Machine Learning in Population Health:

Frequent Emergency Department Utilization Pattern Identification and Prediction

A Thesis Submitted to the

College of Graduate and Postdoctoral Studies

in Partial Fulfilment of the Requirements

for the degree of Master of Science

in Community and Population Health Sciences Program

University of Saskatchewan

Saskatoon, Canada

By

Razieh Safaripour

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Community Health and Epidemiology Department
Health Sciences Building E-Wing, 104 Clinic Place
University of Saskatchewan
Saskatoon, Saskatchewan S7N 2Z4, Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9, Canada

Abstract

Emergency Department (ED) overcrowding is an emerging risk to patient safety and may significantly affect chronically ill people. For instance, overcrowding in an ED may cause delays in patient transportation or revenue loss for hospitals due to hospital diversion. Frequent users with avoidable visits play a significant role in imposing such challenges to ED settings. Non-urgent or "avoidable" ED use induces overcrowding and cost increases due to unnecessary tests and treatment. It is, therefore, valuable to understand the pattern of the ED visits among a population and prospectively identify ED frequent users, to provide stratified care management and resource allocation. Although most current models use classical methods like descriptive analysis or regression modelling, more sophisticated techniques may be needed to increase the accuracy of outcomes where big data is in use.

This study focuses on the Machine Learning (ML) techniques to identify the ED usage pattern among frequent users and to evaluate the predicting ability of the models. I performed an extensive literature review to generate a list of potential predictors of ED frequent use. For this thesis, I used Korean Health Panel data from 2008 to 2015. Individuals with at least one ED visit were included, among whom those with four or more visits per year were considered frequent ED users. Demographic and clinical data was collected. The relationship between predictors and ED frequent use was examined through multivariable analysis. A K-modes clustering algorithm was applied to identify ED utilization patterns among frequent users. Finally, the performance of four machine learning classification algorithms was assessed and compared to logistic regression. The classification algorithms used in my thesis were Random Forest, Support Vector Machine (SVM), Bagging, and Voting. The models' performance was evaluated based on Positive Predictive Value (PPV), sensitivity, Area Under Curve (AUC), and classification error.

A total of 9,348 individuals with 15,627 ED visits were eligible for this study. Frequent ED users accounted for 2.4% of all ED visits. Frequent ED users tended to be older, male, and more likely to be using ambulance as a mode of transport than non-frequent ED users. In the cluster analysis, we identified three subgroups among frequent ED users: (i) older patients with respiratory system complaints, the highest discharged rates who were more likely to visit in Spring and Winter, (ii) older patients with the highest rate of hospitalization, who are also more likely to have used ambulance, and visited ED due to circulatory system complaints, (iii) younger patients, mostly female, with the highest

rate of ED visits in summer, and lowest rate of using an ambulance, who visited ED mostly due to damages such as injuries, poisoning, etc. The ML classification algorithms predicted frequent ED users with high precision (90% - 98%) and sensitivity (87% - 91%), while showed high AUC scores from 89% for SVM to 96% for Random Forest, as well. The classification error varied among algorithms; logistic regression had the highest classification error (34.9%) while Random Forest had the least (3.8%). According to the Random Forest Importance Score, the top 5 factors predicting frequent users were disease category, age, day of the week, season, and sex.

In this thesis, I showed how ML methods applies to ED users in population health.
The study results show that ML classification algorithms are robust techniques with predictive power for future ED visit identification and prediction. As more data are collected and the amount of data availability increases, machine learning approaches is a promising tool for advancing the understanding of such 'Big' data.

Acknowledgment

Throughout the writing of this thesis, I have received a great deal of support and I wish to thank all the people whose assistance was a milestone in the completion of this project.

I would like to express my deepest gratitude to my supervisor, Dr. Hyun J. "June" Lim, whose constructive comments, patience, insightful guidance, and immense knowledge lightened my way through the learning process of this master thesis. I would not have dared to step into this research if it were not for her encouragement, support, and trust in me. I will always be grateful to her for challenging me to conduct this research. I wholeheartedly appreciate that she created a friendly and stress-free environment through her mental and financial support, making it possible for me and my teammates to focus on our research. Thank you, Dr. Lim!

My sincere thanks go to my friend, Dr. Arsia Takeh, for being such a wonderful person whose help, suggestions, and ideas were so important to accomplish this thesis.

My thanks and appreciations go to my supervisory committee members Dr. Nicolas Pena-Sanchez and Dr. Nazeem Muhajarin, for taking the time to assess my thesis and for their helpful suggestions and valuable comments.

My special regards go to the Community Health and Epidemiology department crew and all my professors who provided a home away from home for me. I am very proud to be a member of the CHEP community.

Last but not least, I would like to thank my family, especially my parents, for their pure everlasting love and support. They have always been there for me, giving me hope and strength where I needed most. I can never be grateful enough for what they have done for me.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACSC	Ambulatory Care Sensitive Conditions
AUC	Area Under Curve
CART	Classification and Regression Trees
CDC	Disease Control and Prevention
CERES	Clouds and the Earth's Radiant Energy System
CIHI	Canadian Institute for Health Information
ED	Emergency Department
ICU	Intensive Care Unit
KCD	Korea Classification of Disease
KDD	Knowledge Discovery in Databases
KHPS	Korea Health Panel Study
ML	Machine Learning
NACRS	National Ambulatory Care Reporting System
NHI	National Health Insurance
NHIC	National Health Insurance Corporation
OECD	Organization for Economic Co-operation and Development
OOP	Out-Of-Pocket
OR	Odds Ratio
PPV	Positive Predective Value
RBF	Radial Basis Function
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristics
SDoH	Social Determenant of Health
SVM	Support Vector Machine
VIF	Variance Inflation Factor

Glossary

Accuracy	Proportion of results correctly classified [i.e., (true positives plus true negatives) divided by total number of results predicted]
Big data	Big data refers to the large, divers sets of information that grow at ever-increasing rates. It encompasses the volume of information, the velocity or speed at which it is created and collected, and the variety or scope of the data points being covered (Known as the "three Vs" of big data)
Data mining	Exploratory analysis
Ensemble learning	A machine-learning approach involving training multiple models on data subsets and combining results from these models when predicting for unobserved inputs.
Feature	Measurements recorded for each observation (e.g., participant age, sex, and body mass index are all features)
Label	Observed or computed value of an outcome or other variable of interest
Learning algorithm	The set of steps used to train a model automatically from a data set (not to be confused with the model itself, e.g., there are many algorithms to train a neural network, each with different bounds on time, memory, and accuracy)
Overfitting	Fitting a model to random noise or error instead of the actual relationship (due to having either a small number of observations or a large number of parameters relative to the number of observations)
Precision	Positive predictive value
Recall	Sensitivity
Supervised learning	An analytic technique in which patterns in covariates that are correlated with observed outcomes are exploited to predict outcomes in a data set or sets in which the correlates were observed but the outcome was unobserved. For example, linear regression and logistic regression are both supervised learning techniques.
Unsupervised learning	An analytic technique in which data are automatically explored to identify patterns, without reference to outcome information. Latent class analysis (when used without covariates) and k-means clustering are unsupervised learning techniques.
Training	Fitting a model
Training dataset	A subset of a more complete data set used to train a model whose empirical performance can be tested on a test data set.
Test dataset	A subset of a more complete data set used to test empirical performance of an algorithm trained on a training dataset.

Chapter 1

Introduction

Emergency Department (ED) is a high-cost care setting; the cost associated with visiting ED is a significant burden on the health care system across the globe. For example, the cost of EDs utilization is as high as $1.8 billion per year in Canada (Dawson & Zinck, 2009). In the UK, the annual cost of non-admitted emergency department visits was £2,300 million in 2014-2015, of which about 26.5% were due to unsuccessful access to a primary care(Whittaker et al., 2016). According to the Center for Disease Control and Prevention (CDC), the cost imposed by ED patients with Ambulatory Care Sensitive Conditions (ACSC), which could be prevented with timely primary care, accounts for US$38 billion of total healthcare cost in the United States (Agarwal et al., 2016; Coleman et al., 2001).

EDs overcrowding is another issue that significantly impacts the quality of care for patients. In the year 2014-2015, over 10 million EDs visits were registered in the National Ambulatory Care Reporting System (NACRS), which was about 63% of all EDs visits in Canada (Canadian Institute for Health Information (CIHI), 2015). Another report from the Organization for Economic Co-operation and Development (OECD) in 2015 showed that EDs visits in 21 member countries increased from 29.3% in 2001 to 30.8% in 2011 (≈ 5.2% raise)(Berchet, 2015). The report shows that in the United States, EDs users under six years old (7% of the population) represent 24.2% of the ED visits, and individuals of 75 years old and higher (6% of the population) accounts for 27% of the EDs visits (Berchet, 2015). Injuries such as fracture, dislocation, sprain, or strain accounted for an average of 25% of the visits to EDs (Berchet, 2015). In Canada, trauma, respiratory system complaints, pneumonia, and abdominal pain were among the most common reasons to visit EDs (Canadian Institute for Health Information, 2018).

The most common definition of *frequent ED visits* is four or more visits per year (E. LaCalle & Rabin, 2010; Pham, 2017). In Canada, frequent ED users account for over 30% of all EDs visits, of which the age groups of less than 18 and over 85 contribute to 7% and 15% of visits, respectively (Canadian

Institute for Health Information (CIHI), 2015). Studies show that in addition to medical problems (Fuda & Immekus, 2006; Hunt et al., 2006), frequent ED users are often vulnerable individuals with psychosocial risk factors such as belonging to a minority group, unemployment, alcohol dependence, and number of psychiatric hospitalization (Bieler et al., 2012; E. J. LaCalle et al., 2013; Mandelberg et al., 2000; Soril et al., 2016), and/or experience difficulties with access to care (Burns, 2017; Franchi et al., 2017; Han et al., 2007; Pham, 2017). Although the chief concern of the patient seems to be clinical, the underlying issues go back to their living situation, food insecurities, ect., which together form social determinants of health (SDoH)(Folckele et al., 2019). According to Fockele C. et al., health behaviour (30%), access to and quality of care (20%), Social and economic factors (40%) and physical environment (10%) are all contributing to ED visits. They suggest that identifying SDoH barriers and advocating for patients' SDoH needs could increase the life quality and expectancy of patients, while decreasing the cost of healthcare(Folckele et al., 2019).

Frequent use of ED brings challenging issues to healthcare system authorities (Pines et al., 2011; Raven, 2011). Given the challenges associated with patients' characteristics identification, predictive modelling is an approach to classify ED users who are most likely to enforce heavy burden on health care services in the future (Raven, 2011).

Outcomes of predictive models depend on analytical methods, data source quality, and the extent of accessible data features. Although national administrative databases are very limited in clinical information, they have some advantages such as "*accessibility*" and "*a greater number of features*" (Ohno-Machado, 2011) that make them an appropriate source for predicting future EDs users.

Despite a significant rise in the amount of data (volume) and the speed at which data is generated (velocity) in the population health field, the majority of current predictive models use classical methods like regression (Chiu et al., 2019; Fuller et al., 2017; Raven, 2011). However, to increase the accuracy of outcome prediction for such big data, more sophisticated algorithms of Machine Learning (ML) are appropriate. One of the main advantages of more advanced ML algorithms compared to logistic regression is the absence of assumptions that reduce the capacity of the logistic regression to deal with the unstructured nature of big data (Artificial Intelligence and Population Health, 2017). On the other hand, it is impossible for planners and practitioners across healthcare systems to fully explore high-dimensional population-level data sets using traditional methods (Ravaut et al., 2021). For example, exploring all the possible interactions in high-dimensional datasets is considerably costly in terms of

time and human resources. However, there is a debate regarding the advantages of more advanced ML over logistic regression because of the limited number of studies on health applications of ML (Morgenstern et al., 2020), which makes this thesis even more relevant in its domain. The need to extract valuable knowledge hidden in the ever-growing amount of health-related data to improve work efficiency and enhance the quality of decision-making is inevitably crucial. Understanding the patterns of ED utilization will help provide stratified care management and resource allocation, decrease the cost and overcrowding of ED, and improve the quality of care.

The three objectives of this study are:

1. To analyse factors associated with frequent ED visits in the South Korean general population.
2. To identify patterns of ED usage among frequent users through ML clustering method.
3. To use ML classification algorithms to evaluate their ability in predicting frequent ED users and compare their performances with logistic regression.

In the following, Chapter 2 provides a literature review on ED utilization and those of machine learning in population health, followed by a brief background on data mining and machine learning technology. Chapter 3 explores the method of the study, including study design, data pre-processing, and analysis techniques. Chapter 4 provides the results of the analysis of Korean health panel data. Chapters 5 and 6 discuss the findings of the study, the conclusion, and the future research directions.

Chapter 2

Literature Review

Medical facilities like EDs are essential to providing rapid access to care for patients with acute health conditions. EDs overcrowding imposes a serious threat to patients with urgent needs by deteriorating appropriateness of care (Ng & Jordan, 2002). Multiple studies estimate frequent users contribute to as little as 2.7% to 8% of patients but account for up to 67% of all EDs visits in North America (Griswold et al., 2005; Krieg et al., 2016; Lucas & Sanford, 1998).

This chapter provides a literature review on EDs utilization, summarized in three sections: i) observational studies analyzing EDs visits, ii) pattern identification using ML clustering technique, and iii) frequent EDs users prediction using ML classification algorithms.

2.1 ED Utilization

A variety of studies have explored the characteristics of frequent EDs users within different settings and populations. A systematic review of the characteristics of EDs users among the general adult population shows that in the countries with national as well as private health insurance systems, frequent users were more likely to have mental health, cardiovascular, or respiratory diagnoses and be low-income (Soril et al., 2016). Other factors associated with frequent users were age over 65, being unemployed, substance abuse issues, access to primary care, and being a male (Soril et al., 2016). Another study in South Korea found that frequent EDs users were associated with older age, males, and lower socioeconomic status (Woo et al., 2016). Furthermore, studies show that in addition to medical problems, frequent EDs users are often individuals with psychosocial risk factors (E. J. LaCalle et al., 2013) or challenging access to care (Han et al., 2007).

Table 2-1 summarizes the most relevant articles including study country, target population, cohort size, objective(s) and the data source.

Table 2-1 Literature Review Summary: Emergency department users characteristics

Authors, Years, Country	**Population**	**Cohort size**	**Objective**	**Data source**	**ED Characteristics**
(E. J. LaCalle et al., 2013), USA	All	59,172 unique patients	To describe the demographic and utilization characteristics of patients who visit the ED 20 or more times per year.	An ED setting within an urban county of 1.5 million inhabitants	30-59 years of age, stably insured, and had at least one significant psychosocial cofactor such as homelessness or substance abuse
(Soril et al., 2016), Canada	All	Twenty moderates to high-quality comparative cohort studies	to synthesize and compare population characteristics associated with frequent emergency department (ED)	All	older, female, and had a mental health diagnosis.
(Han et al., 2007), Canada	Patients 17 years and older	894 participants	To assess the frequency and determinants of patients' efforts to access alternative care before ED presentation	2 urban ED sites in Edmonton	Injury presentation, living arrangements, smoking status and whether patients had a family practitioner
(Kim et al., 2018), South Korea	Children under 16 years old	33 765 patients in a one-year period	To identify the Characteristics of recurrent visits.	Medical records of three university hospital (2012)	Using the 119-rescue center service, having a medical condition, with younger age and a higher rate of hospitalization
(Seo et al., 2018), South Korea	Patients under 18 years old with at least one ED visit	3,160 pediatric ED visits	To investigate the characteristics of pediatric ED patients and to determine factors associated with hospital admission after ED	Korea Health Panel data from 2008 to 2013	Male, under 5 years old, lack of private insurance, living in provinces, lower income, due to diseases

			treatment		than injuries.
(Woo et al., 2016), USA	All	156,246 individuals	To understand the characteristics of frequent ED users in Korea	Korea Health Insurance Review Agency (HIRA), records of 2009	Older, male, and of lower socio-economic status, longer stays in the hospital when admitted, higher probability of undergoing an operative procedure, and increased mortality
(Ustulin et al., 2018), South Korea	Patients with type 2 diabetes mellitus	109,412 individuals	To identify the characteristics of frequent ED users among Korean patients with type 2 diabetes mellitus	Health Insurance Review and Assessment Service National Patient Sample	Men, longer treatment duration, more frequent comorbidities (cardiovascular and chronic kidney disease), higher mortality, longer hospitalization duration, higher costs per visit

Although the presented studies have investigated the ED users characteristics in different settings, none has explored ED utilization in different days and/or seasons. This thesis provides a unique finding regarding daily and seasonal pattern of frequent ED users. Moreover, as presented in the table, all the studies have explored the risk factors associated with ED utilization in a conventional approach, whereas, identifying the pattern of visits among a large group of population could inform tailored intervention design. ML clustering method provides such opportunity.

2.2 Pattern Identification Using Clustering Technique

Clustering is a popular technique to identify patterns from massive data. In clinical and health sciences, clustering was applied in different fields, from Alzheimer's disease application to health jurisdiction categorization (Alashwal et al., 2019; Lavergne, 2016). Clustering studies pursued various objectives

such as grouping local health areas based on the distribution of healthcare spending (Lavergne, 2016), identifying previously unknown patterns of clinical characteristics among home-care clients (Armstrong et al., 2012), and to categorized general patient populations into homogeneous groups (Vuik et al., 2016). K-means was the dominant algorithm in the health-related studies, whereas K-modes were mostly seen in non-health-related applications.

One study used clustering to segment high-risk patients visiting all types of care settings, including the emergency department (Vuik et al., 2016). The study obtained data from both clinical and administrative databases. Another study used clustering to identify patterns of cost change in a specific group of patients (Liao et al., 2016). Vranas et al. used clustering to determine if the care program appropriately targets Intensive Care Unit (ICU) patients by identifying patients' characteristics in each subgroup; age, clinical diagnosis, and long-term care were among the factors that separated the clusters of patients (Vranas et al., 2017). Moreover, Armstrong et al. used clustering to explore the characteristics of home-care patients in need of rehabilitation services (Armstrong et al., 2012). They used the clusters to identify the clinical features of each subgroup; age, sex, cognition, and functionality were some of the dissimilarities of the groups.

Table 2-2 summarizes the studies that used clustering techniques outlining the country of the study, target population, their cohort, study objective(s), and clustering algorithm.

Table 2-2: Literature Review: Clustering Analysis

Authors, Years, Country	Population	Cohort Size	Data Source	Objectives of the Study	Algorithms	
					K-means	K-modes
(Liao et al., 2016), USA	Individuals 18 years or more with ≥2 ESRD diagnosis	18,380 individuals	Truven Health MarketScan® Research Databases.	To identify cost change patterns of patients with end-stage renal disease (ESRD) who initiated hemodialysis (HD) by applying different clustering methods	Yes	No
(Lavergne, 2016), Canada	All	89 Local Health Areas (LHAs) in BC, with populations from under 4,000 to over 300,000	Health System Matrix, Health System Planning Division of BC's Ministry of Health	To group Local Health Areas based on the distribution of healthcare spending across service categories	Yes	No
(Armstrong et al., 2012), Canada	Home care clients who use rehab services	150,253 clients	Resident Assessment Instrument–Home Care (RAI-HC)	To examine the heterogeneity of home care clients who use rehabilitation services to identify previously unknown patterns of clinical characteristics	Yes	No
(Vranas et al., 2017), USA	ICU patients aged ≥ 18 years	5,000 patients randomly selected from 24,884 ICU patients	Kaiser Permanente Northern California (KPNC), an integrated healthcare delivery system	To empirically identify ICU patient subgroups and evaluate whether these groups represent appropriate targets for care redesign efforts.	Yes	No
(Vuik et al., 2016), UK	All	a random sample of	Clinical Practice Research Data	To explore the potential of using	Yes	No

		300,000 individual	(CPRD), Hospital Episode Statistics (HES), and the Townsend Index of Deprivation 2001	utilization-based cluster analysis to segment a general patient population into homogeneous groups		

The reviewed studies explore important applications of ML clustering in population health and health services. However, they are all using k-mean algorithm, and none has analyzed the ability of ML clustering in identifying the pattern of ED usage among of frequent users. This thesis examines this method using k-mode algorithm among South Korean population.

2.3 Frequent ED User Prediction Using ML Classification

There has been a variety of research efforts for predicting future EDs visits. A scoping review on statistical tools for analyses of frequent users showed that the most common practices was regression and hypothesis testing (Chiu et al., 2019; Poole et al., 2016), and only a few of the studies used machine learning method for predicting future ED visits (Chiu et al., 2019; Grinspan et al., 2015), out of which we review the five most relevant in this section.

Grinspan et al. present a retrospective cohort study to identify frequent ED utilization among people with epilepsy (Grinspan et al., 2015). They use two years of data gathered through health information exchange in New York City to implement their predictive models. However, their predictive models, including their implemented classification algorithms (i.e., Random Forest, AdaBoost, Support Vector Machine (SVM), and Classification and Regression Tree (CART)), experienced a poor sensitivity score of 20%, which means they can identify a small proportion of the frequent users in the targeted population correctly. Moreover, they did not consider insurance status as a risk factor, whereas other studies had found a significant association between the two (Puka et al., 2016; Soril et al., 2016).

Grinspan Z. et al. also focuses on ED visits by children with epilepsy (Grinspan et al., 2018). They used Health Record Data (HRD) from two centers (Weill Cornell Medical Center and Nationwide Children's Hospital) in 2013 to predict the ED visits for those children in 2014. They find the performance of 3-variable models (i.e. prior ED use, insurance, number of AED) equal or better than the machine learning algorithms and one-variable model. They evaluated the models through the

expected annual ED visits by the top 5% high-risk children in both centers. The same group of researchers also investigated the predictability of different models for frequent ED visits (>=2) among children with asthma (Das et al., 2017). They defined three criteria to select their final model: "Parsimony, Accuracy, and Interpretability." They found that the two-variable model (i.e. type of insurance and ED visit in the first year) meets all the requirements. The logistic regression technique was the best tool to predict the frequent users (AUC= 0.86, PPV=56%) but with a low sensitivity score (23%).

All these studies showed low ML performance; there are some reasons such as overfitting, transformation process, and outliers that could have impacted the learning process. However, we cannot draw a certain conclusion due to the lack of information on the preprocessing stage. Furthermore, when it comes to assessing an intervention, participant selection that is mainly based on the prior visit to EDs (e.g. last 12 months) could be questionable since ED usage of the selected population might have regressed, with no intervention at all (Raven, 2011).

Two other studies investigated models to predict future ED visits using four years of data gathered through the Indiana Public Health Emergency Surveillance System (Poole et al., 2016) and the California Office of nationwide health Planning and development (Pereira et al., 2016). Pereira M. et al. predicted the frequent users through data from 2009 to 2010 and validated the models using the data from 2011 to 2013. In contrast, Pool S. et al. compared several models to predict revisit in the next month, the next three months and the next six months. While Pool S. et al. Focused on either revisit or no revisit, Pereira M. et al. used a multiple-class classification of low, medium and high frequency (Pereira et al., 2016; Poole et al., 2016). The latter concluded that predicting low and high-frequency ED users is more accurate compared to moderate ED users. They also compared a binary classification with different thresholds and found that the AUC improves with increasing the threshold for the number of visits. The results, however, showed a poor precision score for the medium- and high-frequency classes and a low to fair sensitivity score for all three levels regardless of the models. A strength of Pereira M. et al. study was the model validation using data from different years. Moreover, fitting the models over a relatively large dataset with over 14 million data points is a solid practice of ML, although a relatively low classification score is expected. However, another reason for the low scores could be that AdaBoost and Decision Tree used in their studies are no the best ML options for such a large dataset due to their high sensitivity to outliers (AdaBoost) and instability (DT).

Table 2-3 summarizes the studies that used machine learning classification analysis, including the type of ML algorithms they have evaluated.

Table 2-3: Literature Review: Machine Learning Classification Analysis

Authors, Years and Country	Population	Study Cohort Size	Frequent User Definition	Objectives of the Study	Learning Algorithms				
					SVM	Log reg.	Random Forest	Bagging	Voting
(Grinspan et al., 2015), USA	People with Epilepsy	8041	≥4 visits per year	To Predict frequent ED use as a result of inadequate disease control or lack of access to care/ to identify demographics, comorbidity, and use of health services factors associated with ED use	Yes	Yes	Yes	No	No
(Grinspan et al., 2018), USA	Children with Epilepsy	2730 and 786 (Two cohorts)	≥1 visit for one cohort and ≥2 visits for another cohort	To predict ED usage as a result of poor access to care for disease control/ To estimate the break-even cost through reducing ED visit	Yes	Yes	Yes	No	No
(Das et al., 2017), USA	Children with Asthma	2691	≥2 visits per year	Predictability of ED use asthma using Electronic Health Record (EHR)	Yes	Yes	Yes	No	No

(Poole et al., 2016), USA	All	1,125,118	Defined as the revisit within 1 month, 3months, and 6 months, separately	Predicting frequent ED users using routinely recorded registration data to identify if they revisit ED in 1 month, 3 months, and 6 months. (3 separate models)	Yes	Yes	Yes	No	No
(Pereira et al., 2016), USA	All	14,365,975	≤ 1 visit per year (low freq.), 2-4 visits per year (medium freq.), ≥5 visits per year (high freq.)	Predicting future ED users from discharged data of hospitals / To evaluate the impact of different risk factors (i.e. demographics, prior Ed use, distance to ED, and clinical predictors)	No	Yes	No	No	No

2.4 Data Mining

Data mining is gaining popularity in various research fields due to the powerful tools and techniques for knowledge discovery. The more recent definition to distinguish it from previous statistical modelling explains data mining as "the use of *machine learning algorithms* to find faint patterns of relationship between data elements in large, noisy, and messy data sets, which can lead to actions to increase benefit in some form (diagnosis, profit, detection, etc.)" (Baker, 2010). Data mining in population health is gaining momentum because it benefits all stakeholders: communities, care providers, healthcare settings, insurers, policymakers, and researchers. Data mining finds relationships, trends, and models that can help predict outcomes and make better decisions, all while

reducing the burden on the health-care system (Tekieh & Raahemi, 2015).Data mining is technically a tool for Knowledge Discovery in Database (KDD). There are five stages involved in this process:

1) *Selection*: Selects target data from raw data
2) *Pre-processing*: *The* first step of cleaning data, such as detecting outliers and missing data.
3) *Transformation*: Reduces and projects the data to find invariant aspects
4) *Model building*: Obtains the best-fitting model through several ML algorithms
5) *Model evaluation*: Evaluates the model's accuracy (e.g., predictive power) through different evaluation metrics and confusion matrix.

Nisbet et al. have explained the relationship between data mining and KDD and the five stages in his book shown in Figure 2-1(Nisbet et al., 2017).

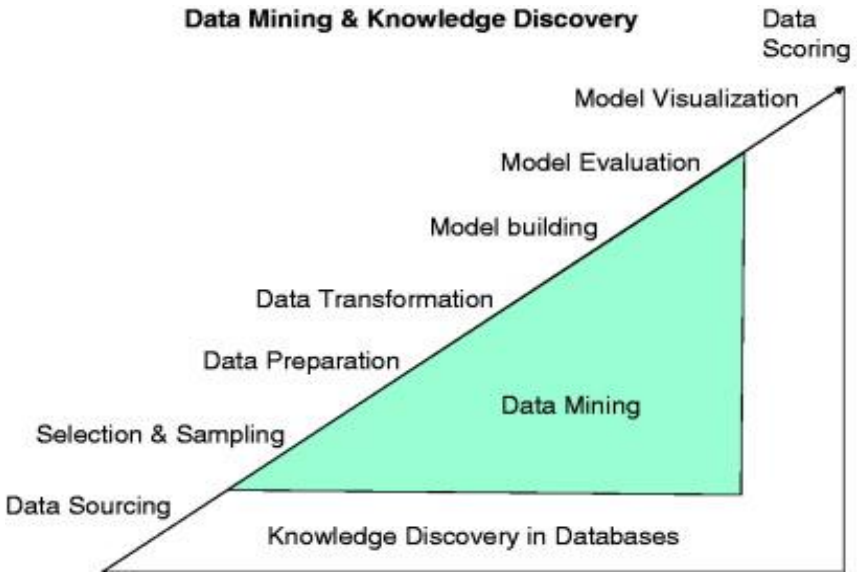


Figure 2-1: Relationship Between Data Mining and Knowledge Discovery in Database. From "Handbook of Statistical Analysis and Data Mining Applications (Second Edition), 2018" by Robert Nisbet Ph.D., ... Ken Yale D.D.S., J.D.

2.4.1 Machine Learning

Arthur Samuel (1901-1990) defines *ML* as a technology that "enables computers to learn from data, and even improve themselves, without being explicitly programmed"(Awad & Khanna, 2015). A person may miss multiple connections and relationships between data, while machine learning technology can recognize them and make a highly accurate decision for future machine's behaviour.

Machine learning uses data mining as a knowledge source to extract patterns and learn from them to adapt to future events. The essence of machine learning is making an accurate prediction of outputs for data that the model has never seen before (i.e. generalisation) (Al-Masri, 2019). In general, machine learning techniques are classified into three groups based on the way they "learn" from data: *Supervised learning (aka Predictive)*, Unsupervised *learning (aka Descriptive)*, and *Reinforcement*. Since *Reinforcement* focuses on the reward/punishment approach for dynamic learning, it is out of the scope of this study.

2.4.1.1 Supervised Learning

Supervised learning is the most common ML method used to predict an outcome of interest from unknown input data. Supervised learning algorithm learns through human guidance (e.g. data scientist) that what results are expected. It requires the possible outputs to be known, and the data for training algorithms is already labelled. A *classification algorithm* is a supervised learning algorithm that learns to identify the target after being trained on a dataset with properly labelled data and identifying characteristics (Géron, 2017) . Figure 2-2 shows the common classification algorithms from which this study evaluates *Support Vector Machine*, *Random Forest*, *Bagging,* and *Voting*.

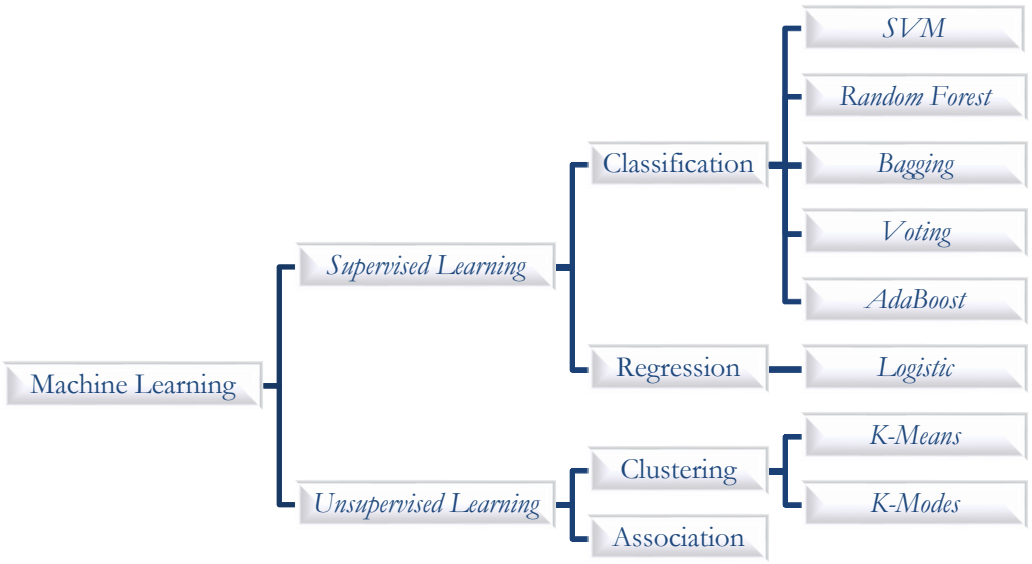


Figure 2-2: Overview Diagram of Machine Learning Common Algorithms

2.4.1.2 Unsupervised Learning

Unsupervised learning is closer to "real" artificial intelligence — "the idea that a computer can learn to identify complex processes and patterns without a human to provide guidance along the way" (Nikki Castle, 2017). Although unsupervised learning is disproportionately complex for some uncomplicated use cases, it opens the doors to solving problems humans usually would not tackle with two learning processes, *clustering* and *association mining rules* (Géron, 2017).

The structure and volume of the data, as well as the research's use case, all influence the decision on choosing supervised or unsupervised learning. Both types of algorithms will be used in a fully developed machine learning method to create predictive data models that aid researchers in overcoming various challenges.

Chapter 3

Methods and Material

3.1 Study Data

3.1.1 South Korea Healthcare System and Korea Health Panel Study

South Korea provides a mandatory National Health Insurance (NHI) covering the entire population. In 2000, all the insurance societies were merged into a single insurer, the National Health Insurance Corporation (NHIC), responsible for providing health care benefits, payment collection and reimbursement. Insurance providers, government, and Out-Of-Pocket (OOP) payments by health services users are the primary funding resource of the NHIC (Song, 2009). In 2007, the government paid about 54.9% of total health care expenditure, while the private sector supported approximately 45.1%. Of the latter, OOP payment contributed 35.7%, private health insurance paid about 4.1%, and charitable funds financed the remainder (Chun et al., 2009). Private sector providers of medical services with approximately 90% of hospital beds dominate South Kora healthcare. Primary and secondary care facilities provide healthcare services to the population. While clinics, hospitals and general hospitals are responsible for primary care, patients can access secondary care through tertiary hospitals (Song, 2009). Patients in South Korea have the freedom to select which medical institution they want to go to for care (Song, 2009).

In some cases, such as emergency medical care, the patient can go to any hospital without a referral slip, which - along with patients' preferences for large medical institutions, causes an overflow of patients in the EDs of those medical institutions. In 2018, over 10 million patients were treated in South Korean emergency rooms, an increase of 1.76% percent over the previous year, and the number of patients admitted to EDs increased by 2.95% from a year earlier (Jung et al., 2021).

The current study used the 2008-2015 Korea Health Panel Study (KHPS) data as the secondary source of information. The KHP is an official database carried out by The Korea Institute for Health and

Social Affairs and the National Health Insurance Service since 2008 (*Korean Health Panel Study (KHPS)*, 2016). The purpose of KHPS is "to generate basic data on individual healthcare behaviours, health level, usage of health services, and healthcare expenditures" (*Korean Health Panel Study (KHPS)*, 2016). The sampling frame is a national representative of the Korean population as it has used 90% of the 2015 Population and Housing Census data and has employed a two-stage probability proportionate and stratified cluster sampling method. The stratification of the population based on geographic areas yielded 237,165 clusters, out of which 350 sample clusters were extracted (*Korean Health Panel Study (KHPS)*, 2016; Seo et al., 2018). Next, households were sampled from those clusters, and finally, family members from these households formed the Korea Health Panel (*Korean Health Panel Study (KHPS)*, 2016; Seo et al., 2018).

A retrospective cohort study was conducted to investigate if additional predictive power is present when using machine learning techniques to identify ED visit patterns and predict frequent ED users. The public administrative database of South Korea Health Panel data collected from 2008 to 2015 was used for the purpose of this thesis.

3.1.2 Data Description

The available data by KHPS was initially stored in 16 separate datasets. This study used *Emergency Department Utilization (er), Household Information (hh)* and *Household Member Information (ind)* datasets. The total number of participants was 28,585, of which 9,348 (28.5 %) visited ED at least once between 2008 and 2015. Figure 3-1 summarizes the data sampling procedure.

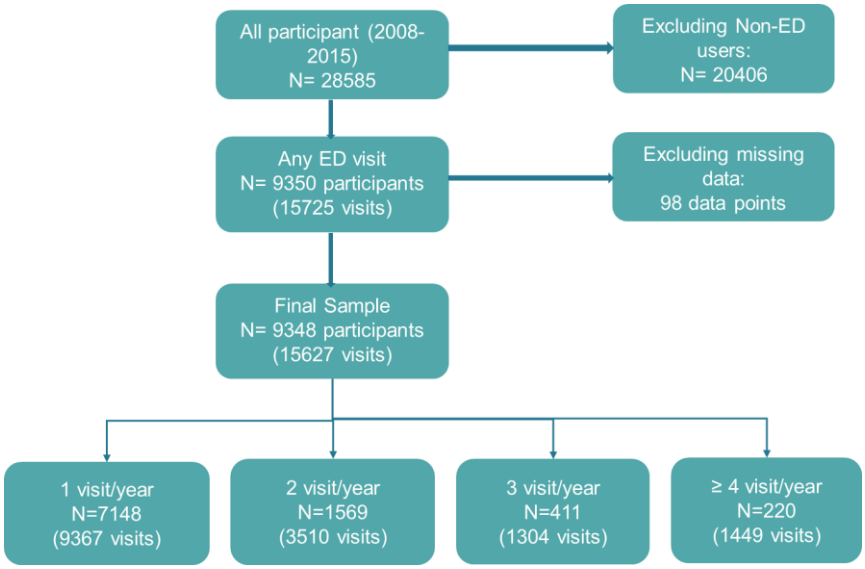


Figure 3-1: Study flowchart

3.1.3 Target

Among all ED users, 220 participants were identified as frequent users (4 or more visits per year), with 1,449 visits in that time (≈9 % of the total ED visits). For the clustering purpose, no target variable was defined. However, the target for ML classification was set as a binary outcome: Frequent ED user (Yes=1, No =0).

3.1.4 Features

The baseline characteristics included in this study were defined in 4 categories:

i) ***Demographics***, which includes age, sex, type of insurance, and living location. The age variable was categorized into five groups (0-14, 15-24, 25-44, 45-64, and 65+) according to the South Korea age structure (*South Korea Age Structure - Demographics*, 2020). The type of insurance had two levels: *National Health Insurance* (NHI) and *private insurance*, including any other insurance than NHI such as long-term care insurance, industrial disaster, and car insurance. The cities were divided based on the administrative districts of residence of patients. The administrative districts of South Korea were classified as the capital city, metropolitan cities, and provinces.

ii) ***Time of Visit***, which includes Season of the visit (i.e. Spring, Summer, Fall and Winter) and Day of Week to examine the possibility of ED visit on the weekends due to lack of out-of-hours access to primary care.

iii) ***Reason for Visit,*** which was categorized into 26 groups according to the South Korea Classification of Diseases (KCD) (*Classification of Diseases-6th Version*, 2010).

iv) ***Access to ED***, which includes the mode of transport (i.e. Ambulance vs. self-transport), the type of ED visited by the patients, and Hospitalization. The latter had three levels: in ED-death, discharged and admitted (i.e., admitted to the arriving hospital or transferred to tertiary care).

Other variables in the SDoH categories such as education or occupation were not included. In the future studies with a larger dataset, more variables in SDoH category could be analyzed.

3.2 Data preprocessing

The required above information was extracted and merged on the *PIDWON* key (i.e., individual unique ID), resulting in 9 new datasets with potential input variables for each year. Some variables were combined to form one feature (e.g., Day of Week). The majority of the ED utilization baseline features were extracted based on conventions from the ED users-related literature. Feature selection is one of the most important pre-processing steps as it directly impacts ML models' performance.

3.2.1 Homogenization and integration

Each of the created datasets had similar variables with different value types (e.g. int, float, string, etc.) and some with different names. Therefore, as a first step, all the variables' names were unified. Some years recorded the diagnosis code as a 4-digit code, some years as 5-digit code and some as alphanumeric code. The 4-digit codes were considered as the baseline, and the others were re-valued accordingly. After that, the disease codes were categorized into 26 groups in line with KCD 6, of which two groups, 'Congenital anomalies' and 'prenatal condition' with a total of three visits, were removed from the final dataset. In the next step, all data types were transformed to numeric and categorical features were "dummified" to fit logistic regression models and the learning algorithms, where applicable.

3.2.2 Missing data

Once all datasets were homogenized, missing data were identified. Out of 15,725 data points, 98 had some missing values. Missing data was registered as '-9' in the original dataset, and they were removed from the data since they contributed to a very small proportion of the final data (0.6%). No influential outliers were detected in the final dataset.

3.2.3 Data resampling

One of the common issues of ML classification is the 'unbalanced classes' issue. Data imbalance usually indicates an uneven distribution of classes within a dataset. Imbalanced data will lead to an 'illusory' high-score accuracy (Géron, 2017). In our data set, the ratio of occasional users (i.e., 1 to 3 visits per year) to frequent users (i.e., 4 or more visits per year) was about 90:10; therefore, a resampling technique was used to balance data for a more accurate and reliable result. There are two resampling techniques: *Under-sampling* and *Oversampling*. Synthetic Minority Over-sampling Technique (SMOTE) was chosen due to the several advantages over under-sampling (Géron, 2017). It should be mentioned that oversampling was executed only on the training data; therefore, no information was drained into the model training from test data. At last, the classification models were trained on a sample of 9,920: 9,920 participants.

3.2.3.1 Overfitting Problem

Overfitting is one of the most common problems in fitting machine learning algorithms. Brownlee J. has defined overfitting as follow:

> *"Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models' ability to generalize."*
> (Brownlee, 2016)

In this study, K-fold cross-validation was implemented to overcome this problem.

3.3 Statistical Analysis

Descriptive statistics were used to summarize the characteristics of the study participants. Means and standard deviations for continuous variables and frequency and percentages for categorized variables were presented. The Chi-square test was used for group comparison of categorical variables. The characteristics were analyzed based on the number of visits (i.e., 1, 2, 3, and ≥ 4 visits per year). The frequency and percentage of each variable were also calculated, and bar charts were used to visualize the results. Univariate and multivariate logistic regression analysis was used to find the relationship between these characteristics and frequent ED users.

3.3.1 Univariate Logistic Regression Analysis

The majority of the ED utilization baseline characteristics were extracted based on conventions from the ED literature (Japkowicz & Shah, 2011; Soril et al., 2016). Odds ratios (OR) and 95% confidence intervals (CI) were calculated. All reported tests were two-sided, and α=0.05 was set for statistical significance.

3.3.2 Multivariable Logistic Regression Analysis

We used the generous threshold of $P < 0.10$ in the univariate analysis to ensure that all the potential useful variables will remain for further assessment in the multivariable model (Newcombe et al., 2018). The absence of multicollinearity was confirmed using Variance Inflation Factor (VIF) scores; variables with high VIF (> 10) were set to be removed. The final multivariate logistic model contained only significant predictors with p-values <0.05, and interactions among the main predictors in the final model were examined. Adjusted Odds ratios (OR) and 95% confidence intervals (CI) were calculated, and α=0.05 was set for statistical significance.

3.3.3 Machine learning models development

An unsupervised model was developed through the clustering technique in order to identify patterns among frequent users. Furthermore, ML classification was employed to evaluate the performance of different learning algorithms and was compared to classical logistic regression. The dataset for ML

classification was randomly split into 70% and 30% of observations for training and test sets, respectively.

3.3.3.1 Clustering

Clustering is the process of grouping data samples together into clusters based on their similarities. The clustering is helpful in population health data when an intervention is supposed to be designed and applied (Wiemken & Kelley, 2019). For example, policymakers may decide to put some patients in clusters with other patients who share similar characteristics on ED utilization. Therefore, the recommended intervention will be somewhat on-point, as current ED users with similar characteristics are likely to be addressed with similar interventions. Moreover, as a new patient visits the ED, that person will be placed within a particular cluster, and the intervention will be applied to them as well. The three clustering algorithms are K-means for numerical data, K-modes for categorical data and K-prototype for mixed data. This study uses K-modes clustering due to the categorical nature of the dataset.

3.3.3.1.1 *K-modes clustering*

K-modes algorithm is used to cluster categorical data, for which it uses *modes* instead of *means*. *K-modes* is an extension of *k-means* which, instead of distances, uses *dissimilarities* (i.e. "quantification of the total mismatches between two objects: Smaller the number of mismatches, more similar the two objects are." (Khan & Ahmad, 2012). The number of modes will be as many as the number of clusters since they act as centroids; the algorithm uses a frequency-based method to update modes in the clustering process to minimize the cost function.

K-modes clustering algorithm was first introduced by Zhexue Huang (Huang, 1998). However, the Huang K-mode algorithm is susceptible to the choice of initial centers; an improper choice may generate objectionable cluster structures. In the Huang algorithm, random initialization is used to choose initial centers due to its simplicity; however, this may lead to non-repeatable clustering results. Therefore, it is not easy to rely on the results obtained, and several re-runs of the K-modes algorithm may be required to arrive at a meaningful conclusion. To address this issue, Cao et al. (Cao et al., 2009) introduced a K-mode algorithm that uses the average density of objects and the distance between objects to initialize cluster centers:

> *"It uses the frequency of categories to define the average density of an object. The first cluster center is formed by selecting the object which has the maximum density. It then extends the MaxMin algorithm in the combination of objects' density and the distance between objects for remaining clusters. The clustering algorithms then use the initialization approach for k-modes."*(Cao et al., 2009)

A drawback for both algorithms, however, is that the number of clusters must be defined manually. Therefore, a try-and-effort process is required to achieve the optimal number of clusters.

In this study, the value of K was set to the numbers ranged from 2 to 5 for selecting the best clustering performance, the more distinct the cluster, the better. The default number of times the K-modes algorithm was run with different centroid seeds was 10, and in our study was set to 20, so the final results will be the best output of 20 consecutive runs in terms of cost. The number of iterations for each run was set to 100 as the default number.

3.3.4 Regression Analysis

3.3.4.1 Logistic regression with Recursive Feature Elimination (RFE)

A regression model using the RFE technique was created to identify the best-fitting regression model by examining all predictor variables that are specified. RFE selects features recursively, considering smaller and smaller sets of variables to comply with the parsimony principle. The performance metric used in this study to evaluate feature performance was p-value; if the p-value was above 0.05, it was removed; else, it was kept in the model.

3.3.5 Classification

3.3.5.1 Random Forest

Random Forest is one of the most flexible and easy-to-use ML algorithms. It creates decision trees on randomly selected data samples drawn from a training set, gets a prediction from each tree and selects the best solution through voting. It is an ensemble method of decision trees generated on a randomly split dataset based on the "divide-and-conquer" concept. Each tree depends on an independent random sample. In a classification problem, each tree "votes" on the outcome when a new example is

introduced, and the most popular class (i.e., the outcome with the majority vote) is the final prediction. Random Forest algorithm has considerable advantages, including but not limited to high accuracy and robustness, resistance to overfitting problem (i.e., it takes an average of all the predictions which cancels out the biases) and providing a good indicator of the feature importance, which helps in selecting the most contributing features for the classifier. Thampi et al. shows the processing steps of the algorithm in Figure 3-2 (Thampi et al., 2013).

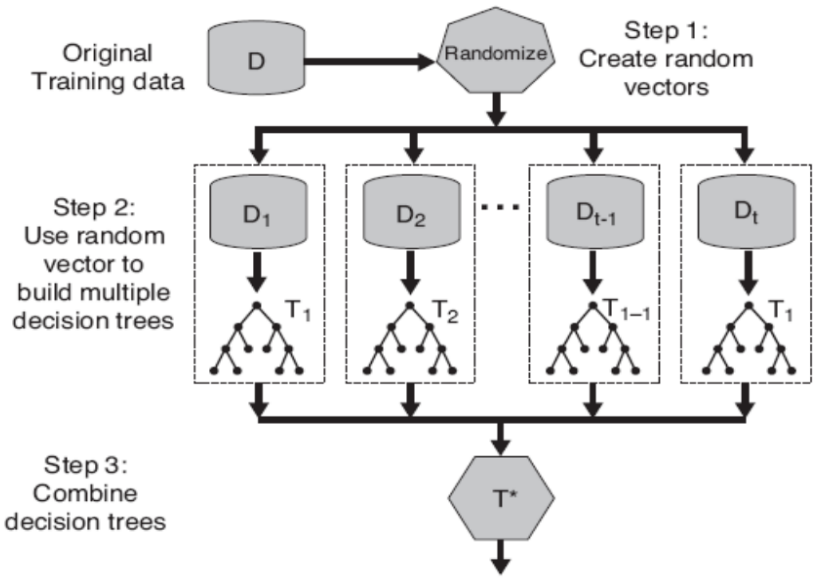


Figure 3-2: Random Forest Flow-Diagram from *"Novel application of Random Forest method in CERES scene type classification"* by Bijoy V. Thampi, Constantine Lukashin and Takmeng Wong, 2013.

3.3.5.2 Support Vector Machine

Support Vector Machine (SVM) is one of the most popular ML algorithms due to its capacity for multi-class classification (Ahmad et al., 2015). Its kernel trick helps build a more accurate classifier by taking a low-dimensional input space and transforming it into a higher-dimensional space. For this study, Radial Basis Function (RBF) kernel was used. RBF is one of the most popular kernels with the gamma parameter, ranging from 0 to 1, that needs to be set in the learning algorithms. A higher value of gamma will cause over-fitting; therefore, the default value, which set gamma to 1/n-features, was used for training. Many prefer SVM because it achieves significant accuracy while using less computing power(Géron, 2017) . However, it is sensitive to data transformation and cannot handle categorical data; therefore, creating dummy variables was required. Figure 3-3 shows the structure of the SVM algorithm and how the choice of kernel affects the accuracy of the algorithm.

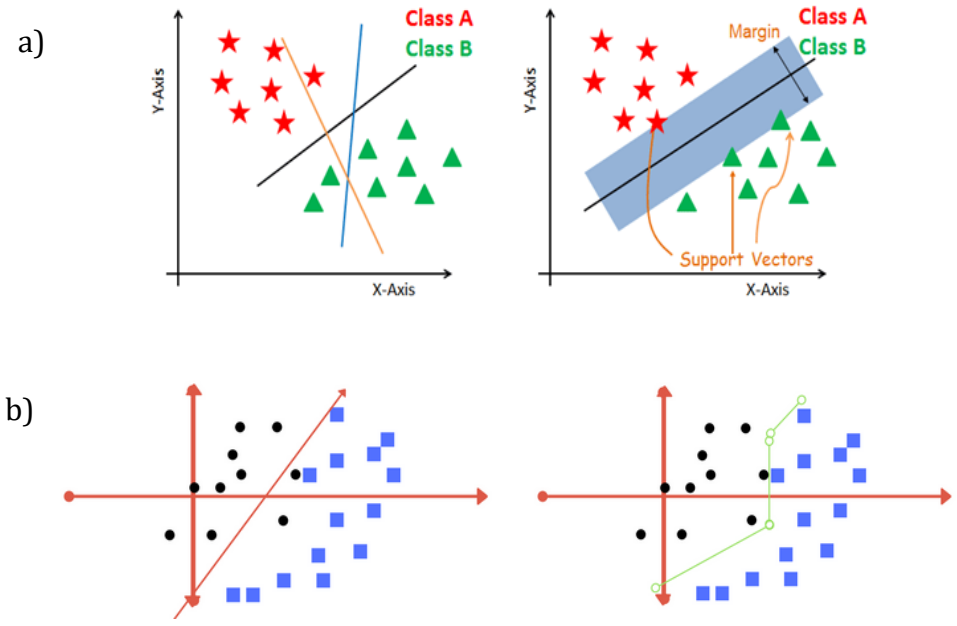


Figure 3-3: a) Support Vector Machine Algorithm Classification Process. **b)** Shows the effect of linear kernel vs Non-linear (e.g., RBF) in classifying data samples. Image downloaded from *https://heartbeat.fritz.ai/understanding-the-mathematics-behind-support-vector-machines-5e20243d64d5*.

3.3.5.3 Ensemble ML algorithms

Ensemble learning methods are "meta-algorithms" that merge a number of machine learning methods into a single predictive model to improve outcome prediction (Smolyakov, 2017). Ensembles ML offer more accuracy than base classifier. In this study, ensemble algorithms are fit to build a predictive model: Bagging and Voting.

3.3.5.4 Bagging

Bagging classifier, acronym of Bootstrap Aggregation, combines multiple learners to reduce the variance of estimates. It uses the bootstrap sampling technique to select "n" observation out of the "n" observation population. The bagging algorithm draw samples from the training set, and -create classifiers using those bootstrap samples. The final prediction is the average of all predictions. K-fold cross-validation is used for the input parameter of the classifier to choose the best sub-model and increase the performance of the final model selection and handling the overfitting problem. Figure 3-4 shows the process flow of the Bagging algorithm.

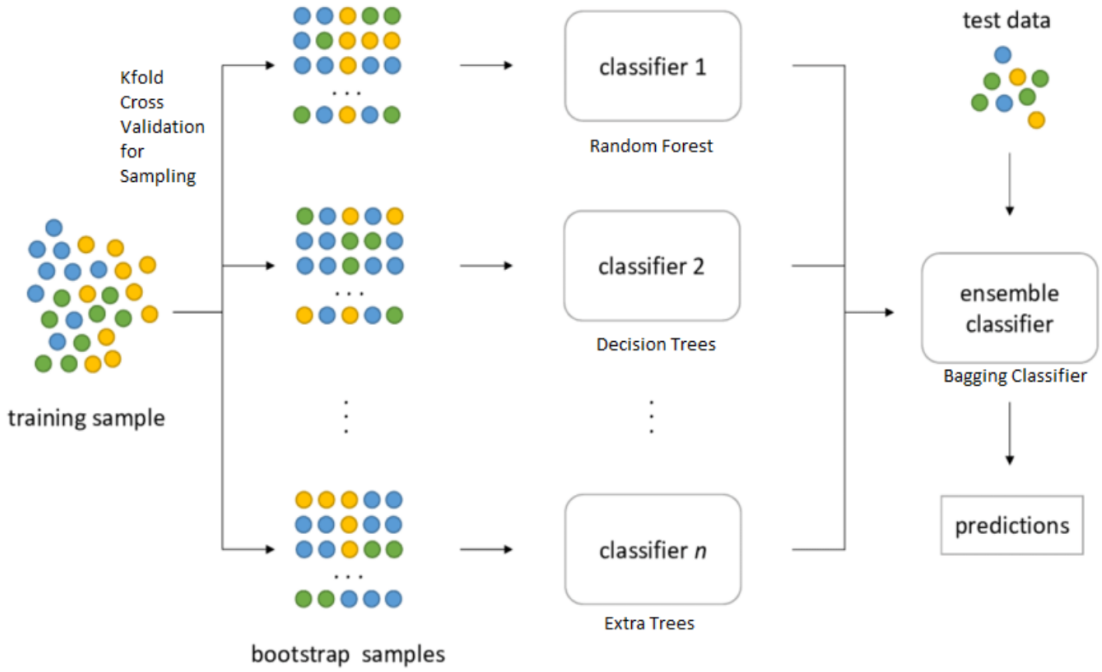


Figure 3-4: Bagging Process Flow. Image downloaded from *https://medium.com/ml-research-lab/bagging-ensemble-meta-algorithm-for-reducing-variance-c98fffa5489f.*

3.3.5.5 Voting

Voting is one of the ensembles learning strategies that aggregate predictions from multiple models. The procedure begins with the development of two or more different models using the same dataset. A voting model may combine the previous models and merge the predictions of those models. The predictions made by base models will be selected in the best possible way, using the *stacked aggregation technique*. In this study, a Voting classifier was created by assembling logistic regression, Classification and Regression Tree (CART)and SVM to train the data. Figure 3-5 shows how Voting take place in the algorithm.

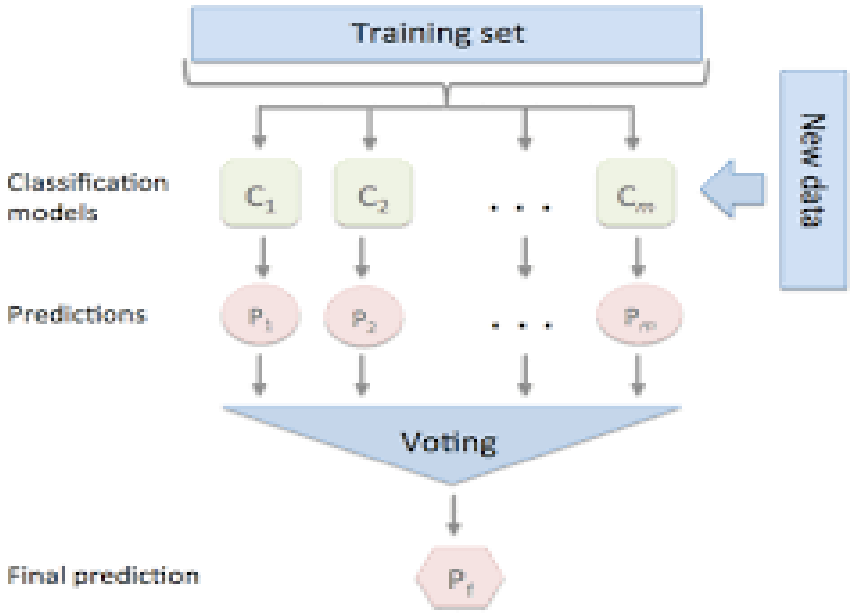


Figure 3-5: Voting Algorithm Process. Image downloaded from *http://rasbt.github.io/mlxtend/user_guide/classifier/EnsembleVoteClassifier/*.

3.4 Performance Evaluation

Models were evaluated on the test set using the standard classification evaluation metrics: (i) AUC, (ii) sensitivity (aka recall or TP/TP+FN), (iii) precision (aka PPV or TP/TP + FP), and (iv) classification error (i.e. the percentage of predictions that are incorrect) (Japkowicz & Shah, 2011). *Classification error* was defined as (1-Accuracy) * 100, where accuracy is the percentage of times the predicted outcome is equal to the observed outcome.

The performance of each model's AUC, sensitivity, and precision were calculated. Interpretation of the evaluated values is as follows: <0.6 for poor, 0.6 - 0.69 for fair, 0.7 - 0.79 for good, 0.8 - 0.89 for very good, and > 0.9 for excellent (Šimundić, 2009). For classification error evaluation the following rubric was also used: 0 - 4% for excellent, 5 -9% for very good, 10-14% for good, 15-20% for fair, and >20% for poor.

3.5 Software

We used Python on Google Colaboratory platform (Google, 2020) with the following libraries for analysis: "numpy", "panda", "scikit-learn", "statsmodels", "matplotlib" and "seaborn". The coding is partially available in Appendix A.

3.6 Ethics

I used anonymized data from KHPS in this thesis. Although Google Colab is a cloud-based platform and subject to some privacy concerns, privacy management settings were set to maximum control, meaning no content and information from Google Colab was allowed to be collected according to Google Privacy Policy. Moreover, information (if any) that could risk participants' anonymity was removed from the applied dataset. In addition, the Ethics application form submitted to the Ethics Board of the University of Saskatchewan indicated the use of Google Colab for data analysis to ensure it is in compliance with ethics requirements; the approval is included in Appendix B.

Chapter 4

Results

4.1 Characteristics of ED users

The distribution of baseline characteristics of ED users showed that non-frequent ED users accounted for 14,178 (90.7%) of total visits (Table 4-1). Of the 9,348 participants included in the current analysis, 220 had made ≥ 4 ED visits and so were categorized as frequent users who contributed to 9.3% of total ED visits in the period of study.

Approximately equal numbers of males and females were non-frequent users while males contributed 61% of frequent users. Patients aged 65 years old or older had the highest proportion of ED use in both frequent and non-frequent groups and were more likely to be the residents in provinces rather than in capital and metropolitan cities. Frequent users were more likely to take an ambulance to EDs than non-frequent users (25% vs 18%). Frequent users were more likely to use NHI compared to non-frequent users (84% vs 75%), while they visited private EDs more than public EDs (87% vs 13%). Frequent users were also less likely to be admitted to the hospital compared to non-frequent users. (27% vs 30%). The highest proportion of ED utilization was seen on Sundays among both frequent (20%) and non-frequent users (21%), followed by Saturdays (15% and 16%, respectively). The proportion of visits in spring and summer was higher than in fall and winter among frequent (10%) and non-frequent (6%) users. Respiratory disease was one of the most common reasons to visit EDs in both groups (27% of frequent ED users vs 15% of non-frequent users). Table 4-1 summarizes the descriptive results of the study.

Table 4-1: Baseline characteristics of emergency department visits between 2008 and 2015 (N=15,627 visits by 9,348 patients)

Variables	Non-frequent visits (n, %) (n=14,178)	Frequent visits (n, %) (n=1,449)	P-Value
Age			
0 -14	3345 (24)	282 (19)	<0.0001
15 -24	1143 (8)	61 (4)	
25 - 44	2624 (19)	171 (12)	
45 - 64	3370 (24)	275 (19)	
65+	3696 (26)	660 (46)	
Sex			
Female	7004 (49)	572 (39)	
Male	7174 (51)	877 (61)	<0.0001
Type of ED			
Private	13045 (92)	1266 (87)	
Public	1133 (8)	183 (13)	<0.0001
Transportation to ED			
Ambulance	2569 (18)	355 (25)	
Self-Transport	11609 (82)	1094 (76)	<0.0001
Insurance Status			
Additional	3544 (25)	226 (16)	
NHI	10634 (75)	1223 (84)	<0.0001
Region			
Capital	1681 (12)	158 (11)	
Metropolitan city	3834 (27)	311 (21)	
Province	8663 (61)	980 (68)	0.039
Hospitalization			
Admitted	4189 (30)	398 (27)	
Discharge	9989 (70)	1051 (73)	0.097
Season of Visit			
Spring	3638 (26)	403 (28)	0.017
Summer	3638 (26)	403 (28)	
Fall	3515 (25)	323 (22)	
Winter	3387 (24)	320 (22)	
Day of Visit			
Monday	1976 (14)	224 (15)	
Tuesday	1712 (12)	211 (15)	
Wednesday	1706 (12)	184 (13)	
Thursday	1771 (12)	162 (11)	0.038
Day of Visit (cont.)			
Saturday	2237 (16)	212 (15)	
Sunday	3032 (21)	292 (20)	
Friday	1744 (12)	164 (11)	

Reason for Visit			
Respiratory System	2092 (15)	396 (27)	<0.0001
Unclassified clinical finding	2015 (14)	218 (15)	
Digestive System	1732 (12)	146 (10)	
Damage (e.g., injuries)	2787 (20)	134 (9)	
Neoplasm	366 (3)	92 (6)	
Circulatory System	779 (5)	89 (6)	
Nervous System	372 (3)	67 (5)	
Fracture	1264 (9)	65 (4)	
Endocrine and Metabolic	149 (1)	54 (4)	
Musculoskeletal System	479 (3)	50 (3)	
Infectious and Parasitic	683 (5)	36 (2)	
Genitourinary tract	482 (3)	35 (2)	
Mental and behavioral disorder	92 (0.7)	23 (2)	
Skin and skin underlying	403 (3)	17 (2)	
Eye and eye attachment	52 (0.4)	7 (0.5)	
Diseases of Ear	150 (1)	6 (0.4)	
Blood and hematopoietic disorder	22 (0.2)	5 (0.4)	
Other Damage/Poison	143 (1)	3 (0.2)	
Pregnancy, childbirth, maternity	64 (0.5)	3 (0.2)	
Other morbidities	52 (0.4)	3 (0.2)	

ED users who were children up to 14 years old were primarily male (57%), whereas females contributed more to ED visits in the age range from 25 to 65 (Figure 4-1). Respiratory system complaints were one of the top reasons to visit EDs among both very young (under 14 years old) and very old (over 65 years old) patients (Figure 4-2). Spring and winter were the seasons with the highest number of visits due to respiratory system diagnosis (Figure 4-3). Most of the visits due to damage (e.g., injuries, poisoning, etc.) took place in summer, and the number was higher in children under 14 and adults between 45-64 years old.

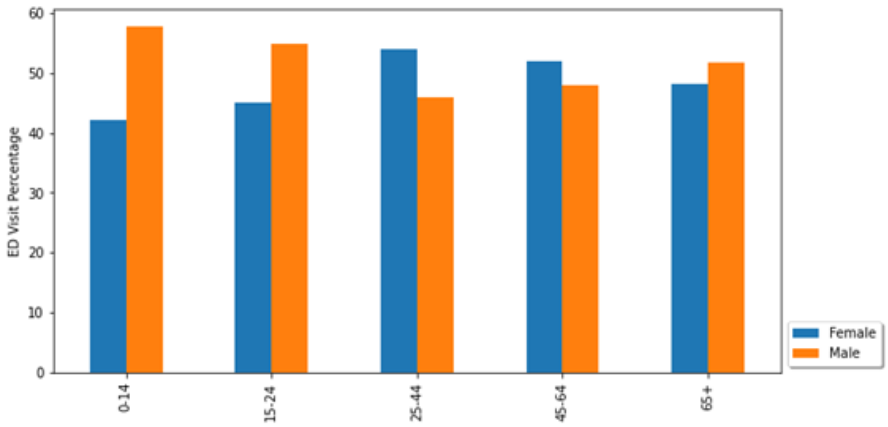


Figure 4-1: Emergency department visits by sex within age groups

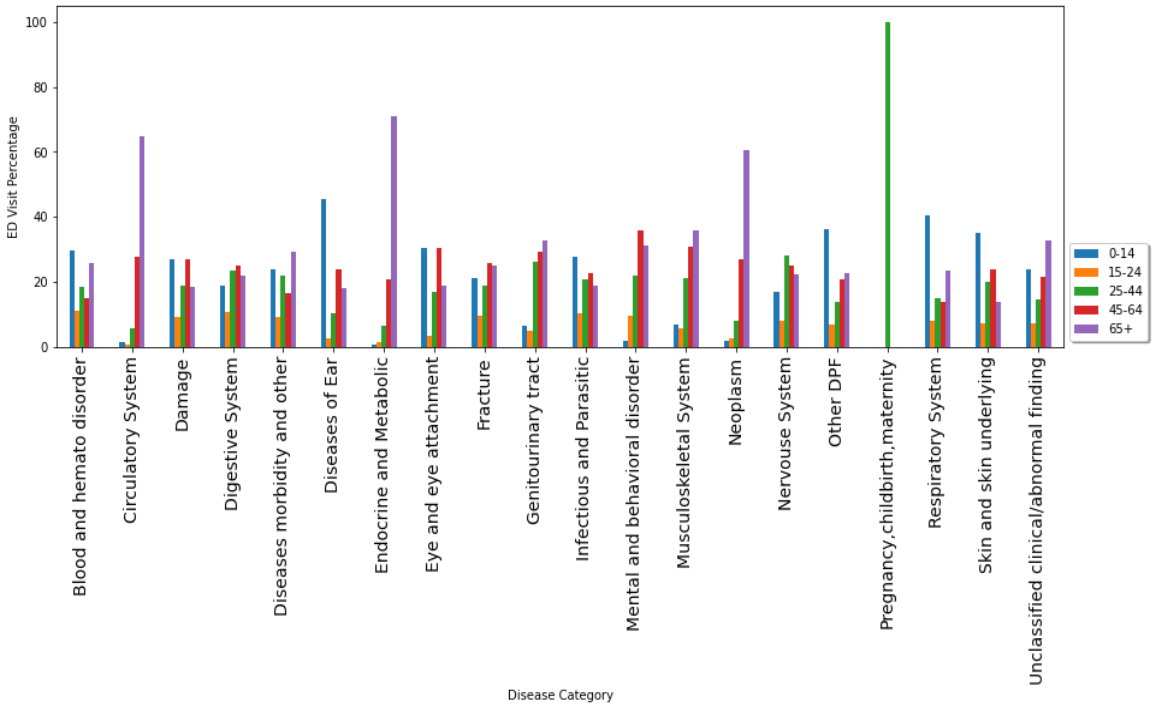


Figure 4-2: Emergency department visits by age groups within disease categories

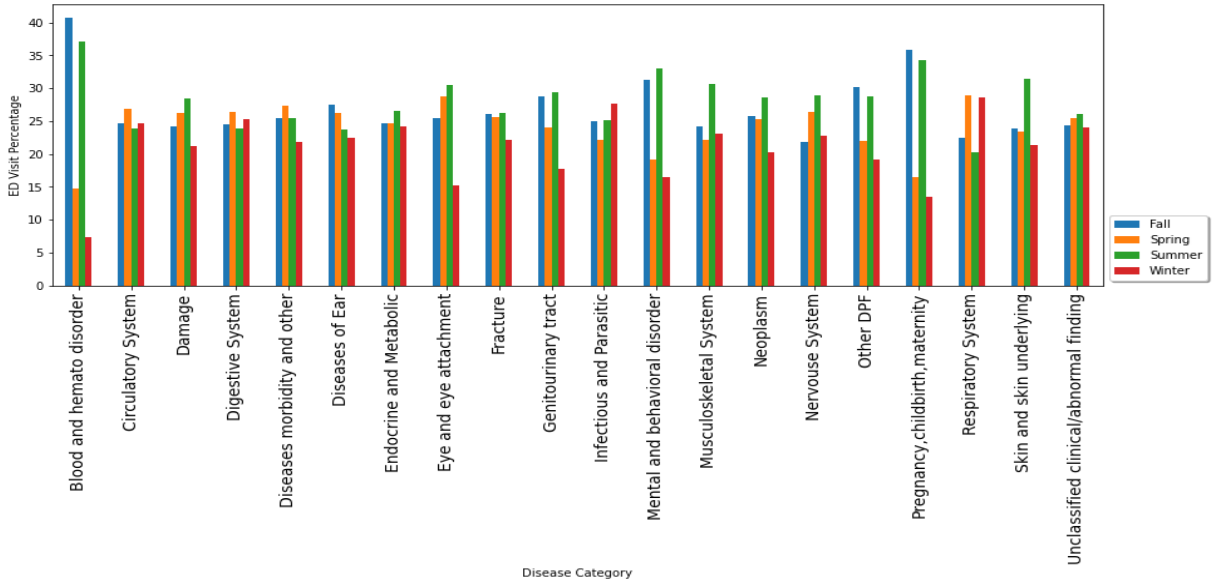


Figure 4-3: Emergency department visits by season of the visit within disease categories

Male users contributed more than females in categories such as damage (56.5%), respiratory system (59.1%), fracture (60.1%), and circulatory system (52.1%). In contrast, females accounted for more visits than males due to the diagnosis such as digestive system complaints (51%), infectious and parasitic diseases (55.6%), mental and behavior disorder (64.8%), nervous system (64.9%), and unclassified clinical/abnormal findings (53.4%) (Figure 4-4).

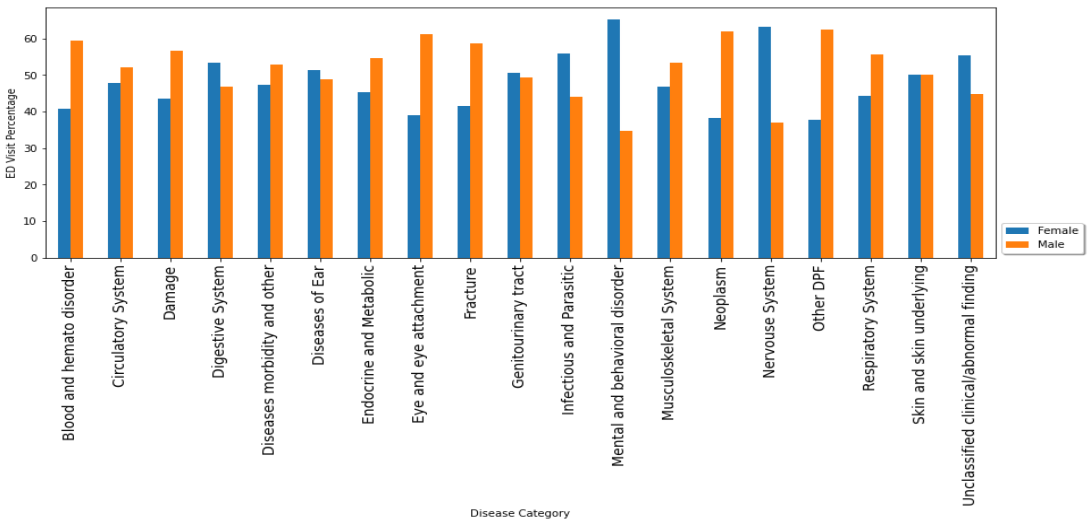


Figure 4-4: Emergency department visits by sex within disease categories

4.2 Univariate Analysis

Univariate logistic regression showed age, sex, reason, season, and day of the visit, as well as region, transportation, type of ED, and insurance status, were associated with ED frequent use. Patients 65+ were about two times more likely to be frequent users than those under 14 years old. Males were 1.5 times more likely to visit EDs than females continually. Those who used self-transportation were less likely to be frequent users than those who took an ambulance to the hospital (OR = 0.68; 95% CI: -0.51 – -0.26; p <0.0001), and ED users with NHI were 1.80 times more likely to be frequent users (OR = 1.8; 095% CI: 1.56 – 2.09; p <0.0001). Frequent users were 2.68 times more likely to visit EDs due to endocrine and metabolic conditions than respiratory system diagnosis, and they had higher odds of living in the provinces than the capital (OR=1.20; 95% CI:1.01 – 1.44; p=0.039).

Hospitalization was not statistically significant and therefore was not included in the multivariate model. Table 4-2 summarizes the analysis results.

Table 4-2: Univariate logistic regression analysis of emergency department frequent visits with odds ratio and 95% confidence interval. (N=9,348 patients)

Covariate	Odds Ratio (95% C.I.)	P-value
Age		
0-14	1.00	
15-24	0.59 (0.44 - 0.79)	<0.0001
25-44	0.73 (0.60 - 0.89)	0.002
45-64	0.95 (0.80 - 1.13)	0.548
65+	2.03 (1.76 - 2.35)	<0.0001
Sex		
Female	1.00	
Male	1.50 (1.34 - 1.67)	<0.0001
Type of ED		
Private	1.00	
Public	1.66 (1.41 - 1.97)	<0.0001
Transportation to ED		
Ambulance	1.00	
Self-Transport	0.68 (-0.51 – - 0.26)	<0.0001
Insurance Status		
Additional	1.00	
NHI	1.80 (1.56 - 2.09)	<0.0001
Region		
Capital	1.00	
Metropolitan city	0.86 (0.71 - 1.05)	0.147
Province	1.20 (1.01 - 1.44)	0.039
Hospitalization		
Admitted	1.00	
Discharge	1.11 (0.98 - 1.25)	0.097
Season of Visit		
Winter	1.00	
Spring	1.21 (1.03 - 1.41)	0.017
Summer	1.21 (1.03 - 1.41)	0.017
Fall	1.03 (0.88 - 1.21)	0.729
Day of Visit		
Monday	1.00	
Tuesday	1.08 (0.88 - 1.31)	0.469
Wednesday	0.96 (0.78 - 1.17)	0.664
Thursday	0.80 (0.65 - 0.99)	0.038

Saturday	0.82 (0.67 - 1.02)	0.072
Sunday	0.84 (0.69 -1.02)	0.081
Friday	0.85 (0.71 - 1.02)	0.076
Reason for Visit		
Respiratory System	1.00	
Unclassified clinical finding	0.71 (0.59 - 0.85)	<0.0001
Digestive System	0.50 (0.40 - 0.61)	<0.0001
Damage (e.g., injuries)	0.34 (0.28 - 0.42)	<0.0001
Neoplasm	1.76 (1.37 - 2.25)	<0.0001
Circulatory System	0.68 (0.53 - 0.88)	0.003
Nervous System	1.25 (0.95 - 1.65)	0.108
Fracture	0.29 (0.21 - 0.38)	<0.0001
Endocrine and Metabolic	2.68 (1.95 - 3.70)	<0.0001
Musculoskeletal System	0.67 (0.49 - 0.91)	0.011
Infectious and Parasitic	0.31 (0.22 - 0.45)	<0.0001
Genitourinary tract	0.47 (0.33 - 0.67)	<0.0001
Mental and behavioral disorder	1.46 (0.90 - 2.36)	0.123
Skin and skin underlying	0.34 (0.22 - 0.54)	<0.0001
Eye and eye attachment	0.99 (0.46 - 2.10)	0.972
Diseases of Ear	0.57 (0.32 - 1.02)	0.058
Blood and hematopoietic disorder	1.09 (0.37 - 3.17)	0.877
Other Damage/Poison	0.52 (0.28 - 0.97)	0.041
Pregnancy, childbirth, maternity	0.78 (0.36 - 1.69)	0.524
Diseases morbidity and other	0.83 (0.36 - 1.91)	0.664

4.3 Multivariable Analysis

Multivariable logistic regression analysis showed that sex, age, insurance, day, season, reason of the visit, and the type of ED were associated with frequent users. Also, male by age interaction was significant in the multivariable model.

The odds ratios of main effects not involved in the interaction terms were extracted directly from the full model table and for the covariates in the interaction terms were calculated separately. The full model is presented below.

The odds of frequent visits to ED were higher for patients with only NHI than those with private insurance (OR=1.63; 95% CI: 1.46 – 1.98; p <0.0001). Frequent ED users were 78% more likely to visit EDs due to endocrine and metabolic complaints than due to respiratory system diagnosis (OR=1.78; 95% CI: 1.26 – 2.49; p <0.0001). They were also 1.26 times more likely to visit EDs in summer than in winter.

$$Logit(p) = \beta 0 + \beta 1 Age2 + \beta_2 Age3 + \beta_3 Age4 + \beta_4 Age5 + \beta_6 Sex + \beta_7 EDType + \beta_8 Ins +$$

$$\beta_9 Winter + \beta_{10} Spring + \beta_{11} Summer + \beta_{12} Fall + \boxed{\beta_{13} + \cdots + \beta_{18}} + \boxed{\beta_{20} + \cdots + \beta_{42}} + \beta_{43} Sex *$$

$$Age2 + \beta_{44} Sex * Age3 + \beta_{45} Sex * Age4 + \beta_{46} Sex * Age5 + \beta_{47} Province + \beta_{48} MetroCity +$$

$$\beta_{49} Transport$$

Day

Diseases

The interaction between sex and age found to be statistically significant. Several odds ratios could be calculated for the interaction. The OR for patients 65+ compared to youngest group (i.e., under 14) among males and the OR of males compared to females among age 65+ were calculated after controlling for the other covariates. The results showed that age was a significant risk factor among males; those in 65+ group were more likely to become frequent ED users compared to males of the younger age with an estimated odds ratio of 3.01 (95% CI: 1.83 – 5.09; $p<0.0001$). The OR for the other age groups compared to the younger age among males found to be not statistically significant. From the other hand, analysis showed that 65+ males were also more likely to become frequent users than 65+ females (OR=2.91; 95% CI: 1.89– 4.84; $p<0.05$). The odds ratio for males vs females among each age category was calculated separately; the results showed that the likelihood of becoming frequent ED user was not statistically significantly different between males and females among other age groups. The other examined interactions were not significant.

In general, frequent users were more likely to have diseases related to neoplasm, nervous system, and endocrine and metabolic complaints than respiratory system complaints. However, they were less likely to visit ED due to complaints such as fracture, digestive system disease, and circulatory system complications. They were also more likely to visit in relatively warm seasons (i.e., spring and summer), and compared to non-frequent users, they were less likely to visit on weekends.

Table 4-3 summarizes the results of the final multivariable logistic model, and Table 4-4 shows the significant interaction effects between sex and age.

Table 4-3: Multivariate analysis of emergency department frequent visits with odds ratio and 95% confidence interval. (N=9,348 participants)

Covariate	Estimates	SE	Odds Ratio (95% C.I.*)	P-value
Insurance Status				
Private			1.00	
NHI	0.54	0.07	1.71 (1.46 -1.98)	0.420
Type of ED				
Public			1.00	
Private	0.38	0.08	1.49 (1.26 - 1.77)	<0.0001
Region				
Capital			1.00	
Metropolitan	-0.31	0.10	0.72 (0.58 - 0.88)	0.003
Provinces	-0.08	0.09	1.08 (0.86 – 1.23)	0.369
Mode of Transport				
Self-transport				
Ambulance	0.20	0.07	1.22 (1.06-1.41)	0.005
Reason for Visit				
Respiratory System			1.00	
Unclassified clinical finding	-0.40	0.09	0.63 (0.52 - 0.76)	<0.0001
Digestive System	-0.56	0.10	0.57 (0.46 - 0.70)	<0.0001
Damage (e.g., injuries)	-1.04	0.10	0.36 (0.29 - 0.44)	<0.0001
Neoplasm	0.31	0.13	1.31 (1.09 - 1.85)	0.008
Circulatory System	-0.52	0.13	0.55 (0.42 - 0.71)	<0.0001
Nervous System	0.37	0.14	1.43 (1.08 - 1.90)	0.012
Fracture	-0.96	0.14	0.35 (0.26 - 0.46)	<0.0001
Endocrine and Metabolic	0.58	0.17	1.78 (1.26 - 2.49)	0.001
Musculoskeletal System	-0.29	0.15	0.70 (0.51 - 0.95)	0.023
Infectious and Parasitic	-0.80	0.17	0.41 (0.29 - 0.58)	<0.0001
Genitourinary tract	-0.69	0.17	0.51 (0.36 - 0.72)	<0.0001
Mental and behavioral disorder	0.33	0.25	1.35 (0.81 - 2.24)	0.248
Skin and skin underlying	-0.74	0.21	0.45 (0.29 - 0.68)	<0.0001
Eye and eye attachment	0.01	0.39	0.98 (0.45 - 2.12)	0.966
Diseases of Ear	-0.36	0.28	0.66 (0.37 - 1.16)	0.153
Blood and hemato disorder	0.08	0.57	1.06 (0.33 - 3.35)	0.920
Other Damage/Poison	-0.40	0.30	0.63 (0.34 - 1.15)	0.133
Pregnancy, childbirth, maternity	-0.07	0.45	0.93 (0.36 - 2.37)	0.883
Diseases morbidity and other	-0.10	0.43	0.88 (0.38 - 2.04)	0.769
Day of Visit				
Monday			1.00	
Tuesday	0.04	0.10	0.95 (0.82 - 1.23)	0.653
Wednesday	-0.19	0.10	0.82 (0.70 - 1.08)	0.72
Thursday	-0.34	0.11	0.71 (0.58 - 0.89)	0.002
Friday	-0.28	0.11	0.75 (0.64 - 0.92)	0.009
Saturday	-0.25	0.10	0.77 (0.61 - 0.93)	0.014
Sunday	-0.24	0.09	0.78 (0.64 - 0.95)	0.009

Season of Visit				
Winter			1.00	
Spring	0.31	0.08	1.37 (1.17 - 1.60)	0.005
Summer	0.13	0.08	1.14 (1.07 - 1.47)	0.003
Fall	0.11	0.09	1.11 (0.91 - 1.27)	0.253

Table 4-4: Significant interaction effect between sex and age

	Odds Ratio (95% C.I.*)	**P-value**
Among males		
15-24 vs 0-14 years	0.59 (0.21– 1.49)	0.771
25- 44 vs 0-14 years	0.92 (0.44 – 1.56)	0.423
44-64 vs 0-14 years	0.86 (0.48 – 1.46)	0.235
65+ vs 0-14 years	3.01 (1.83 – 5.09)	<0.0001
Among females		
15-24 vs 0-14 years	0.64 (0.42 – 0.97)	0.038
25- 44 vs 0-14 years	0.78 (0.59 – 1.04)	0.088
44-64 vs 0-14 years	1.07(0.84 – 1.38)	0.559
65+ vs 0-14 years	0.92 (0.72 – 1.19)	0.555

4.4 Clustering results for patterns identification

The optimal k was determined by comparing costs against each k between 2 and 5, inclusive. The optimal K was chosen by comparing costs against each k and as shown Figure 12, k=3 was the optimal number of clusters for our dataset. The three-cluster model (A=0, B=1, and C=2) for frequent ED users, is shown in Figure 4-5 clear boundaries.

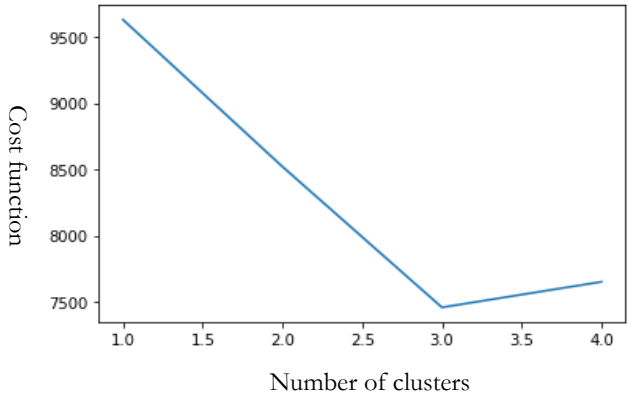


Figure 4-5: Optimal number of k based on cost function, i.e., the dissimilarity rate for the clustering.

Characteristics of the frequent ED users who belonged to each cluster are presented in Table 4-5**.** In short, patients who belonged to cluster A as light pink in Figure 4-6, were older, male, and visited ED for respiratory system complaints. Discharged rates were the highest (93.3%) among patients who belonged to this cluster. They were also more likely to visit ED on Sundays. Hospitalization rates (98.7%) were the highest among the patients who belonged to cluster B as dark pink in Figure 4-6, who were also more likely to used ambulance (60.4%). They were 65+, male and their main reason to visit ED was circulatory system complaints. Patients who belonged to cluster C as purple in Figure 4-6, were younger, female, and more likely to visit ED in summer and on weekends. They were less likely to have used an ambulance to visit the ED (7.9%), more likely to have been discharged, and their main complaints belonged to the damage category, including but not limited to injuries and poisoning.

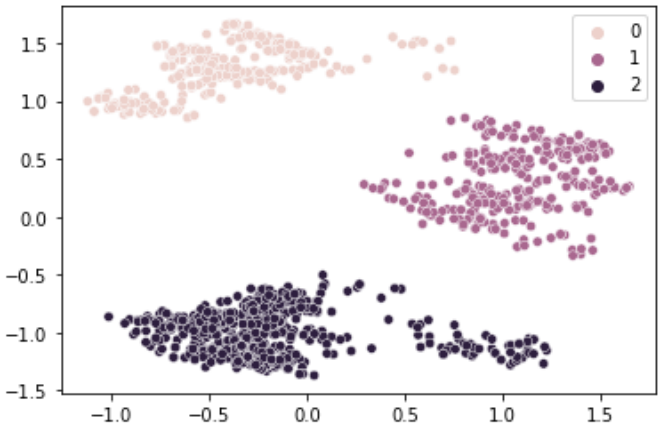


Figure 4-6: K-modes clustering indicate that frequent emergency department users can be clustered into 3 clusters with relatively clear boundaries.

Analysis of chi-square test showed that there was not a statistically significant difference among clusters with respect to the type of ED ($\chi 2$ (2) = 1.28; p-value= 0.526). Table 4-5 reports the p-values associated with each characteristic; distribution of age, sex, mode of transport, insurance, hospitalization, region, time of visits (i.e., day and season), and the reason of visits was different across clusters.

Table 4-5: Characteristics of frequent emergency department users and their subgroups

Features	All N=1449	Cluster A n= 460	Cluster B n=308	Cluster C n=681	p-value
Age					
0-14	282 (19.5)	63 (13.7)	18 (5.9)	201 (29.5)	<0.0001
15-24	61(4.2)	9 (1.9)	6 (1.9)	46 (6.7)	<0.0001
25-44	171 (11.8)	7 (1.5)	29 (9.4)	135 (19.8)	<0.0001
45-64	275 (18.9)	6 (1.3)	52 (16.9)	217 (31.9)	<0.0001
65+	660 (45.6)	375 (81.6)	203 (65.9)	82 (12.1)	<0.0001
Sex					
Male	877 (60.5)	443 (96.3)	215 (69.8)	219 (32.2)	<0.0001
Type of ED					
Private	1266 (87.4)	377 (81.9)	266 (86.4)	623 (91.5)	0.526
Mode of transport					
Ambulance	355 (24.5)	115 (25.0)	186 (60.4)	54 (7.9)	<0.0001
Type of insurance					
NHI	1223 (84.4)	428 (93.0)	216 (70.1)	579 (85.0)	<0.0001
Region					
Province	980 (67.6)	305 (66.4)	219 (71.1)	456 (67.0)	0.561
Metropolitan City	311 (21.5)	124 (26.9)	56 (18.2)	131 (19.2)	0.003
Capital	158 (10.9)	31 (6.7)	33 (10.7)	94 (13.8)	0.002
Hospitalization					
Admitted	391 (27.0)	30 (6.5)	304 (98.7)	57 (8.4)	<0.0001
Discharged	1053 (72.7)	429 (93.3)	0	624 (91.6)	<0.0001
In-ED Death	5 (0.3)	1 (0.2)	4 (1.3)	0	0.006
Season of visit					
Fall	320 (22.1)	84 (18.3)	74 (24.0)	162 (23.8)	0.111
Spring	439 (30.3)	157 (34.1)	95 (30.8)	187 (27.5)	0.346
Summer	365 (25.2)	94 (20.4)	74 (24.1)	197 (28.9)	0.013
Winter	325 (22.4)	125 (27.2)	65 (21.1)	135 (19.8)	0.034
Day of visit					
Monday	224 (15.4)	52 (11.3)	68 (22.1)	104 (15.3)	0.002
Tuesday	211 (14.7)	66 (14.4)	52 (16.9)	93 (13.7)	0.331
Wednesday	184 (12.7)	55 (11.9)	41 (13.3)	88 (12.9)	0.967
Thursday	162 (11.2)	45 (9.8)	40 (13.0)	77 (11.3)	0.128
Friday	164 (11.3)	50 (10.8)	38 (12.3)	76 (11.2)	0.897
Saturday	212 (14.6)	64 (14.0)	36 (11.7)	112 (16.4)	0.198
Sunday	292 (20.1)	128 (27.8)	33 (10.7)	131 (19.2)	<0.0001
Selected reasons for visit					
Respiratory System	396 (27.3)	281 (61.1)	23 (7.7)	92 (13.5)	<0.0001
Digestive System	146 (10.1)	6 (1.3)	46 (14.9)	94 (13.8)	<0.0001
Damage (e.g. injuries)	134 (9.2)	2 (0.4)	18 (5.8)	114 (16.7)	<0.0001
Circulatory System	89 (6.1)	5 (1.1)	52 (16.9)	32 (4.7)	<0.0001
Fracture	65 (4.5)	6 (1.3)	25 (8.1)	34 (5.0)	<0.0001
Mental health and behavioral disorder	23 (1.6)	0	4 (1.3)	19 (2.8)	0.001

4.5 Classification results for ML predictive models

All the classification algorithms predicting frequent ED users showed adequate discriminating power from a very good AUC of 0.89 for SVM to an excellent AUC of 0.96 for Random Forest. Figure 4-7 shows the results of each learning model.

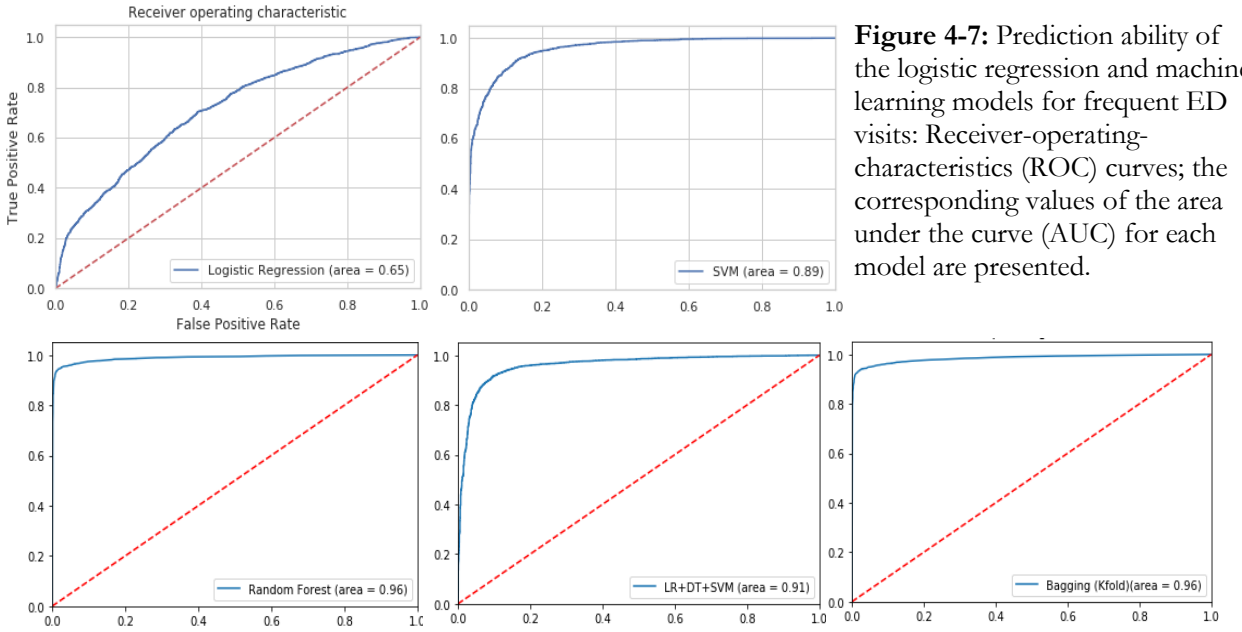


Figure 4-7: Prediction ability of the logistic regression and machine learning models for frequent ED visits: Receiver-operating-characteristics (ROC) curves; the corresponding values of the area under the curve (AUC) for each model are presented.

The best logistic regression model with RFE for training included age, sex, insurance status, season, day, residential region, type of ED and reason for the visit. This model underperformed the machine learning classification algorithms for our data (AUC = 0.65 vs. AUC = 0.89 – 0.96; classification error = 34.9% vs. classification error = 3.8% –11.8%). It worth mentioning that ML algorithms evaluated in this study including SVM with nonlinear kernel support automatic feature interactions.

Random Forest performed best among the machine learning algorithms, with an AUC indicating excellent predictability (0.96) and excellent classification error (3.8%). Table 4-6 summarizes the accuracy of each model.

Table 4-6: Evaluation of predictive models on a test set of 5,952 visits (30% of total data of 19,840: the length of the resampled data with SMOTE technique used to balance original data for more reliable and accurate result; See section 3.3.3)

Models	Classification error (%)	Sensitivity	Precision	Area Under Curve (AUC)
Logistic regression	34.92%	0.67	0.65	0.65
SVM	11.37%	0.87	0.90	0.89
Random Forest	3.77%	0.95	0.98	0.96
Bagging	4.34%	0.94	0.97	0.96
Voting	9.18%	0.91	0.91	0.91

Random Forest with the smallest classification error and highest precision and sensitivity was chosen for further analysis. The Random Forest algorithm has built-in feature importance that use Gini importance to measure how each feature decreases the impurity of a node; the average decrease will be calculated from all the trees in the forest; the higher the impurity decrease, the more important the feature. Figure 4-8 shows the most important features contributing to frequent ED user prediction, according to the Random Forest feature importance score. The top five important features were disease category, age, day of the week, season, and sex.

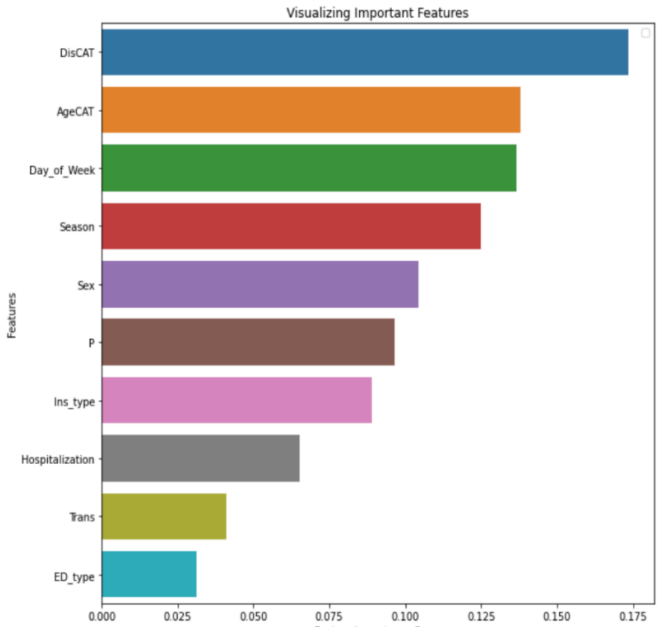


Figure 4-8: Random Forest features importance score based on built-in impurity measure.

Chapter 5

Discussion

This thesis aimed to explore demographic and clinical characteristics of ED patients and frequent ED users in the Korean general population using KHPS, to identify frequent ED utilization pattern, and to evaluate the predictive power of the machine learning techniques in comparison to logistic regression.

We found that frequent ED users accounted for 2.4% of all Korean ED patients and 9.3% of all ED visits, which is 69% and 48% lower than Canada (30%) and Australia (18%), respectively. Both countries are members of the OECD with a similar universal healthcare system to South Korea. A potential reason for such difference could be the high number of 'non-urgent' visit to ED in Canada and Australia (Berchet, 2015), who could have been treated in alternative settings such as primary care, whereas in Korea the accessibility to care even in remote areas through public health centers make the "inappropriate' visits to ED less frequent (*OECD Reviews of Public Health: Korea*, 2020). Our study showed that frequent users of EDs were more likely to be male, very old adults, use public insurance, visit more in summer, and visit more due to diseases such as respiratory system diseases than damages such as injuries or poisoning. This is similar to the findings of other studies (Krieg et al., 2016; Seo et al., 2018; Woo et al., 2016).

Although males made more frequent ED visits than females, females were more likely to visit EDs due to mental health issues. Older adults (age 65+) also visited EDs more frequently than any other age categories. These results are similar to those of Woo J. et al., who have used the Korean National Health Insurance data (Woo et al., 2016). The odds of patients being frequent users were lower for patients living in the metropolitan areas than for those living in Capital. This could imply easier access to ED in the Capital city.

Our study showed that the season and day of visit were significantly associated with frequent ED visits, which was a unique finding; to the best of my knowledge, other studies did not report these associations. However, some studies have investigated the seasonal and weekly change in the total number of ED visits confirming that daily demand for ED services is affected by seasonal and weekly pattern (Jaroudi et al., 2019; Jones et al., 2008). There are several potential reasons for the seasonal or daily pattern of frequent use of ED. For example, the higher rate of frequent visits to ED during Winter or Spring could be due to the potential risk of respiratory system complaints caused by seasonal flu, Pneumonia, allergies, etc. In summer, an increase in outside activities could cause a higher risk of injuries. However, further investigation is required to explore the reasons for such patterns among Korean frequent ED users. Another potential area of investigation is the time of ED visits by the frequent users, which was not available in our database. Analysis of ED visit time among frequent users could further clarify the reason for frequent use among our target population. For example, the time pattern could confirm if the reason is urgent or it is more likely due to the inadequate access to primary care, easy access to ED or the lack of health literacy. Our analysis, similar to other studies (Batal et al., 2001; Jaroudi et al., 2019; Marcilio et al., 2013), showed that calendar variable can be used for developing preventive strategies as well as planning of resources to reduce the burden on ED settings.

Other studies found that an increase in the frequency of ED visits is significantly correlated with a mental health diagnosis(Krieg et al., 2016; Soril et al., 2016). However, our study did not show mental health and behavioral disorder were a significant reason for frequent visits to EDs. Similar to the Chan et al. study in Singapore(Chen et al., 2013), we found that using an ambulance as the means of transport was associated with frequent visits to ED. The reason(s) need to be investigated in further studies; however, we can look at the fact that in South Korea ambulance respond to the centralized number cover "all prehospital transport, free of charge, including basic life support (BLS), intravenous (IV) access, and endotracheal intubation" (Lee et al., 2015). Therefore, further analysis is required to determine what proportion of the ambulance utilization was for urgent need. Other risk factors such as psychosocial characteristics, which was associated with frequent ED users in other studies, were not included in our analysis; therefore, further research is required to investigate the role of these factors in the frequent use of EDs among the Korean population.

Using cluster analysis, frequent ED users were categorized into three meaningful clusters:

A. older patients with respiratory system complaints, the highest discharged rates who were more likely to visit in Winter.

B. older patients with the highest rate of hospitalization, who are also more likely to have used ambulance, and visited ED due to circulatory system complaints.

C. younger patients, mostly female, with the highest rate of visits in summer and lowest rate of using an ambulance, who visited ED mostly due to damages such as injuries, poisoning, etc.

The results of the cluster analysis supported the findings in the descriptive analysis: Patients with circulatory disease (i.e., cluster B) which were attributable to the highest hospitalization rates within frequent users were older compared to the other clusters; NHI was widely used across the cluster by all patients; Private ED was the favorite point of care of patients across the groups, and most of ED frequent user were from provinces. The lowest hospitalization rates emerged for patients with respiratory system complaints (i.e., patients who belonged to cluster A), the oldest group of patients. However, it is worth mentioning that this finding is based on pre-COVID-19, and re-evaluation with post-COVID-19 is advised. In our study, there was no association between frequent users and hospitalization rate. However, a previous study in South Korea showed that frequent ED use was associated with higher hospitalization rate (Woo et al., 2016). These discrepant findings could be due to the number of participants (i.e. 256,246 vs. 9348) and the datasets used in each study. For example, National Health Insurance data could provide richer data on hospital admissions, whereas administrative data tend to be biased concerning the information provided by the participant and interviewers.

A thorough analysis of the status quo is required to design appropriate interventions to improve emergency medical services. In that regard, our investigation results can be utilized as baseline information for future research. For instance, patients with mental health and behavioral disorder mostly belonged to cluster C. Considering they are also female and visited ED due to injuries, poisoning, etc., we could hypothesis that mental health is attributable to frequent visits in this cluster. A detailed study concerning the connection between these factors is required to yield an appropriate intervention to reduce the frequent ED visits and increase the well-being of these patients. In addition, investigating the efficacy of interventional strategies to reduce non-urgent ED visits (e.g., increasing the health literacy to avoid using EDs for non-urgent reasons) could be studied in future research.

Our study also evaluated the performance of four ML classification algorithms and compared them to logistics regression for predicting frequent ED utilization among the Korean population. Random Forest was the best performing method with the highest precision and lower classification error. Other machine learning algorithms also outperformed logistic regression for predicting frequent ED users.

Our machine learning classification models with all variables showed higher accuracy than logistic regression. Interestingly, it is different from the reports in previous studies discussed in the Chapter 2. One possible explanation could be due to the different nature of our data compared to other studies (administrative vs. hospital/EMR data). Also, the data available for this study was not as large and was not affected by outliers or a high number of missing values. A future study with a larger dataset with a higher possibility of outliers and missing values would be required to verify the findings.

Our results have implications for emergency department practices. By more accurate predictions of future ED visits, there is a great potential to reduce preventable and non-urgent ED visits by designing appropriate interventions based on the algorithms' predictions. For example, an increase in the number of visits due to injuries during summer is a potential focus area to design appropriate interventions. Other factors, such as access to care, may also be manageable in health services interventions such as care management.

Strength

Our study has its strengths. First, although several studies have assessed frequent use of EDs, their utility has been limited because most were not population-based or nationally representative. Many studies often included only a subgroup of ED-based data or data only from one medical center. However, our study had participants from a nationally representative sample of the Korean population and was designed to account for differences in the likelihood of selection and differential response rates. Second, our data were collected over a considerable period of time (8 years), with seasonal data reflecting the seasonal ED visits pattern. A study duration less than one year may not reflect patterns of frequent seasonal ED use. The 8-year period of our study included both summer and winter months, which are usually considered the busiest months of the year. Third, our data used medical records and prescriptions, which reduced the recall bias usually associated with self-reported data. Finally, we trained two machine learning classification algorithms (i.e., Bagging and Voting), which

were not previously used in other studies. Both algorithms showed considerably improved performance compared to Adaboost and/or CART used in other studies.

Limitations

This study has limitations. First, there could be some critical missing records, which might not be retrievable or obtainable. Second, our study samples were from medical data, and all patients received some service at ED. However, there would be a small number of patients who visited ED but left without receiving any services, which might cause misclassification of an ED user and/or frequent user categories. Third, while there are some advantages to using administrative data, there is limited information on patients' clinical factors. Clinical factors may be crucial to better understanding the patient's health utilization patterns. Fourth, in this study some socioeconomic factors such as income and occupation were not included, which could be confounders and influence the explanation of the outcome; therefore, additional analysis with the inclusion of socioeconomic factors would further explain the association between the outcome and the risk factors. Finally, our dataset was relatively small for ML application, and our variables were categorical and hence the high accuracy in our ML performance. For future studies, a larger mixed dataset will give a better perspective on the power of machine learning for population health data.

Chapter 6

Conclusion and Future Research

Conclusion

Our study showed that 9.3% of all ED visits were attributed to frequent ED users, lower than other OECD countries. Frequent ED visits were associated with factors such as being a male and age 65+, living in a province, having no private health insurance and having endocrine and metabolic conditions. Based on the findings of this study, future interventional studies may be required to design policies regarding proper and effective care management. For example, primary care settings such as community or public health centres in South Korea could prevent non-urgent frequent ED visits of 65+ males without private health insurance who live in areas other than the Capital and are diagnosed with endocrine and metabolic. Several potential changes could take place to achieve this goal: changing the perception of patients toward primary care settings, improving the quality of care by physicians in primary settings, improving the health literacy in the general public to distinguish between non-urgent and urgent needs, which can increase the use of primary care instead of a hospital, and even additional reimbursement for primary care physicians, so they extend their office hours according to their patients' needs (Ock et al., 2014).

Furthermore, three meaningful clusters of frequent ED users were found. These results highlight the heterogeneity of frequent ED visits. Further research is needed to improve the generalizability of our results. The current research on the characteristics and potential subgroups of frequent emergency department patients can be used to implement multidimensional strategies to minimize ED overcrowding and optimize emergency care.

Our study also found that Random Forest with 98% precision, 95% sensitivity, and 3.8% classification error had the best predictive power. Logistic regression underperformed other algorithms with the lowest precision (65%) and sensitivity (67%) and the highest classification error (34.9%). The results

show that ML classification algorithms are robust techniques with predictive power for future ED visit identification and prediction.

Understanding the characteristics of frequent ED users and potential risk factors associated with frequent ED utilization, identifying the ED utilization pattern, and accurate prediction of frequent ED visits are all essential for designing effective interventions to reduce the number of preventable visits and hence the cost of ED visits for healthcare systems.

Future research

Limited research is performed using machine learning in population health, health services research and health policy, although it allows researchers to have a more in-depth understanding of their data, especially if it falls into the *big data* definition. Most importantly, the post-COVID-19 era is where ML algorithms will be beneficial to dig into the data of where COVID-19 has a potential impact. Since our study focuses more on the method than the data itself, we can extend our work in several directions, two of which are listed below:

1. Non-urgent or avoidable ED visits are one of the main reasons for overcrowding EDs worldwide. Low health literacy has been associated with frequent ED visits due to non-urgent reasons (Griffey et al., 2014). During the pandemic, unnecessary visits to EDs have put the users at higher risk of infection and increase the risk for urgent patients waiting to be admitted to the hospital. One of the most troubling tasks of health authorities has been educating people on when to visit ED during the pandemic. A comparison between the ED usage pattern in countries such as New Zealand and South Korea with the U.S., Canada or a European country could yield meaningful findings to be used in the future. For such a study, big data techniques can be applied, and machine learning is the most reliable and robust method that can analyse a combination of clinical and social factors.

2. Our pre-COVID-19 data showed there were not many visits to ED due to mental health in South Korea. However, we know that the ED visit due to mental health, especially among the younger population, has increased during the pandemic (Leeb et al., 2020). Investigating the change after the pandemic and comparing it to other countries can illuminate a path to better understanding the effect of a pandemic on societies. Machine learning methods have the ability

to perform on a large and wide dataset such as National health insurance and hospitals to identify the characteristics of the patients, find a pattern among them and predict future patients with high accuracy.

References

Agarwal, P., Bias, T. K., Madhavan, S., Sambamoorthi, N., Frisbee, S., & Sambamoorthi, U. (2016). Factors Associated With Emergency Department Visits: A Multistate Analysis of Adult Fee-for-Service Medicaid Beneficiaries. *Health Services Research and Managerial Epidemiology*, *3*(3). https://doi.org/10.1177/2333392816648549

Ahmad, P., Qamar, S., & Qasim Afser Rizvi, S. (2015). Techniques of Data Mining In Healthcare: A Review. *International Journal of Computer Applications*, *120*(15). https://doi.org/10.5120/21307-4126

Al-Masri, A. (2019). What Are Overfitting and Underfitting in Machine Learning? In *Towards Data Science*.

Alashwal, H., El Halaby, M., Crouse, J. J., Abdalla, A., & Moustafa, A. A. (2019). The application of unsupervised clustering methods to Alzheimer's disease. In *Frontiers in Computational Neuroscience* (Vol. 13). https://doi.org/10.3389/fncom.2019.00031

Armstrong, J. J., Zhu, M., Hirdes, J. P., & Stolee, P. (2012). K-means cluster analysis of rehabilitation service users in the home health care system of Ontario: Examining the heterogeneity of a complex geriatric population. *Archives of Physical Medicine and Rehabilitation*, *93*(12). https://doi.org/10.1016/j.apmr.2012.05.026

Awad, M., & Khanna, R. (2015). Efficient learning machines: Theories, concepts, and applications for engineers and system designers. In *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. https://doi.org/10.1007/978-1-4302-5990-9

Baker, R. S. J. d. (2010). Data mining. In *International Encyclopedia of Education* (pp. 112–118). Elsevier Ltd. https://doi.org/10.1016/B978-0-08-044894-7.01318-X

Batal, H., Tench, J., McMillan, S., Adams, J., & Mehler, P. S. (2001). Predicting patient visits to an urgent care clinic using calendar variables. *Academic Emergency Medicine*, *8*(1). https://doi.org/10.1111/j.1553-2712.2001.tb00550.x

Berchet, C. (2015). Emergency Care Services: Trends, Drivers and Interventions to Manage the Demand. *OECD Health Working Papers*, *83*. https://doi.org/10.1787/5jrts344crns-en

Bieler, G., Paroz, S., Faouzi, M., Trueb, L., Vaucher, P., Althaus, F., Daeppen, J. B., & Bodenmann, P. (2012). Social and medical vulnerability factors of emergency department frequent users in a universal health insurance system. *Academic Emergency Medicine*, *19*(1). https://doi.org/10.1111/j.1553-2712.2011.01246.x

Brownlee, J. (2016). Overfitting and Underfitting With Machine Learning Algorithms. *Machine Learning Mastery*.

Burns, T. R. (2017). Contributing factors of frequent use of the emergency department: A synthesis. In *International Emergency Nursing* (Vol. 35). https://doi.org/10.1016/j.ienj.2017.06.001

Canadian Institute for Health Information. (2018). *NACRS Emergency Department Visits and Length of Stay, 2017–2018* . https://www.cihi.ca/en/access-data-

reports/results?fs3%5B0%5D=primary_theme%3A676&fs3%5B1%5D=published_date%3A2018

Canadian Institute for Health Information (CIHI). (2015). *Sources of Potentially Avoidable Emergency Department Visits*. https://secure.cihi.ca/free_products/ED_Report_ForWeb_EN_Final.pdf

Cao, F., Liang, J., & Bai, L. (2009). A new initialization method for categorical data clustering. *Expert Systems with Applications*, *36*(7). https://doi.org/10.1016/j.eswa.2009.01.060

Chen, N. C., Hsieh, M. J., Tang, S. C., Chiang, W. C., Huang, K. Y., Tsai, L. K., Ko, P. C. I., Ma, M. H. M., & Jeng, J. S. (2013). Factors associated with use of emergency medical services in patients with acute stroke. *American Journal of Emergency Medicine*, *31*(5). https://doi.org/10.1016/j.ajem.2013.01.019

Chiu, Y., Racine-Hemmings, F., Dufour, I., Vanasse, A., Chouinard, M. C., Bisson, M., & Hudon, C. (2019). Statistical tools used for analyses of frequent users of emergency department: A scoping review. In *BMJ Open* (Vol. 9, Issue 5). https://doi.org/10.1136/bmjopen-2018-027750

Chun, C., Kim, S., Lee, J., & Lee, S. (2009). Republic of Korea: health system review. *World Health Organization. Regional Office for Europe, European Observatory on Health Systems and Policies*.

Classification of Diseases-6th version. (2010). Korean Standard Statistical Classification. http://kssc.kostat.go.kr/ksscNew_web/ekssc/main/main.do

Coleman, P., Irons, R., & Nicholl, J. (2001). Will alternative immediate care services reduce demands for non-urgent treatment at accident and emergency? *Emergency Medicine Journal*, *18*(6). https://doi.org/10.1136/emj.18.6.482

Das, L. T., Abramson, E. L., Stone, A. E., Kondrich, J. E., Kern, L. M., & Grinspan, Z. M. (2017). Predicting frequent emergency department visits among children with asthma using EHR data. *Pediatric Pulmonology*, *52*(7). https://doi.org/10.1002/ppul.23735

Dawson, H., & Zinck, G. (2009). CIHI Survey: ED spending in Canada: a focus on the cost of patients waiting for access to an in-patient bed in Ontario. *Healthcare Quarterly (Toronto, Ont.)*, *12*(1). https://doi.org/10.12927/hcq.2009.20411

Folckele, C., Janeway, H., & Hiseh, D. (2019). Ch. 28 - Social Determinants of Health. In *Emergency Medicine Advocacy Handbook*. Emergency Medicine Residents' Association.

Franchi, C., Cartabia, M., Santalucia, P., Baviera, M., Mannucci, P. M., Fortino, I., Bortolotti, A., Merlino, L., Monzani, V., Clavenna, A., Roncaglioni, M. C., & Nobili, A. (2017). Emergency department visits in older people: pattern of use, contributing factors, geographical differences and outcomes. *Aging Clinical and Experimental Research*, *29*(2). https://doi.org/10.1007/s40520-016-0550-5

Fuda, K. K., & Immekus, R. (2006). Frequent Users of Massachusetts Emergency Departments: A Statewide Analysis. *Annals of Emergency Medicine*, *48*(1). https://doi.org/10.1016/j.annemergmed.2006.03.001

Fuller, D., Buote, R., & Stanley, K. (2017). A glossary for big data in population and public health:

Discussion and commentary on terminology and research methods. *Journal of Epidemiology and Community Health*, *71*(11). https://doi.org/10.1136/jech-2017-209608

Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems. In *O'Reilly Media*.

Google. (2020). *Colaboratory – Google*. Colaboratory Frequently Asked Questions.

Griffey, R. T., Kennedy, S. K., McGownan, L., Goodman, M., & Kaphingst, K. A. (2014). Is low health literacy associated with increased emergency department utilization and recidivism? *Academic Emergency Medicine*, *21*(10). https://doi.org/10.1111/acem.12476

Grinspan, Z. M., Patel, A. D., Hafeez, B., Abramson, E. L., & Kern, L. M. (2018). Predicting frequent emergency department use among children with epilepsy: A retrospective cohort study using electronic health data from 2 centers. *Epilepsia*, *59*(1). https://doi.org/10.1111/epi.13948

Grinspan, Z. M., Shapiro, J. S., Abramson, E. L., Hooker, G., Kaushal, R., & Kern, L. M. (2015). Predicting frequent ED use by people with epilepsy with health information exchange data. *Neurology*, *85*(12). https://doi.org/10.1212/WNL.0000000000001944

Griswold, S. K., Nordstrom, C. R., Clark, S., Gaeta, T. J., Price, M. L., & Camargo, C. A. (2005). Asthma exacerbations in North American adults: Who are the "frequent fliers" in the emergency department? *Chest*, *127*(5). https://doi.org/10.1378/chest.127.5.1579

Han, A., Ospina, M. B., Blitz, S., Strome, T., & Rowe, B. H. (2007). Patients presenting to the emergency department: The use of other health care services and reasons for presentation. *Canadian Journal of Emergency Medicine*, *9*(6). https://doi.org/10.1017/S1481803500015451

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, *2*(3). https://doi.org/10.1023/A:1009769707641

Hunt, K. A., Weber, E. J., Showstack, J. A., Colby, D. C., & Callaham, M. L. (2006). Characteristics of Frequent Users of Emergency Departments. *Annals of Emergency Medicine*, *48*(1), 1–8. https://doi.org/10.1016/j.annemergmed.2005.12.030

Japkowicz, N., & Shah, M. (2011). Evaluating learning algorithms: A classification perspective. In *Evaluating Learning Algorithms: A Classification Perspective* (Vol. 9780521196000). https://doi.org/10.1017/CBO9780511921803

Jaroudi, S., Yang, S., & Berdine, G. (2019). Trends in emergency department visits in Lubbock from 2011-2017. *The Southwest Respiratory and Critical Care Chronicles*, *7*(27). https://doi.org/10.12746/swrccc.v7i27.513

Jones, S. S., Thomas, A., Evans, R. S., Welch, S. J., Haug, P. J., & Snow, G. L. (2008). Forecasting daily patient volumes in the emergency department. *Academic Emergency Medicine*, *15*(2). https://doi.org/10.1111/j.1553-2712.2007.00032.x

Jung, H. M., Kim, M. J., Kim, J. H., Park, Y. S., Chung, H. S., Chung, S. P., & Lee, J. H. (2021). The effect of overcrowding in emergency departments on the admission rate according to the emergency triage level. *PLoS ONE*, *16*(2 February).

https://doi.org/10.1371/journal.pone.0247042

Khan, S. S., & Ahmad, A. (2012). Cluster Center Initialization for Categorical Data Using Multiple Attribute Clustering. *MultiClust@ SDM*.

Kim, B. S., Kim, J. Y., Choi, S. H., & Yoon, Y. H. (2018). Understanding the characteristics of recurrent visits to the emergency department by paediatric patients: A retrospective observational study conducted at three tertiary hospitals in Korea. *BMJ Open*, *8*(2). https://doi.org/10.1136/bmjopen-2017-018208

Korean Health Panel Study (KHPS). (2016). Korea Health Panel Study. https://www.khp.re.kr:444/eng/main.do

Krieg, C., Hudon, C., Chouinard, M. C., & Dufour, I. (2016). Individual predictors of frequent emergency department use: A scoping review. In *BMC Health Services Research* (Vol. 16, Issue 1). https://doi.org/10.1186/s12913-016-1852-1

LaCalle, E. J., Rabin, E. J., & Genes, N. G. (2013). High-frequency users of emergency department care. *Journal of Emergency Medicine*, *44*(6). https://doi.org/10.1016/j.jemermed.2012.11.042

LaCalle, E., & Rabin, E. (2010). Frequent Users of Emergency Departments: The Myths, the Data, and the Policy Implications. In *Annals of Emergency Medicine* (Vol. 56, Issue 1). https://doi.org/10.1016/j.annemergmed.2010.01.032

Lavergne, M. R. (2016). Identifying distinct geographic health service environments in British Columbia, Canada: Cluster analysis of population-based administrative data. *Healthcare Policy*, *12*(1). https://doi.org/10.12927/hcpol.2016.24717

Lee, Y. J., Shin, S. Do, Lee, E. J., Cho, J. S., & Cha, W. C. (2015). Emergency department overcrowding and ambulance turnaround time. *PLoS ONE*, *10*(6). https://doi.org/10.1371/journal.pone.0130758

Leeb, R. T., Bitsko, R. H., Radhakrishnan, L., Martinez, P., Njai, R., & Holland, K. M. (2020). Mental Health–Related Emergency Department Visits Among Children Aged <18 Years During the COVID-19 Pandemic — United States, January 1–October 17, 2020. *MMWR. Morbidity and Mortality Weekly Report*, *69*(45). https://doi.org/10.15585/mmwr.mm6945a3

Liao, M., Li, Y., Kianifard, F., Obi, E., & Arcona, S. (2016). Cluster analysis and its application to healthcare claims data: A study of end-stage renal disease patients who initiated hemodialysis Epidemiology and Health Outcomes. *BMC Nephrology*, *17*(1). https://doi.org/10.1186/s12882-016-0238-2

Lucas, R. H., & Sanford, S. M. (1998). An analysis of frequent users of emergency care at an urban university hospital. *Annals of Emergency Medicine*, *32*(5). https://doi.org/10.1016/S0196-0644(98)70033-2

Mandelberg, J. H., Kuhn, R. E., & Kohn, M. A. (2000). Epidemiologic analysis of an urban, public emergency department's frequent users. *Academic Emergency Medicine*, *7*(6). https://doi.org/10.1111/j.1553-2712.2000.tb02037.x

Marcilio, I., Hajat, S., & Gouveia, N. (2013). Forecasting daily emergency department visits using calendar variables and ambient temperature readings. *Academic Emergency Medicine*, *20*(8). https://doi.org/10.1111/acem.12182

Newcombe, P. J., Connolly, S., Seaman, S., Richardson, S., & Sharp, S. J. (2018). A two-step method for variable selection in the analysis of a case-cohort study. *International Journal of Epidemiology*, *47*(2). https://doi.org/10.1093/ije/dyx224

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. Generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems*.

Nikki Castle. (2017, July 13). *Supervised vs. Unsupervised Machine Learning*. https://blogs.oracle.com/ai-and-datascience/post/supervised-vs-unsupervised-machine-learning

Nisbet, R., Miner, G., & Yale, K. (2017). Handbook of statistical analysis and data mining applications. In *Handbook of Statistical Analysis and Data Mining Applications*. https://doi.org/10.1016/c2012-0-06451-4

Ock, M., Kim, J. E., Jo, M. W., Lee, H. J., Kim, H. J., & Lee, J. Y. (2014). Perceptions of primary care in Korea: A comparison of patient and physician focus group discussions. In *BMC Family Practice* (Vol. 15, Issue 1). https://doi.org/10.1186/s12875-014-0178-5

OECD Reviews of Public Health: Korea. (2020). OECD. https://doi.org/10.1787/be2b7063-en

Ohno-Machado, L. (2011). Realizing the full potential of electronic health records: the role of natural language processing. In *Journal of the American Medical Informatics Association : JAMIA* (Vol. 18, Issue 5). https://doi.org/10.1136/amiajnl-2011-000501

Pereira, M., Singh, V., Hon, C. P., Greg McKelvey, T., Sushmita, S., & De Cock, M. (2016). Predicting future frequent users of emergency departments in California state. *ACM-BCB 2016 - 7th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. https://doi.org/10.1145/2975167.2985845

Pham, J. C. (2017). *Characteristics of Frequent Users of Three Hospital Emergency Departments | Agency for Healthcare Research & Quality*. Agency for Healthcare Research and Quality.

Pines, J. M., Asplin, B. R., Kaji, A. H., Lowe, R. A., Magid, D. J., Raven, M., Weber, E. J., & Yealy, D. M. (2011). Frequent users of emergency department services: Gaps in knowledge and a proposed research agenda. *Academic Emergency Medicine*, *18*(6). https://doi.org/10.1111/j.1553-2712.2011.01086.x

Poole, S., Grannis, S., & Shah, N. H. (2016). Predicting Emergency Department Visits. *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science*, *2016*.

Puka, K., Smith, M. Lou, Moineddin, R., Snead, O. C., & Widjaja, E. (2016). The influence of socioeconomic status on health resource utilization in pediatric epilepsy in a universal health insurance system. *Epilepsia*, *57*(3). https://doi.org/10.1111/epi.13290

Raven, M. C. (2011). What we don't know may hurt us: Interventions for frequent emergency department users. In *Annals of Emergency Medicine* (Vol. 58, Issue 1).

https://doi.org/10.1016/j.annemergmed.2011.04.009

Seo, D. H., Kim, M. J., Kim, K. H., Park, J., Shin, D. W., Kim, H., Jeon, W., Kim, H., & Park, J. M. (2018). The characteristics of pediatric emergency department visits in Korea: An observational study analyzing Korea Health Panel data. *PLoS ONE*, *13*(5). https://doi.org/10.1371/journal.pone.0197929

Šimundić, A.-M. (2009). Measures of Diagnostic Accuracy: Basic Definitions. *EJIFCC*, *19*(4).

Smolyakov, V. (2017). Ensemble Learning to Improve Machine Learning Results. *Stats & Bots*.

Song, Y. J. (2009). The South Korean health care system. In *Japan Medical Association Journal* (Vol. 52, Issue 3).

Soril, L. J. J., Leggett, L. E., Lorenzetti, D. L., Noseworthy, T. W., & Clement, F. M. (2016). Characteristics of frequent users of the emergency department in the general adult population: A systematic review of international healthcare systems. In *Health Policy* (Vol. 120, Issue 5). https://doi.org/10.1016/j.healthpol.2016.02.006

Tekieh, M. H., & Raahemi, B. (2015). Importance of data mining in healthcare: A survey. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015*. https://doi.org/10.1145/2808797.2809367

Thampi, B. V, Lukashin, C., & Wong, T. (2013). Novel application of Random Forest method in CERES scene type classification. In *National Aeronautics and Space Administration*. https://ceres.larc.nasa.gov/documents/STM/2013-10/27_Bijoy_Random_Forest.pdf

Ustulin, M., Woo, J., Woo, J. T., & Rhee, S. Y. (2018). Characteristics of frequent emergency department users with type 2 diabetes mellitus in Korea. *Journal of Diabetes Investigation*, *9*(2). https://doi.org/10.1111/jdi.12712

Vranas, K. C., Jopling, J. K. S., Sweeney, T. E., Ramsey, M. C., Milstein, A. S., Slatore, C. G., Escobar, G. J., & Liu, V. X. (2017). Identifying Distinct Subgroups of Intensive Care Unit Patients: a Machine Learning Approach HHS Public Access. *Crit Care Med*, *45*(10).

Vuik, S. I., Mayer, E., & Darzi, A. (2016). A quantitative evidence base for population health: Applying utilization-based cluster analysis to segment a patient population. *Population Health Metrics*, *14*(1). https://doi.org/10.1186/s12963-016-0115-z

Whittaker, W., Anselmi, L., Kristensen, S. R., Lau, Y. S., Bailey, S., Bower, P., Checkland, K., Elvey, R., Rothwell, K., Stokes, J., & Hodgson, D. (2016). Associations between Extending Access to Primary Care and Emergency Department Visits: A Difference-In-Differences Analysis. *PLoS Medicine*, *13*(9). https://doi.org/10.1371/journal.pmed.1002113

Wiemken, T. L., & Kelley, R. R. (2019). Machine learning in epidemiology and health outcomes research. In *Annual Review of Public Health* (Vol. 41). https://doi.org/10.1146/annurev-publhealth-040119-094437

Woo, J. H., Grinspan, Z., Shapiro, J., & Rhee, S. Y. (2016). Frequent users of hospital emergency departments in Korea characterized by claims data from the National Health insurance: A cross

sectional study. *PLoS ONE*, *11*(1). https://doi.org/10.1371/journal.pone.0147450

Appendix A

Python Coding for Machine Learning Methods

Random Forest

```
from sklearn.ensemble import RandomForestClassifier
import pandas as pd
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
%matplotlib inline
from sklearn.model_selection import train_test_split

from sklearn.model_selection import cross_val_score
from sklearn.metrics import classification_report, confusion_matrix

X = JoinAll.loc[:, JoinAll.columns != 'ERCOUNT']
y = JoinAll.loc[:, JoinAll.columns == 'ERCOUNT']

from imblearn.over_sampling import SMOTE
os = SMOTE(random_state=0)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
columns = X_train.columns
os_data_X,os_data_y=os.fit_sample(X_train, y_train.values.ravel())
os_data_X = pd.DataFrame(data=os_data_X,columns=columns )
os_data_y= pd.DataFrame(data=os_data_y,columns=['ERCOUNT'])

#to fit Votting
X=os_data_X[columns]
y=os_data_y['ERCOUNT']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

classifier=RandomForestClassifier()

rfc_cv_score = cross_val_score(classifier, X, y, cv=10, scoring='roc_auc')
from sklearn.model_selection import RandomizedSearchCV
# number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]
# number of features at every split
```

```
max_features = ['auto', 'sqrt']

# max depth
max_depth = [int(x) for x in np.linspace(100, 500, num = 11)]
max_depth.append(None)
# create random grid
random_grid = {
 'n_estimators': n_estimators,
 'max_features': max_features,
 'max_depth': max_depth
 }

# Random search of parameters
rfc_random = RandomizedSearchCV(estimator = classifier, param_distributions = random_grid, n_iter = 100, cv
= 3, verbose=2, random_state=42, n_jobs = -1)
# Fit the model
rfc_random.fit(X_train, y_train)
# print results
print(rfc_random.best_params_)

#Create a Classifier
classifier=RandomForestClassifier(n_estimators=1400, max_depth=26, max_features='auto')
#Train the model
classifier.fit(X_train,y_train)

from sklearn.metrics import classification_report, confusion_matrix
y_pred= classifier.predict(X_test)

feature_imp = pd.Series(classifier.feature_importances_, index=X.columns).sort_values(ascending=False)
feature_imp

import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
# Creating a bar plot
plt.figure(figsize=(10, 10))
sns.barplot(x=feature_imp, y=feature_imp.index)
# Add labels to your graph
plt.xlabel('Feature Importance Score')
plt.ylabel('Features')
plt.title("Visualizing Important Features")
plt.legend()
plt.show()
```

```
#The receiver operating characteristic (ROC)
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
mix_roc_auc = roc_auc_score(y_test, classifier.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, classifier.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Random Forest (area = %0.2f)' % mix_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('RF_ROC')
plt.show()
#from sklearn.metrics import accuracy_score
print(classification_report(y_test,y_pred))
#print(confusion_matrix(y_test,y_pred))
#print("Accuracy:",metrics.accuracy_score(y_test, y_pred))

from sklearn.metrics import accuracy_score

accuracy= accuracy_score(y_test, y_pred, normalize=True, sample_weight=None)

error_rate = 1 - accuracy
print (error_rate)
```

SVM

```
#importing libraries
import pandas as pd
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
%matplotlib inline
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.feature_selection import RFE
from sklearn.preprocessing import StandardScaler
from sklearn import svm
#from sklearn.svm import SVC
#Loading the dataset

X = data_final.loc[:, data_final.columns != 'ERCOUNT']
y = data_final.loc[:, data_final.columns == 'ERCOUNT']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

print("Number transactions X_train dataset: ", X_train.shape)
print("Number transactions y_train dataset: ", y_train.shape)
print("Number transactions X_test dataset: ", X_test.shape)
print("Number transactions y_test dataset: ", y_test.shape)

from imblearn.over_sampling import SMOTE
os = SMOTE(random_state=0)
columns = X_train.columns
os_data_X,os_data_y=os.fit_sample(X_train, y_train.values.ravel())
os_data_X = pd.DataFrame(data=os_data_X,columns=columns )
os_data_y= pd.DataFrame(data=os_data_y,columns=['ERCOUNT'])

# we can Check the numbers of our data
print("length of oversampled data is ",len(os_data_X))
print("Number of less than 4 in oversampled data",len(os_data_y[os_data_y['ERCOUNT']==0]))
print("Number of 4 or more",len(os_data_y[os_data_y['ERCOUNT']==1]))
print("Proportion of less than 4 data in oversampled data is ",len(os_data_y[os_data_y['ERCOUNT']==0])/len(os
_data_X))
print("Proportion of 4 or more data in oversampled data is ",len(os_data_y[os_data_y['ERCOUNT']==1])/len(os_
data_X))

#to fit SVM
```

```
X=os_data_X[columns]
y=os_data_y['ERCOUNT']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
#Create a svm Classifier
clf = svm.SVC(kernel='rbf',probability=True,  class_weight='balanced',  C=1.0, gamma= 'scale')

#Train the model using the training sets
clf.fit(X_train,  y_train)

#Predict the response for test dataset
y_pred = clf.predict(X_test)

from sklearn import metrics

from sklearn.metrics import classification_report, confusion_matrix

print(confusion_matrix(y_test,y_pred))
print(classification_report(y_test,y_pred))
#The receiver operating characteristic (ROC)
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
svm_roc_auc = roc_auc_score(y_test, clf.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, clf.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='SVM (area = %0.2f)' % svm_roc_auc)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic example')
plt.legend(loc="lower right")
plt.show()

from sklearn.metrics import accuracy_score

accuracy= accuracy_score(y_test, y_pred, normalize=True, sample_weight=None)
error_rate = 1 - accuracy
print (error_rate)
```

Voting

```
# Voting Ensemble for Classification

from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.ensemble import VotingClassifier

X = data_final_vot.loc[:, data_final_vot.columns != 'ERCOUNT']
y = data_final_vot.loc[:, data_final_vot.columns == 'ERCOUNT']

from imblearn.over_sampling import SMOTE
os = SMOTE(random_state=0)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
columns = X_train.columns
os_data_X,os_data_y=os.fit_sample(X_train, y_train.values.ravel())
os_data_X = pd.DataFrame(data=os_data_X,columns=columns )
os_data_y= pd.DataFrame(data=os_data_y,columns=['ERCOUNT'])

#to fit Votting
X=os_data_X[columns]
y=os_data_y['ERCOUNT']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

kfold = model_selection.KFold(n_splits=10, random_state=7)
# create the sub models
estimators = []
model1 = LogisticRegression()
estimators.append(('logistic', model1))
model2 = DecisionTreeClassifier()
estimators.append(('cart', model2))
model3 = SVC(probability = True)
estimators.append(('svm', model3))
# create the ensemble model
ensemble = VotingClassifier(estimators , voting='soft')
results = model_selection.cross_val_score(ensemble, X_train ,y_train.values.ravel(), cv=kfold)
print(results.mean())
ensemble.fit(X_train , y_train)
y_pred= ensemble.predict(X_test)

#The receiver operating characteristic (ROC)
from sklearn.metrics import roc_auc_score
```

```
from sklearn.metrics import roc_curve
mix_roc_auc = roc_auc_score(y_test, ensemble.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, ensemble.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='LR+DT+SVM (area = %0.2f)' % mix_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('LR+DT+SVM_ROC')
plt.show()

print(classification_report(y_test,y_pred))
print(confusion_matrix(y_test,y_pred))
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

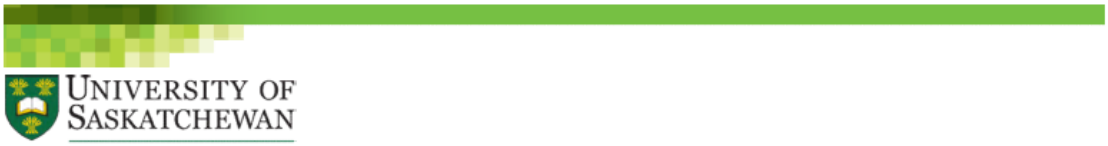
Bagging

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import Imputer
from sklearn.preprocessing import MinMaxScaler
from sklearn import model_selection
from sklearn.ensemble import BaggingClassifier
from sklearn.tree import DecisionTreeClassifier

X = data_final_bag.loc[:, data_final_bag.columns != 'ERCOUNT']
y = data_final_bag.loc[:, data_final_bag.columns == 'ERCOUNT']

from imblearn.over_sampling import SMOTE
os = SMOTE(random_state=0)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
columns = X_train.columns
os_data_X,os_data_y=os.fit_sample(X_train, y_train.values.ravel())
os_data_X = pd.DataFrame(data=os_data_X,columns=columns )
os_data_y= pd.DataFrame(data=os_data_y,columns=['ERCOUNT'])

#to fit Bagging
X=os_data_X[columns]
y=os_data_y['ERCOUNT']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

kfold = model_selection.KFold(n_splits=10, random_state=7)
cart = DecisionTreeClassifier()
num_trees = 100
model = BaggingClassifier(base_estimator=cart, n_estimators=num_trees, random_state=7)
results = model_selection.cross_val_score(model, X, y.values.ravel(), cv=kfold)
print(results.mean())

model.fit(X_train, y_train)
y_pred = model.predict(X_test)

#The receiver operating characteristic (ROC)
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
bag_roc_auc = roc_auc_score(y_test, model.predict(X_test))
fpr, tpr, thresholds = roc_curve(y_test, model.predict_proba(X_test)[:,1])
plt.figure()
plt.plot(fpr, tpr, label='Bagging (Kfold)(area = %0.2f)' % bag_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
```

```
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('Log_ROC')
plt.show()


print(confusion_matrix(y_test,y_pred))
print(classification_report(y_test,y_pred))
```

Appendix B

Ethics Approval Letter



To:	Hyun Lim, Department of Community Health and Epidemiology
Sub-Investigators:	Razieh Safaripour, College of Medicine Cheng Yanzhao Cheng, School of Public Health Kabir Md Rasel Kabir, School of Public Health Kim Min Young Kim, School of Public Health
Date:	February 13, 2020
RE:	Behavioural Ethics Application ID 1759

Thank you for submitting your project entitled: "Statistical methods in epidemiology using South Korean Health Panel (KHP) Data". This project meets the requirements for exemption status as per **Article 2.2 of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 (2018)**, which states "Research does not require REB review when it relies exclusively on information that is:

a. publicly available through a mechanism set out by legislation or regulation and that is protected by law; or
b. in the public domain and the individuals to whom the information refers have no reasonable expectation of privacy."

It should be noted that though your project is exempt of ethics review, your project should be conducted in an ethical manner (i.e. in accordance with the information that you submitted). It should also be noted that any deviation from the original methodology and/or research question should be brought to the attention of the Behavioural Research Ethics Board for further review.

Digitally Approved by Vivian Ramsden, Vice-Chair
Behavioural Research Ethics Board
University of Saskatchewan