TABLE OF CONTENT

COVER

ORIGINALITY STATEMENT OF THESIS

THESIS SUPERVISOR'S APPROVAL

THESIS EXAMINATION COMMITEE

ABSTRACT ..... **v**

ABSTRAK ..... **vi**

FOREWORD ..... **vii**

TABLE OF CONTENT ..... **viii**

LIST OF FIGURES ..... **x**

LIST OF TABLES ..... **xi**

CHAPTER I : INTRODUCTION **1**

1.1 Background 1

1.2 Problem Identification 3

1.3 Scope of Work 3

1.4 Research Questions 3

1.5 Research Objectives 4

1.6 Structure Overview 4

CHAPTER II : LITERATURE REVIEW **6**

2.1 Customer Retention and Churn 6

2.2 Customer Retention and Churn in Education Sector 9

2.3 Machine Learning for Student Retention Prediction 12

2.3.1 Random Forest 12

2.3.2 Logistic Regression 13

2.3.3 Support Vector Machine 14

2.3.4 Neural Network 14

2.4 Evaluation Methods 15

CHAPTER III : RESEARCH METHODOLOGY **17**

3.1 Research Design 17

3.2 Data Collection 17

3.3 Data Pre-Processing 19

3.3.1 Feature Extraction 19

3.3.2 Imputation 22

3.3.3 Feature Selection 23
CHAPTER IV : EXPERIMENT AND RESULT 25
4.1 Data Splitting 25
4.2 Experiments on Models 25
4.2.1 Random Forest 26
4.2.2 Logistic Regression 28
4.2.3 Support Vector Machine 29
4.2.4 Neural Network 30
4.3 Evaluation of the Models 32
4.4 Best Predictor of Student Retention 34
4.5 Managerial Implications 35
CHAPTER V : SUMMARY AND FUTURE RESEARCH 37
5.1 Summary 37
5.2 Future Research 37
REFERENCES 39
PUBLICATION A-1
FINAL ASSIGNMENT STATEMENT AND UPLOAD AGREEMENT .. A-11
TURNITIN A-12
LETTER OF ACKNOWLEDGEMENT A-13
BIOGRAPHY A-14

LIST OF FIGURES

Figure 3.1 Design science research 17

Figure 4.1 Comparison of the Random Forest parameters. 26

Figure 4.2 Confusion matrix of RF3 on training data. 27

Figure 4.3 Confusion matrix of RF3 on testing data. 28

Figure 4.4 Comparison of the Logistic Regression parameters. 28

Figure 4.5 Confusion matrix of LR1 on training data. 28

Figure 4.6 Confusion matrix of LR1 on testing data 29

Figure 4.7 Comparison of the Support Vector Machine parameters. 30

Figure 4.8 Confusion matrix of SVM3 on training data. 31

Figure 4.9 Confusion matrix of SVM3 on testing data 31

Figure 4.10 Comparison of the Neural Network parameters 33

Figure 4.11 Confusion matrix of NN2 on training data 33

Figure 4.12 Confusion matrix of NN2 on testing data 34

Figure 4.14 Comparison of all models on training data 36

Figure 4.15 Comparison of all models on testing data 37

LIST OF TABLES

Table 2.1 List of studies of retention and churn predictions. 8

Table 2.2 List of studies that used machine learning in education sector. 10

Table 3.1 Features obtained from dataset. 20

Table 3.2 Features with missing values. 22

Table 3.3 Results of features correlations with Retain variable. 23

Table 4.1 Random Forest parameters. 26

Table 4.2 Logistic Regression parameters. 28

Table 4.3 Support Vector Machine parameters 30

Table 4.4 Neural Network parameters. 33

Table 4.5 Confusion matrix on training data. 36

Table 4.6 Confusion matrix on testing data. 36

Table 4.7 Comparison of all models on training data. 37

Table 4.8 Comparison of all models on testing data. 38

Table 4.9 Top 5 predictors of student retention. 39