b) Decision tree original data

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 9 | 0 |
| | Retain | 194 | 7,963 |

(c) Naive Bayes original data

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 7,405 | 840 |
| | Retain | 558 | 7,123 |

(d) SVM over sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 7,184 | 1,405 |
| | Retain | 779 | 6,558 |

(e) Decision tree over sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 6,222 | 1,848 |
| | Retain | 1,741 | 6,115 |

(f) Naive Bayes over sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 154 | 35 |
| | Retain | 49 | 168 |

(g) SVM random under sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 153 | 47 |
| | Retain | 50 | 156 |

(h) Decision tree random under sampling

| | | Actual | |
|-----------|--------|--------|--------|
| | | Close | Retain |
| Predicted | Close | 146 | 47 |
| | Retain | 57 | 156 |

(i) Naive Bayes random under sampling

Figure B.3 Confusion matrix in cross validation technique choose $k = 5$ by using SVM, Decision tree and Naive Bayes classifier

CURRICULUM VITAE

NAME : Jirakit Boonmunewai

GENDER : Male

EDUCATION BACKGROUND:

- Bachelor of Science (Mathematics Education), Nakhon Ratchasima Rajabhat University, Thailand, 2018

SCHOLARSHIP:

- Outstanding Study Performance Scholarship for Graduate Student of Suranaree University of Technology.

CONFERENCE:

- 5th National Conference on Quality Management and Technology Innovation Proceeding, The Eastern University of Management and Technology, Ubon Ratchathani, May 23rd, 2020.

EXPERIENCE:

- Teaching assistant in Suranaree University of Technology, Term 2019-2020.
- Assistant Lecturer in Mathematics Training Camp for High School Students, Technopolis, Suranaree University of Technology, 2017-2018.