Predictive Modelling for
30-Days Hospital Risk of Readmission

Student Name: Farhein Akmal

Degree: Master of Health Informatics (MHI)

Year: 2018

Table Of Contents

List of Tables

List Of Figures

List of Abbreviations

ED: Emergency Department

LACE: Length of the stay, Acuity of the admission, Comorbidity index score, Emergency department

OHIP: Ontario Health Insurance Plan

RHR-30: Risk of 30-days Hospital Readmissions

U30R: Unplanned 30-days Hospital Readmission

WDHB: Waitemata District Health Board

Chapter 1 - Introduction

This research report involves developing a predictive model to identify patients which are at high risk of unplanned 30-days hospital readmission (U30R). A variety of diverse criteria is studied to seek out the best model for predicting patients at risk of U30R.

When a patient is admitted as an acute admission into a hospital within a short period of time post the discharge from the hospital is considered as an unplanned hospital readmission. These unplanned hospital readmissions are burden to patients as well as the healthcare system. An increasing unplanned readmissions are not only costly for healthcare providers however are also perceived as an indicator for the quality of care that is being provided by the hospital and are also costly for the healthcare (Chassin, Loeb, Schmaltz, & Wachter, 2010; Zhou, Della, Roberts, Goh, & Dhaliwal, 2016).

In 2016, acute readmission rate within 28-days of discharge for all the District Health Boards (DHBs) nationwide was 7.8%, Canterbury having the highest readmission rate of 8.9% (Health, 2016). In the United Kingdom, the emergency 30-days readmission rate between 2004 and 2010 was 7% and it was estimated that 2% were potentially preventable. (Blunt, Bardsley, Grove, & Clarke, 2014) .

There are many factors other than medical information which adds to unplanned readmissions for example social and culture factors. Hence, not all readmission can be prevented and estimating of how many can be avoided remain controversial (Glass, Lisk, & Stensland, 2012; Joynt & Jha, 2013).

Still there are interventions which could decrease the readmission rates post discharge, guaranteeing cost effective use of those interventions needs the ability to accurately predict those patients which are most at-risk for readmission. Research has been done to predict the readmission rate with the help of machine learning algorithms. Since the time of these research the advances in processing language and in machine learning algorithms have been made, hence this could head to significant improvement in predictive capabilities.

1.1 Aim and Objectives:

The aim of this project is to conduct an in-depth preprocessing of the Waitemata District Health Board (WDHB) data and build a machine learning model to predict the risk of hospital readmissions within the 30 days of discharge of the patient. The achieve this aim there were three phases:

- Analyzing key factors through literature review
- Investigating various machine learning models
- Evaluating LACE model on WDHB data. The LACE model uses four attributes: length of the stay (L), acuity of the admission (A), comorbidity index score (C) and emergency department use which measures as the number of visits in the six months before admissions (E).

  (The LACE model is explained in detail in Chapter 4 section 4.1)

1.2 Outcome Achieved:

Initially the project focus was to validate the current clinical risk of readmission assessment – LACE score used in the acute care settings in New Zealand. This study designed and developed a better predictive model, RHR-30, for the clinical readmission risk assessment in order to test the prediction accuracy within a short period of time i.e. 30 days. The RHR-30 model is a result of REP Tree algorithm with bagging technique. A few of the significant features which contributes the most in building the RHR-30 are count of emergency department visits in the six months before admission, length of the stay, age and acuity of the admission. The discrimination score (also known as ROC area and c-statistic) of RHR-30 is 0.72 which is better than the LACE model.

Chapter 2 - Literature Review

In July and August 2018, Google Scholar and AUT Library were used to search for scholarly articles relating to unplanned hospital readmission. The selection of the research for this project, were focus more to 28-days and 30-days all-cause hospital readmission. The researchers where the proposed model was compared with LACE index for validation, were preferred as this research projects intends to perform a similar procedure. In total 13 studies were assessed for this literature review. Many researchers build a logistic regression model for predicting hospital readmission risk and a few of the research experimented with different machine learning algorithms like decision tree, XGBoost, support vector machine and neural network.

Machine learning algorithms are popular tools for predicting hospital readmissions (Krumholz et al., 2011; Stiglic, Wang, Davey, & Obradovic, 2014). Some researchers have created models which aims to predict hospital readmissions in default population settings. Research suggests these models have been used and have significantly established over time. LACE is one of the popular model which is further discussed in chapter 4 section 4.1 (van Walraven et al., 2010).

2.1 Key Features for “Readmission”:

Different research has different definition for unplanned hospital readmission i.e. different inclusion and exclusion criteria were used in each study. In the study (van Walraven et al., 2010) and (van Walraven, Wong, & Forster, 2012b), samples were selected randomly for one admission per patient after excluding patients who were less than 18 years old, died in the hospital, were psychiatric or obstetric admissions and discharged to long-term care, rehabilitation or other hospitals. (van Walraven et al., 2010) and (van Walraven et al., 2012b) also excluded patients who were ineligible for Ontario Health Insurance Plan (OHIP) coverage at discharge or during 30-day post discharge period. Similarly (van Walraven, Wong, Forster, & Hawken, 2013) excluded psychiatric and obstetrical admissions, patients who died in the hospital, patients discharged to long term care, rehabilitation or other hospital. (van Walraven et al., 2013) also excluded patients who had urgent hospitalization occurring within 30 days of previous (i.e.

admissions that were counted as a result for that patient's previous admission) and the study only considered one observation per patient.

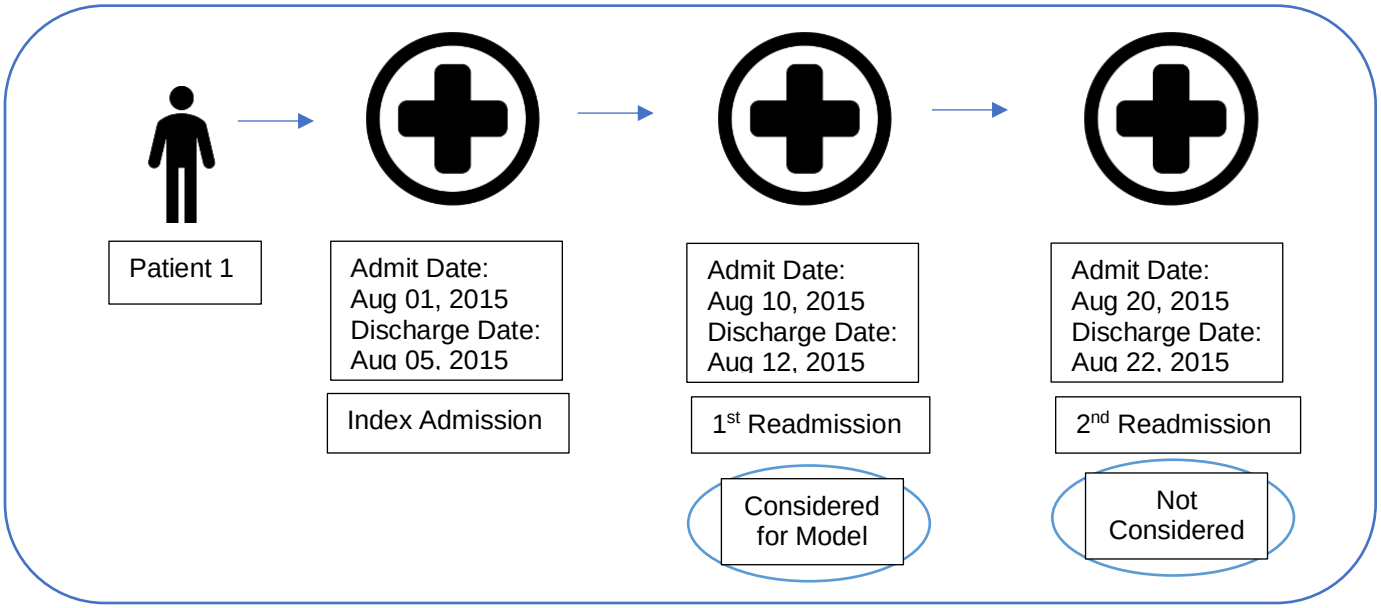

Figure 1. A scenario of a patient having multiple admissions in 30 days

One of the common inclusion criteria seen in the studies as a definition for readmission was considering patient's first admission as an index admission and any admission within 30 days following the initial hospitalization was considered as readmission. If the patients had more than two visits within 30 days, only the first visit after the discharge was considered as readmission, see Figure 1. Although further different inclusion criteria were applied in different studies. (Khan, Malone, Pagel, Vollbrecht, & Baumgardner, 2012) and (Yu et al., 2015) selected samples where patient's age was 65 or above. (Yu et al., 2015) excluded in-hospital mortality visits and patients who were not insured by Medicare or Medicaid.

(Escobar et al., 2015) and (Wakefield & Mehr, 2013) considered patients who were 18 years or above and who had overnight stay hospitalization, the study excluded 1-day surgeries and hospitalization for childbirth, whereas in (Escobar et al., 2015), post-delivery complications were included. The research (Escobar et al., 2015) included data from 21 Hospitals and one of the important selecting condition for inclusion was Epic EMR (internally known as KPHC) was functioning at the hospital for at least 3 months.

age range for the study was 18 or above at the time of admission. (Wakefield & Mehr, 2013) included patients with psychiatric conditions requiring temporary acute care such as alcohol detoxification, drug detoxification or stabilization following a suicide attempt; the study included inpatients admission with at least one of each of the following: diagnosis or procedure, medication order and laboratory order; and excluded elective readmissions unless it was from the emergency department.

Similarly, (Choudhry et al., 2013) excluded patients for psychiatry, skilled nursing, hospice, rehabilitation, patients died during the index admission and maternal and newborn visits. On contrary, (Berry et al., 2018) included newborns, children, cancer patients and mental health condition patients.

(Maali et al., 2018) included first admission to any hospital in NSW which were initiated via emergency department within 60 days of being discharge alive from the index admission, any subsequent readmission by the same patient or readmissions beyond 60 days were not included.

The study done in the (Lee, 2012) holds a different kind of inclusion criteria where patients admitted under 8 different categories of diseases (“neoplasms”, “endocrine, nutritional and metabolic”, “circulatory system”, “respiratory system”, “digestive system”, “musculoskeletal system and connective tissues”, “genitourinary system” and “symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified.”) were selected for the sample considering the exclusion of the diseases that occupied less than 5% of the sample size among the 22 disease categories of the 10th revision of the International Classification of Diseases (ICD-10) code.

The significant inclusion and exclusion criteria were extracted from the 13 studies included in this literature review for 30 days unplanned hospital readmissions are shown in Table 1.

Table 1. A summary of significant inclusion and exclusion criteria on 28-day or 30-day unplanned hospital readmission

References	Inclusion Criteria	Exclusion Criteria
(Choudhry et al., 2013)		• Psychiatry admissions • Skilled nursing admissions • Hospice admissions • Rehabilitation admissions • Maternal and newborn visits • Patients died during the index hospitalization
(Berry et al., 2018)	• Newborns • Children • Cancer patients • Mental health patients	
(Khan et al., 2012)	• Age: 65 and older	• Outpatient stays • Patient died during the hospital stay
(van Walraven, Wong, & Forster, 2012a)	• One admission per patient • Age: 18 and older	• Patients died during the hospital stay • Psychiatric or obstetric admissions • Discharge to long-term care, rehabilitations or other hospitals
(van Walraven et al., 2012b)	• One admission per patient • Age: 18 and older	• Patients died during the hospital stay • Psychiatric or obstetric admissions • Discharge to long-term care, rehabilitations or other hospitals
(Escobar et al., 2015)	• Overnight stay hospitalization • Age: 18 and older • Post-delivery complications	• One-day surgeries which didn't result in overnight stay • Hospitalization for childbirth
(Maali et al., 2018)	• Index hospitalization • Initiated via ED	• Patient died during the index hospitalization
(Shulan, Gao, & Moore, 2013)	• Index hospitalization	• Patients transferred from other hospitals
(van Walraven et al., 2013)	• One observation per patient	• Patients died during the hospital stay • Psychiatric or obstetric admissions • Discharge to long-term care, rehabilitations or other hospitals • Urgent hospitalization occurring within 30 days of previous
(Yu et al., 2015)	• Index hospitalization • If patient had multiple visits within 30days, only first one was considered as readmission • Age: 65 and older	• In-hospital mortality visits • Patients not insured by Medicare/Medicaid
(Wakefield & Mehr, 2013)	• Age: 18 and older • Psychiatric conditions requiring temporary acute care such as alcohol detoxification, drug detoxification or stabilization following a suicide attempt	• Outpatient stays • LOS 0 days • Maternal and newborns • Psychiatric or obstetric admissions • Planned readmissions unless it was from the ED

(Lee, 2012)	• When a patient had multiple visits, each visit evaluated separately	
(Baillie et al., 2013)	• All adult admissions	• Short procedure admissions • Hospice admissions • Rehabilitation admissions

2.2 Critical Analysis:

As (Swinscow & Campbell, 1997) says, design of the study is more important than the analysis as a poorly analyzed study can usually be reanalyzed however a badly designed study can never be retrieved. It's the design of the study which will govern how the data are to be analyzed.

Majority of the research have performed cohort retrospective study. A cohort retrospective study is comparing groups of individuals who are alike in many ways but differ by a certain characteristic in terms of the outcome (Institute, 2016). The data is collected from a health organization for each patient from the existing records and is be analyzed through machine learning models to determine the relative risk of the cohort compared to the control group.

Two of the studies which were reviewed in this literature review also included prospective study with a retrospective study. Prospective cohort study follows over time and compares the predicted outcome (Institute, 2016).

An overview of which research used retrospective design of study and prospective design of study is been shown in Table 2. The table also mentions the durations of the time period of the collected data.

Table 2. Study design of 13 included studies on 28-day or 30-days unplanned hospital readmission

References	Design of Study
(Choudhry et al., 2013)	12 months of retrospective data
(Berry et al., 2018)	Retrospective data
(Khan et al., 2012)	Retrospective data
(van Walraven et al., 2012a)	Retrospective data
(van Walraven et al., 2012b)	Retrospective data

(Escobar et al., 2015)	Retrospective data
(Maali et al., 2018)	12 months of retrospective data + 2 months of prospective data
(Shulan et al., 2013)	Retrospective data
(van Walraven et al., 2013)	Retrospective data
(Yu et al., 2015)	2 years, 6 years and 10 years retrospective data for Hospital 1, Hospital 2 and Hospital 3 respectively
(Wakefield & Mehr, 2013)	Retrospective data
(Lee, 2012)	Retrospective data
(Baillie et al., 2013)	24 months of retrospective data + 12 months of prospective data

2.3 Data and Model Settings:

Table 3 summarizes the characteristic of the final included studies in this literature review. The studies conducted is from various parts of the world like USA, Canada, UK and Australia. The duration of the retrieved data source ranged from one single day data collected across ten hospitals to data collected for 6 years from four health database in Ontario. Every study had a different sample size and accordingly a different derivation and validation split, for example (Shulan et al., 2013; van Walraven et al., 2012a, 2012b) has 50-50 split in derivation and validation split, on the other hand (Lee, 2012) has 70-30 split, (Choudhry et al., 2013) has 75-25 split and (Wakefield & Mehr, 2013) has 90-10 split keeping the higher number as a derivation sample percentage and smaller number as validation sample percentage. A total of 29 models were derived or validated using administrative and/or clinical/medical data. The sample size varied from 227 patients to 31729762 patients. The unplanned hospital readmission rate for 28-days or 30-days ranged from 6.60% (*n*= 62255) to 23% (*n*=2441), '*n'* being the sample size for the respective study. Table 3 also includes all predictive models made/used in the respective studies and compares its performances. In logistic regression, the result variable is the log of the odds of the event ie readmission probability/ 1- readmission probability. After determining the final model, the multi-variable logistic regression allows for the calculation of cohort studies' readmission probability. The majority of the studies reported c-statistic (also known as ROC area) which is a measure for model

discrimination and Hosmer-Lemeshow value as a measure for calibration (Steyerberg et al., 2010). The discriminative ability for all-cause unplanned hospital readmission ranged from 0.55 to 0.80 and calibration ranged from 15.11(p=0.0569) to 40.14(p<0.0001)

Table 3. Characteristics and performance of predictive models of 13 included studies on 28-day or 30-day unplanned hospital readmission predictive models

Ref.	• Data Source • Time Period • Sample Size	Model Name	Results
(Choudhry et al., 2013)	• 8 Hospital, Chicago • 1 March 2010 – 31 July 2012 • 126479	ACC Admission Model	• Discrimination: 0.75 • Calibration: 23.5 (p=0.003)
		ACC Discharge Model	• Discrimination: 0.77 • Calibration: 19.9
(Berry et al., 2018)	• US Agency for Healthcare Research and Quality Nationwide Readmissions Database • 2013 • Index Admissions: 31729762	Logistic Regression Model	Odds ratio for readmission: • Age 16-20: (range 0.70 (95% confidence interval 0.68 to 0.71) to 1.04 (1.02 to 1.06)) • Age 21-45: (range 1.02 (1.00 to 1.03) to 1.12 (1.10 to 1.14)) • Age 46-64: (range 1.02 (1.00 to 1.04) to 0.91 (0.90 to 0.93)) • Age 65 & above: (0.78 (0.77 to 0.79))
(Khan et al., 2012)	• 10 Hospitals/ EMRs • 26-Jan-11 • 227	Rehospitalisation Risk Score	• Cutoff: 7 • Sensitivity: 61% • Specificity: 22% • PPV: 12% • NPV: 77% • p=0.001
(van Walraven et al., 2012a)	• 4 Health Database, Ontario • 1 April 2003 - 31 March 2009 • Random Patients: 200000	CMG score (case-mix groups)	• Discrimination: 0.65 • Calibration: 15.11 (p=0.0569)
		LACE index (validation)	• Discrimination: 0.735 • Calibration: 21.19 (p=0.0067)
		Combined CMG score and LACE index	• Discrimination: 0.759 • Calibration: 40.14 (p<0.0001)
(van Walraven et al., 2012b)	• 4 Health Database, Ontario • 1 April 2003 - 31 March 2009 • Random Patients: 500000	LACE+	• Discrimination: 0.759
		LACE+ with CMG Score	• Discrimination: 0.771

(Escobar et al., 2015)	• 21 Hospitals EMRs • 1 June 2010 - 31 Dec 2013 • 360036	ED 30	• R square: 0.158 • Discrimination: 0.739
		Discharge Day 30	• R square: 0.174 • Discrimination: 0.756
		LACE	• R square: 0.145 • Discrimination: 0.729
(Maali et al., 2018)	• 1 Hospital, Australia • 1 July 2008 - 31 Dec 2012 • 62255	RETURN 30 (gradient tree boosting algo: XGBoost) (95% CI) (10-fold CV)	• Cutoff: 12 • Sensitivity: 52.9% • Specificity: 77.4% • PPV: 14.8% • Discrimination: 0.71
		Logistic Regression (selected variables) (10-fold CV)	• Sensitivity: 59.0% • Specificity: 73.3% • PPV: 16.5% • Discrimination: 0.72
(Shulan et al., 2013)	• Veterans Healthcare Network Upstate New York • 2011 • 8718	Unnamed (losgistic regression)	• Discrimination: 0.8
(van Walraven et al., 2013)	• Centralised Database, Ontario • 2004 – 2009 • Patients: 499996; Index Admissions: 858410	LACE+ (extension of a validation index) (95% CI)	• Discrimination: 0.73
(Yu et al., 2015)	• 3 Hospitals, USA • Not Reported • Hospital 1: 2441; Hospital 2: 26520; Hospital 3: 45785	Generic Model	• Discrimination: 0.85
		Admission: Linear SVM	Hospital2 • Recall: 0.36(0.02) • Precision: 0.18(0.01) • Discrimination: 0.60 Hospital3 • Recall: 0.34(0.01) • Precision: 0.19(0.01) • Discrimination: 0.64
		Discharge: Linear SVM	Hospital2 • Recall: 0.68(0.02) • Precision: 0.34(0.01) • Discrimination: 0.74 Hospital3 • Recall: 0.54(0.01) • Precision: 0.30(0.01) • Discrimination: 0.72
		Admission: Poly SVM	Hospital2 • Recall: 0.31(0.02) • Precision: 0.15(0.01) • Discrimination: 0.57 Hospital3

			• Recall: 0.31(0.02) • Precision: 0.17(0.01)
		Discharge: Poly SVM	Hospital2 • Recall: 0.67(0.01) • Precision: 0.33(0.01) • Discrimination: 0.70 Hospital3 • Recall: 0.52(0.01) • Precision: 0.29(0.01) • Discrimination: 0.69
		Admission: Cox PH	Hospital2 • Recall: 0.34(0.02) • Precision: 0.17(0.01) • Discrimination: 0.59 Hospital3 • Recall: 0.26(0.06) • Precision: 0.15(0.03) • Discrimination: 0.57
		Discharge: Cox PH	Hospital2 • Recall: 0.66(0.02) • Precision: 0.33(0.01) • Discrimination: 0.73 Hospital3 • Recall: 0.49(0.04) • Precision: 0.27(0.02) • Discrimination: 0.67
		Discharge: LACE	Hospital2 • Recall: 0.31(0.02) • Precision: 0.15(0.01) • Discrimination: 0.55 Hospital3 • Recall: 0.27(0.01) • Precision: 0.15(0.01) • Discrimination: 0.60
(Wakefield & Mehr, 2013)	• 91 Hospitals - Health Facts Database • 1 October 2008 - 31 August 2010 • Index Admissions: 463351	Unnaned (logistic regression)	• Discrimination: 0.657
(Lee, 2012)	• 1 Tertiary Hospital • Jan 2009 - Dec 2009 • Patients: 11951	Logistic Regression Model	• Root ASE (Asymptotic Standard Error): 0.385 • Misclassification Rate: 0.180
		Decision Tree	• Root ASE (Asymptotic Standard Error): 0.369 • Misclassification Rate: 0.177
		Neural Network	• Root ASE (Asymptotic Standard Error): 0.383 • Misclassification Rate: 0.211

(Baillie et al., 2013)	• 3 Hospitals • August 2009 - Sept 2012 • 120396	Prediction Model	• F-score: 0.339 • Discrimination: 0.614

Chapter 3 - Data, Data Handling and Analysis

3.1 Project's Objective:

The objectives of this research project are to:

1. Generate the LACE index for the given data
2. Develop all-cause 30-days hospital readmission risk prediction model
3. Compare the prediction model's performance with existing model i.e. LACE index

3.2 Inclusion and Exclusions Criteria for "Readmission":

The inclusions and exclusion criteria for readmission for this project is influenced by several previous research.

The inclusion criteria for readmission are:

1. Only acute (non-elective) admissions are considered i.e. only unplanned admissions are considered.
2. The discharge location of the index admission should be "Home" (van Walraven et al., 2012a, 2012b; van Walraven et al., 2013)
3. The patient who are 18 years old or above at the time of readmission. (Escobar et al., 2015)
4. If the patient has multiple visits within 30 days of index hospitalization, only first visit was considered as readmission (van Walraven et al., 2012a, 2012b; van Walraven et al., 2013; Yu et al., 2015)

The exclusion criteria for readmission are:

1. Hospital visits which did not result in overnight stay in the hospital (Escobar et al., 2015; Khan et al., 2012; Wakefield & Mehr, 2013)
2. Arranged admissions i.e. elective admissions (Khan et al., 2012)
3. Psychiatric admissions returned from leave (Choudhry et al., 2013; van Walraven et al., 2012a, 2012b; van Walraven et al., 2013)

3.3 Data Source:

The data used for this project is gathered from Waitemata District Health Board (WDHB). Waitemata District covers the city center and border retail and commercial areas, and the inner-city residential suburbs. The WDHB data includes admission instances of two years i.e. 2015 and 2016.

The dataset also includes six months of retrospective data and one month of prospective data. The dataset includes six months of retrospective data is included so the number of emergency department visits can be studied for previous six months if the admission date is in early January 2015 and it includes one month of prospective data to check for readmissions for the patients whose initial admission happened in late December 2016.

3.4 Data Description:

The data provided was in unstructured manner and includes three different datasets which are discussed below:

3.4.1 Input_Events:

Patient_Identifier	encounter	admitdate	dischargedate	lengthofstay	EthnicGroup	Age	gender	Female	Admit_Type	admittype	AdmissionSource	Discharge_Destination	Costweight	DRG	hospital
1000002006	H1401695571	8-Jan-15	8-Jan-15	0	Other	39	M	0	Elective	WN	Home	Home	0.948	L07B	3262
1000002006	H1401709258	29-Jan-15	29-Jan-15	0	Other	39	M	0	Acute	AC	Home	Home	0.2151	Z61B	3215
1000002006	H1401854224	17-Sep-15	18-Sep-15	1	Other	40	M	0	Elective	WN	Home	Home	0.7111	L67B	3262
1000002006	H1402155018	22-Sep-16	22-Sep-16	0	Other	41	M	0	Elective	WN	Home	Home	0.2189	L67B	3262

Figure 2. Sample of raw input_events

a. The *input_events* dataset (see Figure 2) consisted of
    - 16 attributes including two identifiers variable (patient and encounter) and
    - 213440 records originally.

Three out of four attributes were categorical variables and the rest of the attributes were numerical.

b. The dataset includes every admissions (encounters) in the year 2015 and 2016 with details like
    - patient ID,
    - gender,
    - age,

- ethnic group,
- admit type,
- admission source,
- discharge destination,
- length of the stay,
- diagnose-related group, etc.

It also includes six months of retrospective data and one month of prospective data. The prospective data was also provided by WDHB in order to confirm the U30R for patients admitted in the last days of December 2016

c. There are no missing values in the dataset, however one record showed length of stay as negative.

d. The detailed description of some of the important attributes in the dataset are shown in Table 4:

Table 4. Description of important attributes in the dataset

Attributes	Observations
Number of Patients	Out of 213440 records, there were 110611 unique patients in the dataset
Gender	Female = 58.75% Male = 41.24%
Age Group	39 years and below = 30.1% 40 to 49 years = 10.8% 50 to 59 years = 13.1% 60 to 64 years = 6.7% 65 to 69 years = 7.8% 70 to 74 years = 7.8% 75 to 79 years = 7.6% 80 to 84 years = 6.6% 85 to 89 years = 5.5% 90 to 94 years = 2.8% 95 to 99 years = 0.6% 100 years and above = 0.05%
Ethnic Group	Asians = 11.9% Maori = 7.6% Pacific Island = 7.4% Others = 72.9%
Admission Source	From Home = 96.8% From North Shore Hospital = 1.6% From Waitakere hospital = 0.96%

Discharge Destination	To Home = 87.6% To North Shore Hospital = 4.32% To Auckland Hospital = 1.29% To Waitakere Hospital = 1.28%
Admit Type	Acute = 75.6% (AC = 61.4% , AA = 14.1% , RL = 0.003%) Elective = 24.3%
Length of Stay	No overnight stay = 44.5% 1 night stay = 21.3% 2 nights stay = 9.1% 3 nights stay = 6.3% 4 to 6 nights stay = 9.2% 7 to 13 nights stay = 5.9% More than 13 night stay = 3.4%
Readmission Rate	Readmission rate = 9.83%

3.4.2 ED_events:

Patient_Identifier	ENCOUNTER	ARRIVED_DTTM	SEEN_DTTM
1000002362	A1401275732	03/31/15	3/31/15
1000002362	A1401275732	03/31/15	3/31/15
1000002812	A1401243386	12/29/14	12/29/14

Figure 3. Sample of raw ED_events

a. The *ED_events* dataset (see Figure 3) consisted of
    - Four attributes and
    - 294926 records originally.

b. The dataset includes patient ID, encounter ID, arrived date and seen date for all the visits in the emergency department.

c. 93% records in the *ED_events* had same arrived date and seen date whereas the rest 7% i.e. 19245 out of 294926 records from *ED_events* has different arrived date and seen date.

d. There are a few duplicate records as well.

e. There is no particular pattern seen between
    - The dataset, *ED_events*' arrive / seen date and
    - another dataset, *input_events*' admit date.

Three different scenarios were observed within the relation of the two datasets: *ED_events* and *input_events*

- *ED_events*' attribute "SEEN_DTTM" is equal to *input_events*' attribute "admitdate",
- *ED_events*' attribute "ARRIVED_DTTM" is equal to *input_events*' attribute "admitdate" and
- *ED_events*' attribute "SEEN_DTTM" is one day ahead to *input_events*' attribute "admitdate".

Therefore, it's difficult to comment on the relationship between two dataset *ED_events* and *input_events* dataset.

3.4.3 Clinical_codes:

a. The *clinical_codes* dataset (see Appendix 1) consisted of
   - 41 attributes and
   - 133616 records originally.

Out of the 41 attributes, nine were categorical variables and the rest of the attributes were numerical

b. The dataset includes international classification of diseases (ICD) codes for patients and similarly includes scores for various conditions like cerebrovascular, stroke, asthma, diabetes, arthritis, renal failure, sickle cell, alcoholism, cancer, metastatic carcinoma, HIV AIDS, etc.

c. There are multiple records for some patients as their scores were different on different dates.

d. Not all patients in the *input_events* dataset were present in *clinical_codes* dataset, for which it is concluded that those patients didn't have any of the mentioned diseases, hence the score for those records were assigned as zero.

3.5 Workflow for the Project:

In this project, three datasets has been undergone intense cleaning procedure to be ready for modelling as the raw data consist noises like outliers, missing values etc. Figure 4 briefly outlines the steps planned for this particular project. These steps are further explained in Figure 5 Figure 6 and Figure 7.

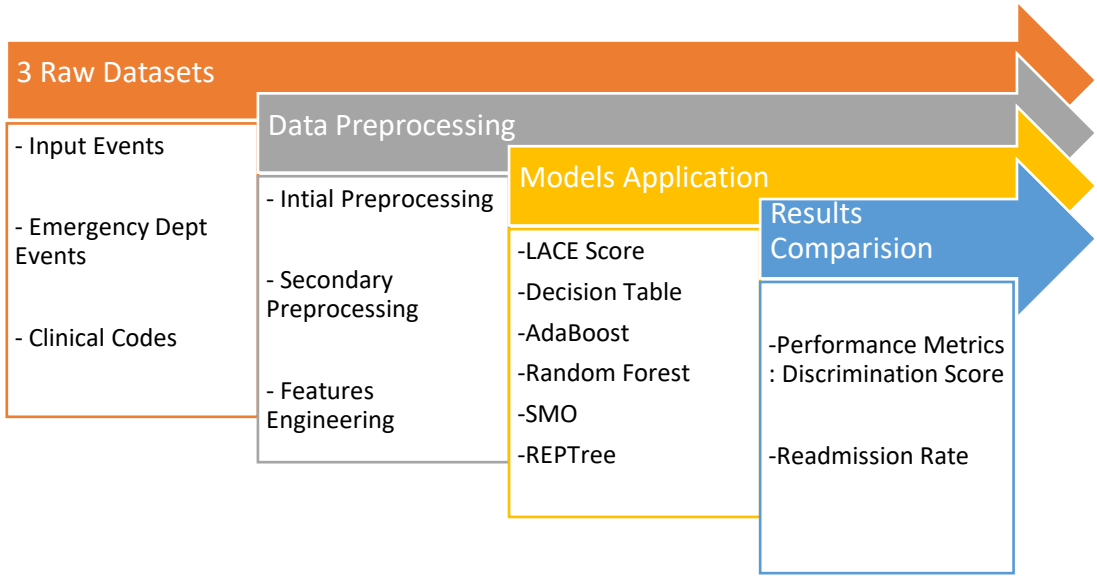

Figure 4. Flowchart of the steps involved in developing and validating the predictive model for risk of 30-days hospital readmission

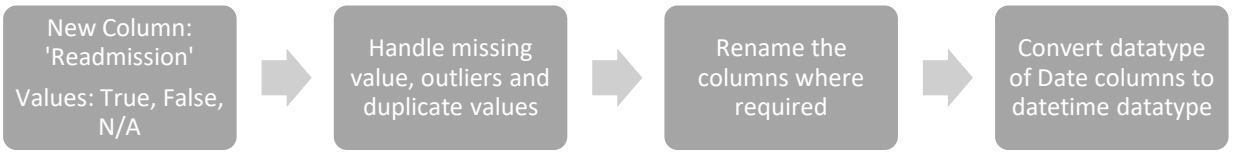

Figure 5. Flowchart of the steps involved in initial preprocessing of data in developing and validating the predictive model for risk of 30-days hospital readmission

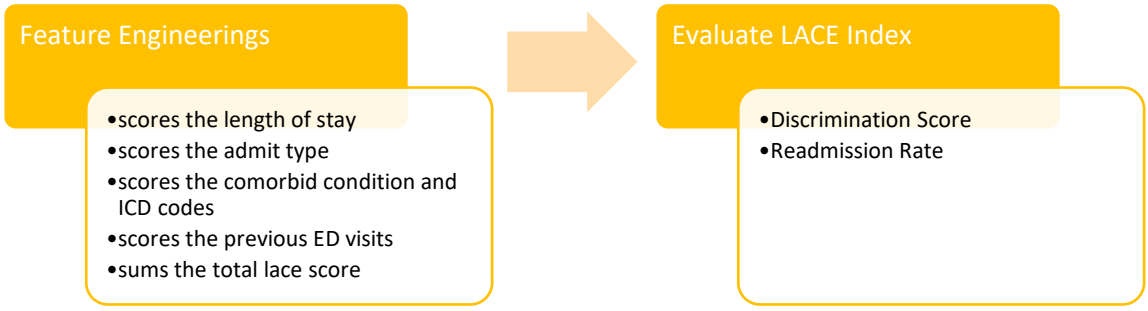

Figure 6. Flowchart of the steps involved in preparation and application of LACE model in developing and validating the predictive model for risk of 30-days hospital readmission

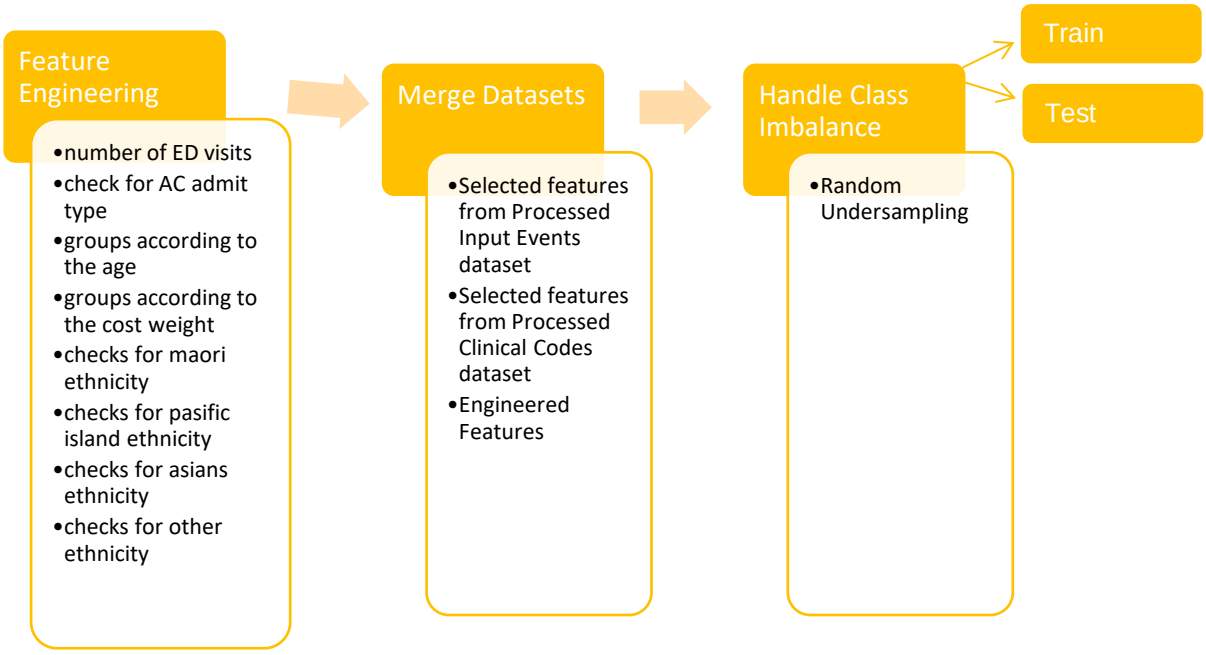

Figure 7. Flowchart of the steps involved in preparation and application of derived model in developing and validating the predictive model for risk of 30-days hospital readmission

3.6 Initial Data Pre-processing:

As mentioned above, all three datasets have to be prepared for modeling, hence has undergone intense cleaning procedure i.e. pre-processing. A few steps were taken in the data cleaning process for each of the three datasets. These steps are discussed below under their respective dataset's name title:

5.6.1 Input_events:

1. The admit date and the discharge date given in *input_events* dataset was converted into datetime datatype.
2. The dataset was sorted with patient identifier column to ease future analysis of records for the same patients.
3. There were no missing values in the dataset.
4. Removing noisy data by removing instances which doesn't satisfy the readmission criteria, which includes the following conditions:
    a. Discharge location is 'Home'
    b. Discharge date is greater than 31st December 2016
    c. Length of stay is negative
5. Creating a new column "Readmitted" with True-False value for readmissions. A true readmission will be which satisfy the following conditions:
    a. The readmission event must be within 30 days of the patient's index discharge date
    b. The admit type for admission (other than the index admission) is 'AC'
    c. The length of stay should be at least overnight.

    (Note: These conditions were based on the literature review and on the inclusion and exclusion criteria for U30R which were discussed previously in section 3.2)
6. The preprocessed *input_events* dataset is then saved as a new csv file, *processed_inputevents*.
7. Observations made based on *processed_inputevents* dataset is seen in Figure 8

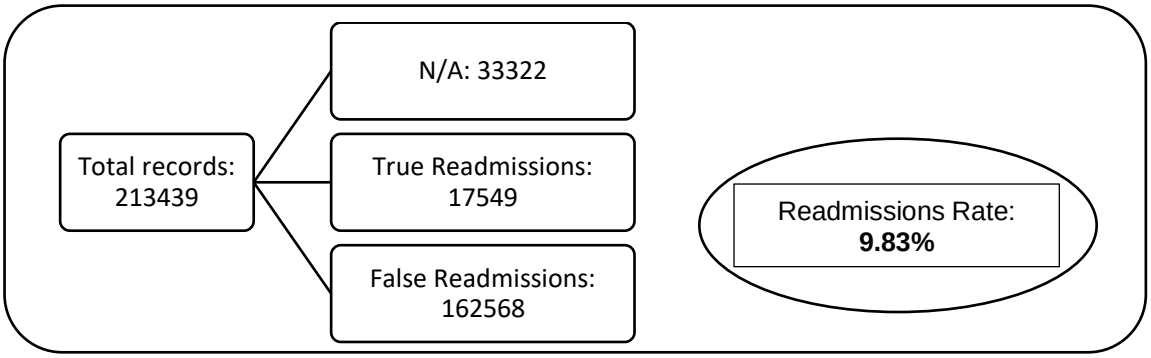

Figure 8. Outline on records of processed_inputevents dataset

3.6.2 ED_events:

1. The arrived date and seen date for the patient and encounter given in the dataset was converted into datetime datatype.
2. The encounter column in the *ED_events* dataset has a default name as “ENCOUNTER” which was renamed to lower case, “encounter”, so it matches the name format of encounter column in *processed_inputevents*.
3. The duplicate records were removed, which left us with 155590 records in the processed dataset
4. The processed *ED_events* were then saved as a new csv file, *processed_edevents*.

3.6.3 Clinical_codes:

1. The encounter column in the *clinical_codes* dataset has a default name as “PiMSEncounter” which was renamed to “encounter”, so it matched the name format of *processed_inputevents* and *processed_edevents*.
2. Columns like “ICDCode”, “ICD_Description”, “sequencenumber”, “DRG”, “DRG_FirstLetter”, “Primary_DX”, “version”, “Primary_DX_Desc” and “Condition_Group” were removed as valuable information from these columns were already present in the subsequent columns in the dataset.
3. There were multiple clinical code records for the same encounter, with different max score/CCI. The instance with the maximum score for same encounter were kept and rest duplicate instances for that encounter were removed from the dataset.
4. The preprocess *clinical_codes* dataset is then saved as a new csv file, *processed_clinicalcodes*.

The three datasets are now noise-free i.e. processed and clean. The observations and attributes of the processed datasets are mentioned in the Table 5. The sample of initial processed dataset for *input_events*, *ED_events* and *clinical_codes* can be been in Figure 9, Figure 10 and Appendix 2 respectively.

Table 5. Total number of observation and attributes in the 3 processed datasets

Dataset	Observation	Attributes
processed_inputevents	213439	17
processed_edevents	155590	4
processed_clinicalcodes	72729	30

	Patient_Identifier	encounter	admitdate	dischargedate	lengthofstay	EthnicGroup	Age	gender	Female	Admit_Type	admittype	AdmissionSource	Discharge_Destination	Costweight	DRG	hospital	Readmitted
0	1000002006	H1401695571	2015-01-08	2015-01-08	0	Other	39	M	0	Elective	WN	Home	Home	0.948	L07B	3262	False
1	1000002006	H1401709258	2015-01-29	2015-01-29	0	Other	39	M	0	Acute	AC	Home	Home	0.2151	Z61B	3215	False
2	1000002006	H1401854224	2015-09-17	2015-09-18	1	Other	40	M	0	Elective	WN	Home	Home	0.7111	L67B	3262	False
3	1000002006	H1402155018	2016-09-22	2016-09-22	0	Other	41	M	0	Elective	WN	Home	Home	0.2189	L67B	3262	False

Figure 9. Sample of processed input_events data

	Patient_Identifier	encounter	ARRIVED_DTTM	SEEN_DTTM
0	1000002362	A1401275732	03/31/15	3/31/15
2	1000002812	A1401243386	12/29/14	12/29/14
3	1000002812	A1401250690	01/19/15	1/19/15
4	1000003019	A1401275862	04/01/15	4/1/15

Figure 10. Sample of processed ED_events data

The LACE index was calculated on the bases of these three datasets. The process of the LACE index model is been discussed in the following chapter in section 4.1.

The dataset formed during the evaluation of LACE index (*lace_score*), *processed_inputevents, processed_edevents and processed_clinicalcodes.* went through further processing, feature engineering and merging datasets according to the new model build in this report.

3.7 Feature Engineering and Merging Dataset:

Considering the three processed datasets (*processed_inputevents*, *processed_edevents* and *processed_clinicalcodes*), eight new attributes were created. These new attributes were created with the intension to include all the information in such a way that all the attributes in the datasets are of numerical type. It is comparatively easy to build and run a machine learning model where the dataset

has all attributes with numerical type. These attributes and the description of these attributes are presented in Table 6.

Table 6. The list and description of new featured added into the dataset

<table>
<tr><th>New Attributes</th><th>Description</th></tr>
<tr><td>ed_visits_count</td><td>A total count of ED visits for the patient in the previous 6 months before the current admission.</td></tr>
<tr><td>ac_check</td><td>A new column which includes values 0 and 1, where 1 is for AC admit type</td></tr>
<tr><td>age_group</td><td>Categorically divides the age into 10 groups. Below is shown age groups and value assigned for the group.
<table>
<tr><th>Age Group</th><th>Value</th></tr>
<tr><td>39 years and below</td><td>0</td></tr>
<tr><td>40 to 49 years</td><td>1</td></tr>
<tr><td>50 to 59 years</td><td>2</td></tr>
<tr><td>60 to 64 years</td><td>3</td></tr>
<tr><td>65 to 69 years</td><td>4</td></tr>
<tr><td>70 to 74 years</td><td>5</td></tr>
<tr><td>75 to 79 years</td><td>6</td></tr>
<tr><td>80 to 84 years</td><td>7</td></tr>
<tr><td>85 to 89 years</td><td>8</td></tr>
<tr><td>90 to 94 years</td><td>9</td></tr>
<tr><td>95 to 99 years</td><td>10</td></tr>
<tr><td>100 years and above</td><td>11</td></tr>
</table></td></tr>
<tr><td>cw_group</td><td>Categorically divides the cost weight into 4 groups. Below is shown the groups and the value for the same.
<table>
<tr><th>Cost Weight Group</th><th>Value</th></tr>
<tr><td>Less than 0.5</td><td>0</td></tr>
<tr><td>0.5 to 0.9</td><td>1</td></tr>
<tr><td>1.0 to 4.9</td><td>2</td></tr>
<tr><td>5.0 to 9.9</td><td>3</td></tr>
<tr><td>10 and above</td><td>4</td></tr>
</table></td></tr>
<tr><td>ethnicity_maori</td><td>A new column which includes values 0 and 1, where 1 is for Maori ethnicity</td></tr>
<tr><td>ethnicity_pi</td><td>A new column which includes values 0 and 1, where 1 is for Pacific Island ethnicity</td></tr>
<tr><td>ethnicity_asians</td><td>A new column which includes values 0 and 1, where 1 is for Asians ethnicity</td></tr>
<tr><td>ethnicity_others</td><td>A new column which includes values 0 and 1, where 1 is for other ethnicities</td></tr>
</table>

3.7.1 Merged_selected_features:

A new dataset, *merged_selected_features* was created by merging the two datasets, *processed_inputevents* and *processed_clinicalcodes* joining at the common feature i.e. encounter ID. *Merged_selected_features* dataset also included above-mentioned new attributes in it. *Merged_selected_features* was then pre-processed to make it ready for the machine learning models to apply on.

1. The *merged_selected_features* dataset consisted of
    a. 54 attributes (see Table 7) and
    b. 180117 records originally (162404 False Readmissions and 17713 True Readmissions).

3.8 Secondary Data Pre-Processing:

1. The newly mergred dataset *merged_selected_features* have the 54 attributes originally. These attributes are mentioned in the Table 7.

Table 7. The list of 54 attributes present in the merged dataset

encounter	Patient_Identifier	admitdate	dischargedate
lengthofstay	EthnicGroup	Age	gender
Female	Admit_Type	admittype	AdmissionSource
Discharge_Destination	Costweight	DRG	hospital
Readmitted	IschaemicScore	CerebrovascularScore	StrokeScore
COPDScore	AsthmaScore	RespiratoryScore	DiabetesScore
ArthritisScore	DevelopmentalDisScore	PeripheralVascScore	RenalFailureScore
SickleCellScore	AlcoholismScore	CCI_MI	CCI_CHF
CCI_PVD	CCI_Cerebrovascular_Dis	CCI_Dementia	CCI_Chronic_Pulmonary_Dis
CCI_Connective_Tissue_Dis	CCI_Peptic_Ulcer_Dis	CCI_Mild_Liver_Dis	CCI_Diabetes_wo_Chronic_Comp_Dis
CCI_Paraplegia_Hemiplegia	CCI_Renal_Disease	CCI_Cancer	CCI_Mod_or_Severe_Liver_Dis
CCI_Metastatic_Carcinoma	CCI_HIV_AIDS	ed_visits_count	ac_check
age_group	cw_group	ethinicity_maori	ethinicity_pi
ethinicity_asian	ethinicity_others		

2. Degenerated attributes like ‘ArthritisScore’, ‘DevelopmentalDisScore’ and ‘PeripheralVasScore’ were then removed as it had only one value i.e zero for all the observations.
3. Attributes like ‘EthinicGroup’, ‘Age’, ‘gender’, ‘Admit_Type’, ‘admittype’ and ‘Costweight’ were removed as information from these attributes were reframed into a new feature within the datasets as discussed in section 3.7
4. Attributes like ‘admitdate’, ‘dischargedate’, ‘AdmissionSource’ and ‘Discharge Destination’ were also removed as the information in these attributes are contributing through a different feature i.e. ‘Readmission’. A record is said to be a True Readmission if it satisfies criteria (mentioned in section 2.1) which includes information collected from four attributes (‘admitdate’, ‘dischargedate’, ‘AdmissionSource’ and ‘Discharge Destination’).
5. Attributes like ‘encounter’, ‘Patient_Identifier’, ‘DRG’ and ‘hospital’ were removed as the information from these attributes didn’t seem to contribute in generating a risk of readmission model
6. Instances for which scores and CCIs were not recorded, were then filled with ‘0’, implicating that they are negative for all conditions.

The *merged_selected_features* datatset was deliberately structured in a format where it deals with numeric value for every attribute which contributes with some information so that it can be easily imported into Weka and explored further. After importing *merged_selected_features* dataset into Weka, following steps were executed.

7. Due to high class imbalance in our dataset, SpreadSubsample in Weka for random under sampling was applied.
8. The dataset after removing unwanted attributes and handling class imbalance, the dataset was saved as *processed_merged_selected_features*

3.8.1 Processed_merged_selected_features:

lengthofstay	ed_visits_count	ac_check	age_group	Female	ethinicity_maori	ethinicity_pi	ethinicity_asian	ethinicity_others	cw_group	IschaemicScore	CerebrovascularScore	StrokeScore
0	1	0	0	0	0	0	0	1	1	0	0	0
0	1	1	0	0	0	0	0	1	0	0	0	0
1	2	0	1	0	0	0	0	1	1	0	0	0
0	4	0	1	0	0	0	0	1	0	0	0	0

COPDScore	AsthmaScore	RespiratoryScore	DiabetesScore	RenalFailureScore	SickleCellScore	AlcoholismScore	CCI_MI	CCI_CHF	CCI_PVD	CCI_Cerebrovascular_Dis	CCI_Dementia	CCI_Chronic_Pulmonary_Dis
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0

CCI_Connective_Tissue_Dis	CCI_Peptic_Ulcer_Dis	CCI_Mild_Liver_Dis	CCI_Diabetes_wo_Chronic_Comp_Dis	CCI_Diabetes_w_Chronic_Comp_Dis	CCI_Paraplegia_Hemiplegia	CCI_Renal_Disease	CCI_Cancer	CCI_Mod_or_Severe_Liver_Dis	CCI_Metastatic_Carcinoma	CCI_HIV_AIDS	Readmitted
0	0	0	0	0	0	0	0	0	0	0	FALSE
0	0	0	0	0	0	0	0	0	0	0	FALSE
0	0	0	0	0	0	0	0	0	0	0	FALSE
0	0	0	0	0	0	0	0	0	0	0	FALSE

Figure 11. Sample of processed_merged_selected_features data

1. The dataset consists of
    a. 38 attributes (see Table 8)
    b. 180117 instances (162404 False Readmissions and 17713 True Readmissions)

    The dataset shown in Figure 11 is a sample of processed_merged_selected_features has been divided into three tables images however are in continuation.
2. The 38 attributes present in the final *processed_merged_selected_features* dataset are listed below in Table 8.

Table 8. The list of 38 attributes present in the processed merged dataset

lengthofstay	ed_visits_count	ac_check	age_group
Female	ethnicity_maori	ethnicity_pi	ethnicity_asian
ethnicity_others	cw_group	IschaemicScore	CerebrovascularScore
StrokeScore	COPDScore	AsthmaScore	RespiratoryScore
DiabetesScore	RenalFailureScore	SickleCellScore	AlcoholismScore

CCI_MI	CCI_CHF	CCI_PVD	CCI_Cerebrovascular_Dis
CCI_Dementia	CCI_Chronic_Pulmonary_Dis	CCI_Connective_Tissue_Dis	CCI_Peptic_Ulcer_Dis
CCI_Mild_Liver_Dis	CCI_Diabetes_wo_Chronic_Comp_Dis	CCI_Paraplegia_Hemiplegia	CCI_Renal_Disease
CCI_Cancer	CCI_Mod_or_Severe_Liver_Dis	CCI_Metastatic_Carcinoma	CCI_HIV_AIDS
Readmitted			

3. The detailed description of some of the important attributes in the dataset are mentioned in Table 9.

Table 9. Description of important attributes in the final dataset

Attributes	Observations
Number of total records	180117
Gender	Female = 58.69% Male = 41.30%
Age Group	39 years and below = 30.5% 40 to 49 years = 11.4% 50 to 59 years = 13.7% 60 to 64 years = 6.9% 65 to 69 years = 8.07% 70 to 74 years = 7.9% 75 to 79 years = 7.5% 80 to 84 years = 6.2% 85 to 89 years = 4.8% 90 to 94 years = 2.2% 95 to 99 years = 0.4% 100 years and above = 0.03%
Ethnic Group	Asians = 12.2% Maori = 7.5% Pacific Island = 7.3% Others = 72.8%
Length of Stay	No overnight stay = 46.3% 1 night stay = 21.1% 2 nights stay = 9.4% 3 nights stay = 6.6% 4 to 6 nights stay = 9.2% 7 to 13 nights stay = 5.06% More than 13 nights stay = 2.08%
Admit Type: AC	59.27%

Cost Weight	Less than 0.5 = 55.26% 0.5 to 0.9 = 24.02% 1.0 to 4.9 = 20.11% 5.0 to 9.9 = 0.50% 10 and above = 0.08%
Emergency Dept Visits	0 visit = 64.41% 1 visit = 19.27% 2 visits = 8.08% 3 visits = 3.64% 4 visits = 1.78% 5 visits = 0.97% 6 visits = 0.58% 7 visits = 0.36% 8 visits = 0.21% 9 visits = 0.15% More than 9 visits = 0.50%
Readmission Rate	Readmission rate = 9.8%

3.9 Data Visualization:

3.9.1 WBHB Dataset

Figure 12 shows readmission rate and no readmission rate for the raw 2015 and 2016 WDHB dataset provided for the project. The readmission rate of the given 2 years of WDHB data is 9.83%. The part in the pie chart in Figure 12, where it shows 9.83% of readmissions is then further broken down in the following section to understand unplanned 30-days hospital readmissions (U30R) records and their patters in gender and age.

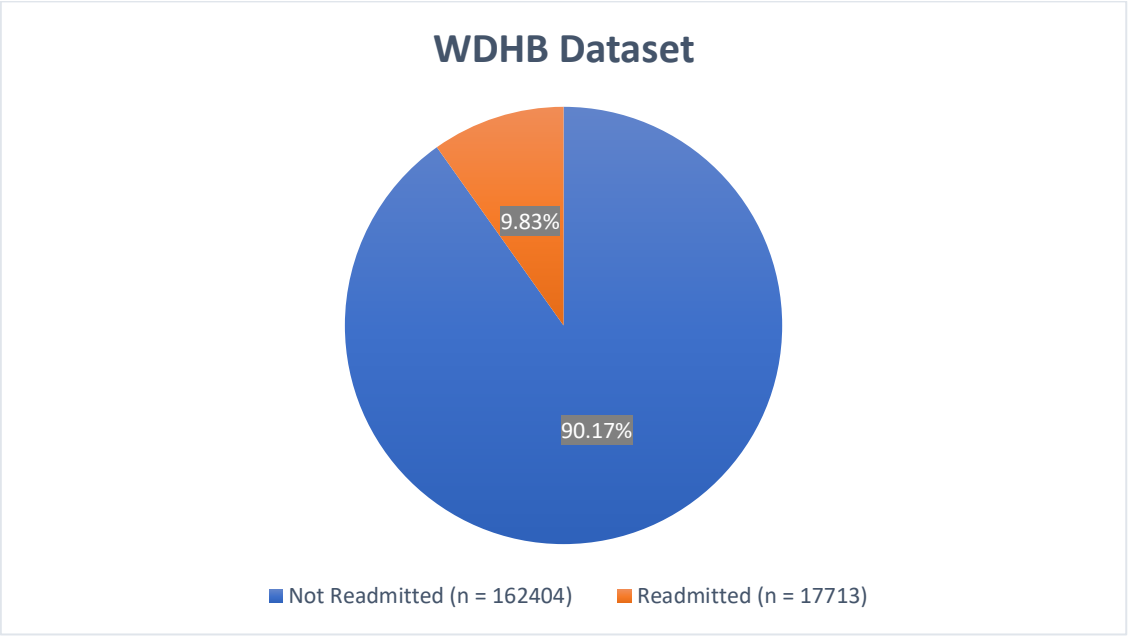

Figure 12. Pie chart of rate of readmission and no readmission for the provided WDHB data

3.9.2 Train and Test Set

The dataset of total valid records of 180117 instances, were split into train set and test set with the ration of 70:30. The train set has 70% of the total records i.e. 126081 and test set has rest 30% of the records i.e. 54036. This split was done using unsupervised instance weka filter, Resample. Figure 13 shows the split percentage and the number of instances in training set and test set.

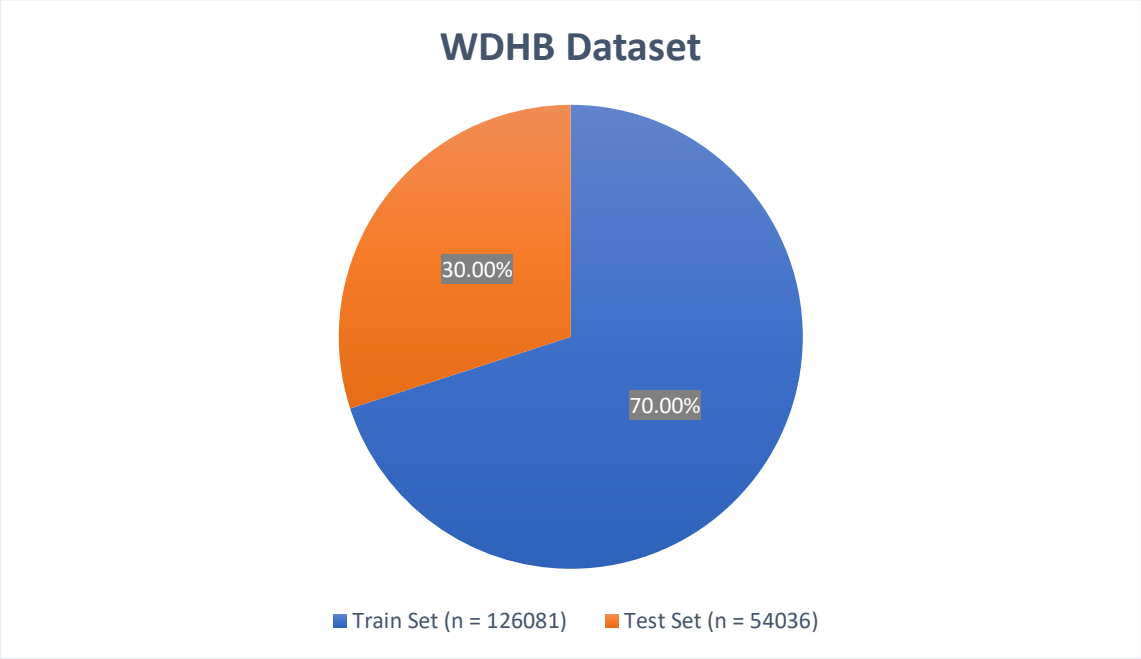

Figure 13. Pie chart of train dataset and test dataset ratio used in this project

After splitting the data into train and test set, the readmission rate was calculated for both the set individually and the following points were observed:

- Training Set:
    - Readmission Rate = 9.86% (n = 12436)
- Test Set:
    - Readmission Rate = 9.77% (n = 5277)

Due to high class imbalance and high instances, the training set which has readmission rate of 9.86% (n = 126081) was gone through under-sampling by applying supervised instances weka filter, Spread Subsample. Under-sampling is a technique for resampling, which leads to random information loss for the majority class (Chawla, 2009). Under-sampling was done in order to have same ratio for true class and false class in the dataset, so that the building model is not biased. After under-sampling, the train set has readmission rate of 50% (n = 5277) with the total number of instances of 10554.

3.9.3 Gender Distribution from Readmission Data

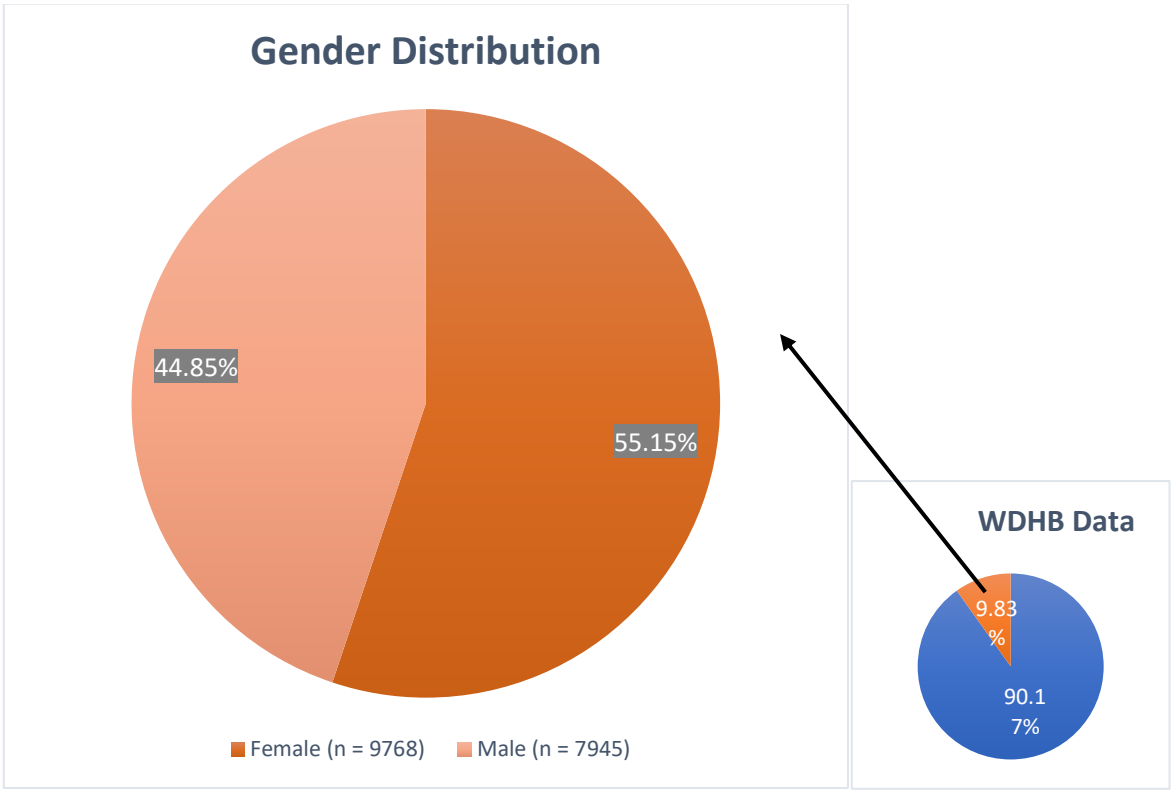

Figure 14. Pie chart of the dataset showing percentage of readmitted males and females

Out of 17713 patients in the WDHB data for 2015 and 2016, who were readmitted in the hospital within 30-days of the initial discharge, 9768 were females and 7945 were males. The pie chat in Figure 14 shows the percentage division of readmitted patients in regard to gender of the patient. Seeing the Figure 14, it is concluded that there is not a major difference between males and females true U30R.

3.9.4 Age Distribution from Readmission Data

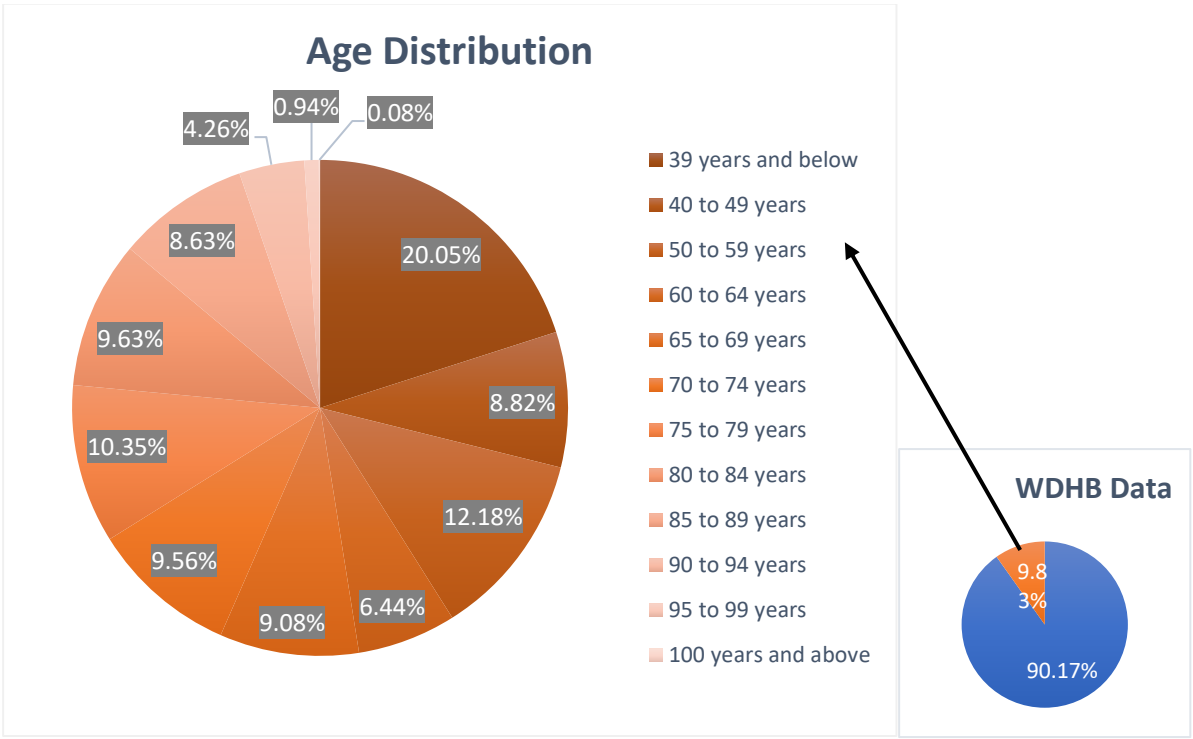

Figure 15. Pie chart of the dataset showing percentage of readmitted patients in different age group

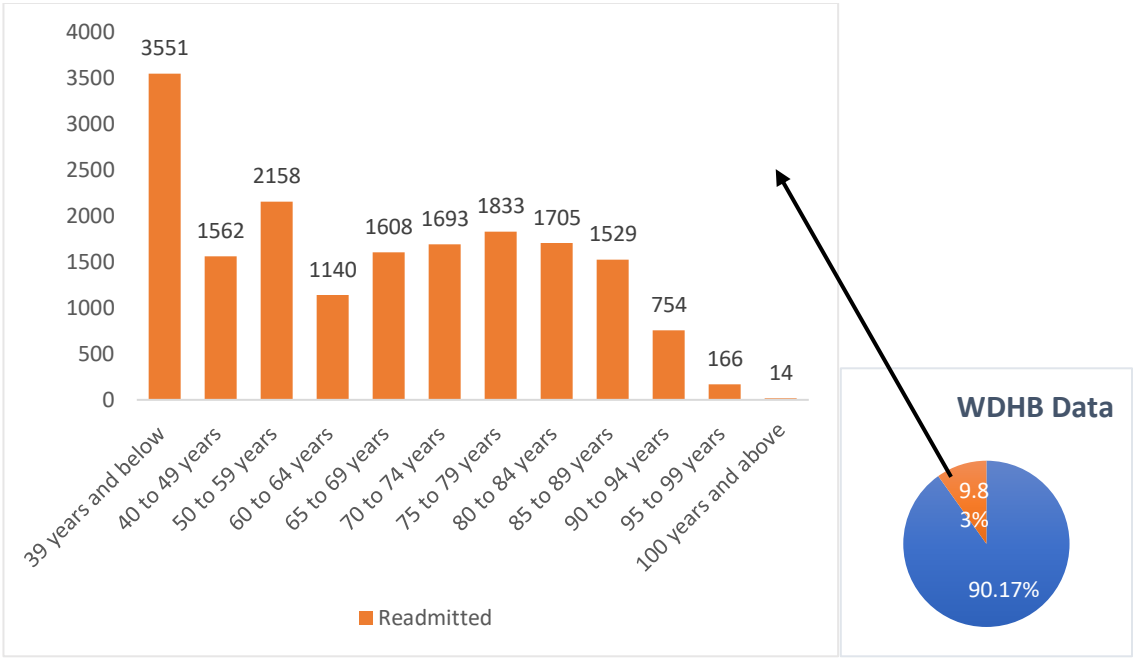

Figure 16. Bar chart of the dataset showing numbers of readmitted patients in different age group

The age has been divided into 12 different age groups. The first age group includes patients of age 39 years and below. The next two age range of 10 years each i.e. 40 to 49 years and 50 to 59 years. The following ages up to 99 years were divided into the range of 5 years each and then the last age group includes patients with 100 years and above. The pie chart in Figure 15 provides an overall view on the population by showing the percentage split for true U30R in 12 age groups. The bar chart in Figure 16 gives an in-depth information on the total number of readmitted patients in the 12 age groups for true U30R.

Chapter 4 – Methodologies and Prediction Model

After the final pre-processing of the merged dataset, several machine learning models and LACE Index model were applied. These models have been discussed in the following sections.

4.1 LACE Index Mode

In response to the challenge of detecting high readmission risk patient, (van Walraven et al., 2010) developed a predictive risk model, LACE Index.

The study was performed on data of 4000 patients extracted from 11 hospitals in Ontario, Canada between October 2002 and July 2006 (van Walraven et al., 2010). It created an index to predict early death or unplanned readmissions after discharge from hospitals. The LACE index uses four attributes which were:

- length of stay (“L”),
- acuity of the admission (“A”),
- comorbidity index score (“C”) and
- emergency department use which measured as the number of visits in the six months before admission (“E”)

Scores using the LACE index ranged from 0 to 19; 0 being 2.0% expected risk of death or unplanned readmission within 30 days and 19 being 43.7% expected risk of death or unplanned readmission within 30 days. The final score is calculated by adding the points for each attribute which ranged from 0 to 19. The index was validated externally using 1000000 randomly selected administrative data and reported accuracy of 0.68 in AUC.

LACE index has its own limitations as well which is discussed below (van Walraven et al., 2010):

- The calculated index could not be reliable in-patient populations which were not involved in its derivation.

- Further work is required to detect additional factors which would increase the discrimination or accuracy of the index.
- Clinicians were expected to find it difficult to commit to memory the point system and its expected risk.

LACE Index was derived and validated as an easily used index that is moderately discriminative and accurate for predicting the risk of early death or unplanned readmission after discharge from hospital to the society (van Walraven et al., 2010). It was also proposed that this index can be used with primary and administrative data.

4.2 Risk Prediction Methods

In this section we give an overview of the machine learning algorithms used in this study. For prediction of risk of readmission, the following machine learning models were applied on Weka and were compared:

- Support Vector Machine (SVM)
- Random Forest
- AdaBoost
- Decision Table
- REP Tree

These algorithms were chosen as they are quite popular in the previous studies done on prediction problems. The accuracy and root mean square error of the following algorithms are quite similar to each other. However, we cannot rely on the accuracy as it can be considered as misleading due tot the nature of the data present i.e. highly imbalanced data. These results were discussed with the intension to bring the attention on amazingly similar accuracy however significantly different discrimination score.

The discrimination score (c-statics) of all these algorithms are discussed in Chapter 5 where we compare these models on the bases of our performance metrics i.e. discrimination score. We will learn more about our performance metrics in section 4.3.

4.2.1 Support Vector Machine (SVM)

It is a supervised machine learning algorithm used for classification problems. It is a discriminative classifier formally defined by a separating hyperplane. In this model, each data item as a point in the n-dimensional space, where n is the number of attributes, with the value of each feature being the value of the predictor coordinate. Then we perform classification by finding the hyperplane that differentiate the two classes very well. We used 10-fold cross validation technique to assess the predictive performance of the model generated. This classifier achieved an accuracy of 66.58% with the root mean squared error of 0.578. Out of all the machine learning models applied, SVM has the highest accuracy but also the highest root mean squared error. In (Yu et al., 2015), SVM was applied on USA data, where the discrimination score achieved was 0.60. Moreover, the discrimination score achieved for WDHB's data is 0.66, which is remarkably close.

4.2.2 Random Forest

Random Forest is an extension of decision tree in an assembled way. It is a combination of tree predictors where every tree depends on the values of a random vector sampled independently and with the same distribution for all trees within the forest. In standard trees, every node is split using the most effective split among all variables. in a random forest, every node is split using the best among a subset of predictors randomly chosen at that node. This somewhat counterintuitive strategy seems to perform very well compared to several other classifiers and is powerful against overfitting (Mesarić & Šebalj, 2016). The error rate of a random forest depends on the strength of each tree and correlation between any two trees. This classifier achieved an accuracy of 66.19% with the root mean squared error of 0.480.

4.2.3 AdaBoost

It is a popular and effective prediction algorithm. The essence of the AdaBoost algorithm is that a combination of many 'weak' learning algorithms, each performing just slightly better than a random guessing algorithm, will generate a 'string' learning algorithm. It is been observed in this project as well that AdaBoost outperforms other predictors such as SVM (support vector machine) which is considered as a powerful algorithm and widely used in biological literatures (Niu, Cai, Lu, Li, & Chou, 2006; Niu et

al., 2008). AdaBoost holds a high potential for improving the quality in predicting feature. This classifier achieved an accuracy of 65.55% with the root mean squared error of 0.466.

4.2.4 Decision Table

It uses recursive partitioning algorithm, which works by splitting the dataset recursively, where the subset which arise from the split are further split till a predetermined termination occurs. At each step, the split is made based on the independent variable that results in the largest possible reduction in heterogeneity of the predicted variables. We used Information Gain to determine the split at each level. An inherent problem with the decision tree is that, it tends to over fit the data, and to overcome this issue, we have pruned the decision tree to get a generalized model. This classifier achieved an accuracy of 65.90% with the root mean squared error of 0.464.

4.2.5 REP Tree

Reduces Error Pruning (REP) Tree Classifier is a fast decision tree learning model and is based on the principle of computing the information gain with entropy and minimizing the error arising from the variance. REPTree is mixture of decision tree and linear regression algorithm, where each leaf node corresponds to a linear regression algorithm. This model constructs the regression/decision tree using variance and information gain. Also, this model prunes the tree using reduced-error pruning with back fitting technique. At the start of the model preparation, it sorts the values of numeric attributes once (Lakshmi Devasena, 2014). This classifier achieved an accuracy of 66.42% with the root mean squared error of 0.464. Among the five models, this model shows the best accuracy and root mean squared error pair.

4.3 Performance Metrics

From the section 2.3 of this report where 13 studies were summarized in Table 3, it was concluded that comparing machine learning models by their discrimination value for prediction of risk of hospital readmission would be most appropriate.

Even through in the previous section we discussed about the accuracy and root mean squared error by each model, we cannot trust the accuracy when choosing the best model as it could be misleading. Due to the nature of the real-time data, where instances of readmissions are evidently very low compare to the patients who were not readmitted, the accuracy of the model will be a result of highly imbalanced data. Hence, the comparison of the significance of the models will be done on the bases of discrimination score.

4.3.1 Discrimination: ROC area

The receiver operating characteristic (ROC) analysis is used to quantify the accuracy in discrimination between two states, in this case we can refer to 'readmitted' and 'not-readmitted'. A ROC curve is based on the notion of a 'separator' scale, on which readmitted, and not-readmitted results forms a pair of overlapping distribution. The complete separation of the two underlying distributions implies a perfectly discriminating test while complete overlap implies no discrimination (Hajian-Tilaki, 2013)

. The ROC curve shows the relationship between the true positive rate (sensitivity) and false positive rate (1- the specificity). These results are plotted against each other at all risk cut off levels and the area under the curve which is also known as curve statistic (C-statistic), is a value that lies between 0 and 1 so one can compare the sensitivities and specificities between models. If the ROC curve is greater than 0.5, it can be concluded that the model is performing better than a random guess. In Figure 17, the ROC curve is plotted with true positive rate (TPR) against the false positive rate (FPR) where TPR is on y-axis and FPR is on the x-asis.

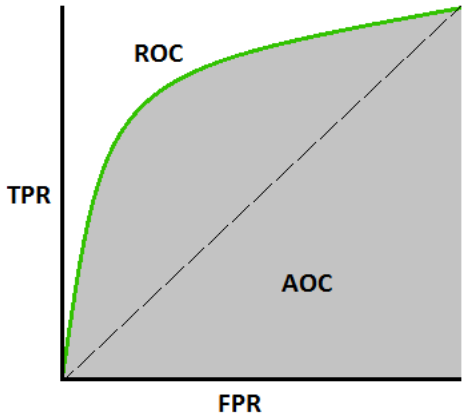

Figure 17. AUC ROC Curve (Narkhede)

4.4 Cross Validation

In order to test the performance of a machine learning model on the data, it is important to assess the error rate on a dataset (test data) which had no relation with the dataset (train data) which was used to during the formation the model. It is believed that the larger the training data, the better the model (Witten, Frank, Hall, & Pal, 2016). Hence, our train set was 70% of the whole dataset and the rest 30% of the data was used as a test dataset. During each model application, 10-fold cross-validation was performed, as 10 folds is the standard method as it results in the best error estimate. This splits the data into 10 groups where $9/10^{th}$ of the data is used of training and $1/10^{th}$ used for testing and then it is repeated ten times till each instance was a part of testing once. Applying cross validation on the dataset effects in an error rate with a small standard deviation, which is then decreased as the validation is repeated 10 times. After calculating the final performance measures from the cross validation, the results to the model on the train and test dataset were compared.

4.5 Bagging Techniques

In order to make reliable decisions, numerous learning techniques are combined to form an ensemble of models. Bagging is one of the most effective computationally intensive procedures to improve on unstable estimators or classifiers, useful especially for high dimensional data set problems (Bühlmann & Yu, 2002). Bagging is a simple and effective way to reduce the error rate of many classification learning algorithms (Domingos, 1997). Hence, after running five different machine learning models (mentioned in section 4.2) on the train dataset, bagging was applied on the best model out of those five models, i.e. REPTree. This model after bagging achieved an accuracy of 66.28% with the root mean squared error of 0.462, which is similar to the model without bagging. It was then applied on both train and test dataset, to validate the model's results.

Chapter 5 - Results and Outcomes

In this chapter, the results of the five machine learning models will be compared, and the best model will be discussed further.

5.1 Comparing different models with each other

The five machine learning models (Decision Table, AdaBoost, Random Forest, SVM and REPTree) were applied on the train set and the discrimination scores were compared (see Table 10). All the model's ROC area was greater than 0.5, which means all the models were performing better than a random guess. Out of these five, REPTree showed the best discrimination score of 0.717 ROC area. Therefore, bagging technique was combined with REPTree model and applied on the train set, which gave better discrimination score i.e. 0.721 ROC Area.

As discussed, REPTree + Bagging performed significantly better than other models on the train set. The same model (REPTree + Bagging) was applied on test set which has imbalanced class and compared it with the score obtained on the train set (see Table 11). The ROC area for the test set was similar i.e. 0.728 hence, REPTree + Bagging was selected as the best model and named **RHR-30** (Risk of Hospital Readmission within 30-days).

Table 10. Comparation of discrimination score and time taken on the train dataset by different models

	Decision Table (train)	**Ada Boost (train)**	**Random Forest (train)**	**SVM (train)**	**REP Tree (train)**	**REP Tree + Bagging (train)**
ROC Area	0.712	0.707	0.687	0.666	0.717	0.721
Time Taken (seconds)	7.04	1.81	16.51	58.15	0.86	6.19

Table 11. Comparation of discrimination score and time taken on the train and test dataset by the best model, RHR-30 (Risk Of 30-days Hospital Readmissions) Model

	REP Tree + Bagging (train)	**REP Tree + Bagging (test)**
ROC Area	0.721	0.728
Time Taken (seconds)	6.19	24.76

5.1.1 RHR-30 – Confusion Matrix and ROC curve

The confusion matrix for RHR-30 model (see Figure 18), describes the performance on the test data for which readmitted instances are known. Out of 54,036 records in the test data,

- 297 records were predicted positive and were true (True Positive)
- 48431 records were predicted negative and were true (True Negative)
- 328 records were predicted positive but were false (False Positive – Type 1 Error)
- 4980 records were predicted negative but were true (False Negative – Type 2 Error)

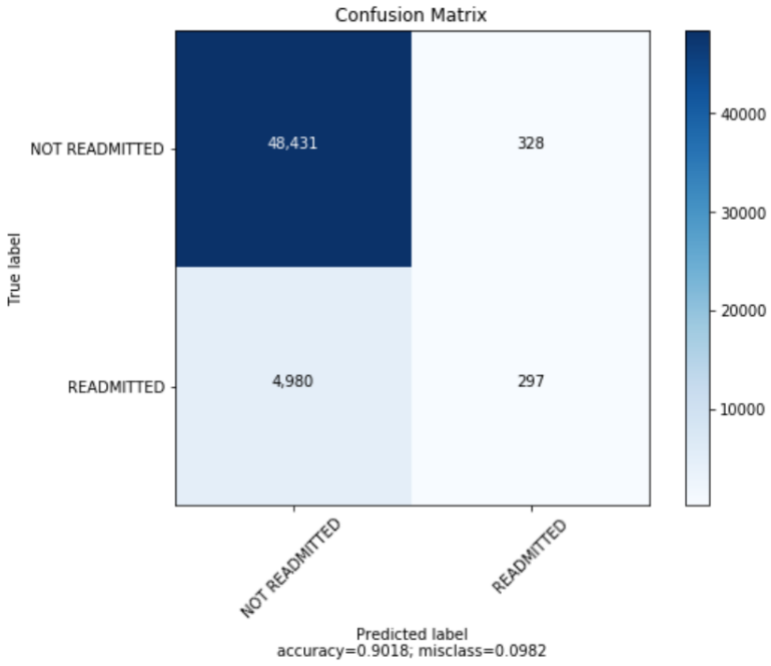

Figure 18. Confusion Matrix shows actual and predicted readmitted and not readmitted patients

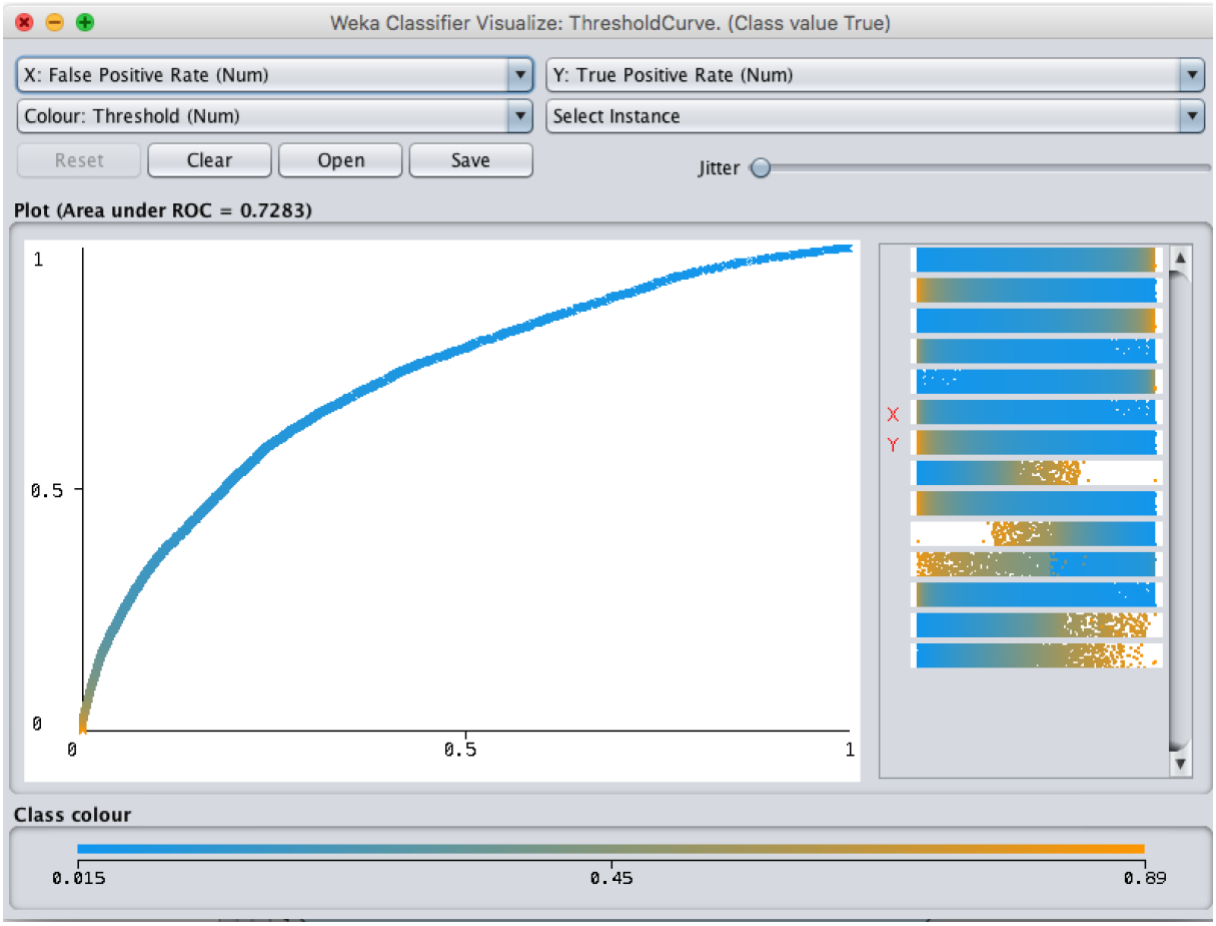

Figure 19. ROC curve for RHR-30 model

As previously discussed in Chapter 4 section 4.3.1, the ROC curve shows the relation between the true positive rate (sensitivity) and false positive rate (1- the specificity). These results are plotted against each other at all risk cut off levels and the area under the curve which is also known as curve statistic (C-statistic), is a value that lies between 0 and 1 so one can compare the sensitivities and specificities between models. For RHR-30, the ROC curve is 0.7283 (see Figure 19), which is the best performance so far in this project.

5.1.2 RHR-30 - Significant Attributes

In order to analyze which attributes, contribute more for RHR-30, Information Gain attribute evaluator with ranker method was applied in Weka. The weightage of top 20 significant attributes can be seen in Figure 20. The top four significant attributes were:

- Count of Emergency Department visits in the six months before admission
- Length of the stay
- Age
- Acuity of the admission

The emergency department visits count attribute dominates the most. It is interesting to consider that the intervention which will reduce re-admissions will also end up reducing emergency visits slightly if not considerably. We can observe the dependable changes once the model is in use for months in a healthcare.

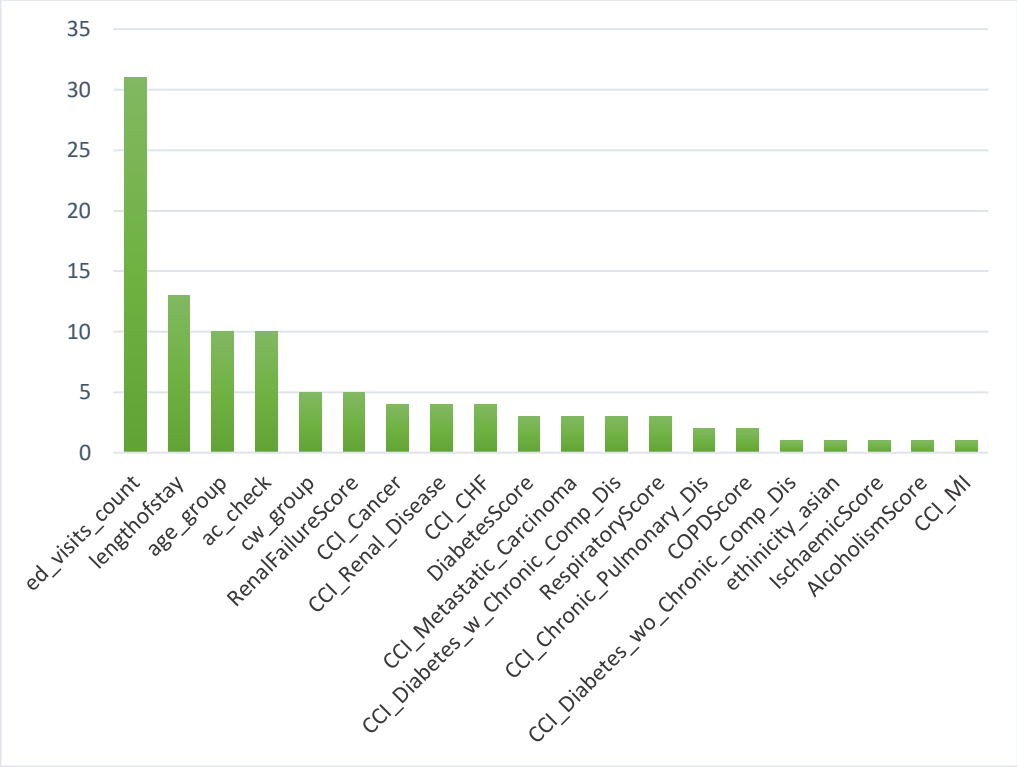

Figure 20. Pareto chart of significant variables includes in the predictive model, RHR-30

5.2 Comparing RHR-30 with LACE Index model

After selecting the best model, RHR-30, the model was compared with a famous pre-existing model, LACE.

5.2.1 ROC Area

As mentioned in chapter 4 section 4.1, LACE was developed on data from Ontario, Canada hence three models' results were compared: LACE on Canada's data, LACE on WDHB's data and RHR-30 on WDHB's data. Table 12 shows the comparison in the discrimination score (ROC area) and Figure 21 shows the bar graph for the compared scores.

Table 12. Comparation of discrimination score of LACE Index on original data i.e. Canada's healthcare's data, LACE Index on WDHB data and RHR-30 model on WDHB data

	LACE on original (Canada's) data	**LACE on WDHB data**	**RHR-30 on WDHB data**
ROC Area	0.68	0.64	0.72

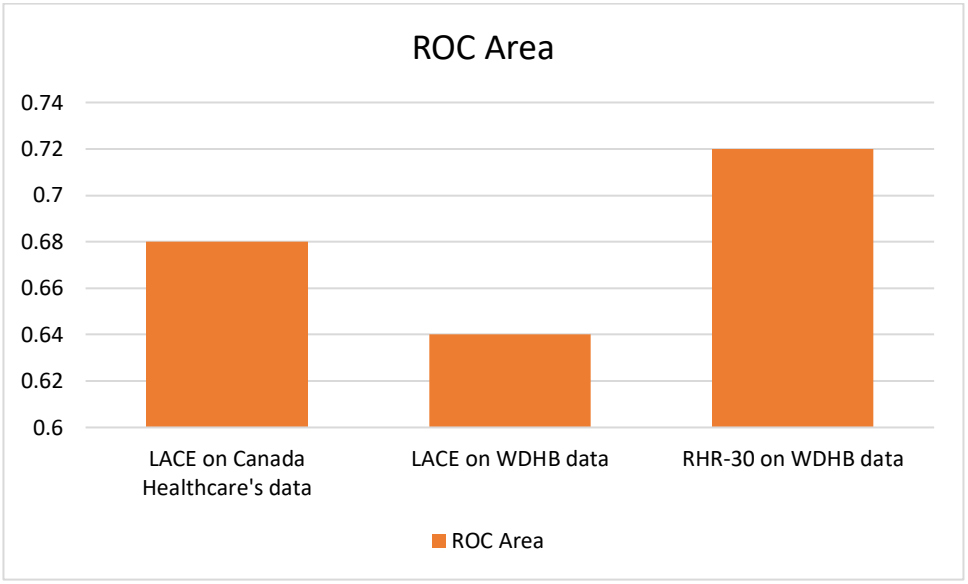

Figure 21. Bar graph of discrimination score of LACE Index on original data i.e. Canada's healthcare's data, LACE Index on WDHB data and RHR-30 model on WDHB data

As seen, performance of LACE Index model on WDHB's data was poorer than LACE Index model on original data. It can be justified as LACE Index model was developed on Canada's data whereas applied on New Zealand's data.

5.2.2 Readmission Rate

As mentioned before, the actual readmission rate for the given WDHB's data is 9.83%. After applying LACE Index model and RHR-30 model, the readmission rate was calculated keeping the best threshold of given WDHB's data (see Table 13). The readmission rate for LACE index model was 10.90% with the threshold of 10 with the minimum and maximum threshold score of 0 and 19 respectively. Similarly, for RHR-30 model the readmission rate was 9.83% with the score threshold of 0.02 with the minimum and maximum threshold score of 0 and 1. It is important to note that the threshold for different healthcare will be different. The fact that we were aware of the readmission rate for WDHB's data in 2015 and 2016, the threshold was adjusted to get the accurate result. However, while doing the real-time prediction on real-world data, it will be important to study the different threshold and continue with the best one.

Table 13. Comparison of LACE model's and RHR-30 model's readmission rate with actual readmission rate

	Actual	**LACE**	**RHR-30**
Readmission Rate	**9.83%**	10.90%	9.83%

Chapter 6 – Conclusion and Future Work

6.1 Conclusion

The unplanned readmissions are those which are acute (non-elective) admissions. Elective admissions are those admissions when a doctor requests a bed be reserved for a patient on a particular day, hence acute admissions excludes planned admissions. Readmissions are unplanned overnight stay in the hospital within a specific period of time from an initial admission. Hospital readmissions are upsetting to patients and costly to the healthcare systems too. Some of the readmissions that do occur are avoidable, hence whether its New Zealand or any other country in the world, readmission rates are one of the very important health quality measure. There are various ways to identify patients at the high risk of hospital readmissions which are, clinical knowledge, threshold modelling and predictive modelling (Purdy, 2010). This project report focuses on developing a prediction model for unplanned hospital readmissions within 30 days of discharged using Waitemata District Health Board (WDHB) records. The time period of 30 days was chosen because it is considered the most likely time setting that the two occasions can be relatable. As the time frame increases the chances of picking up admissions unrelated to the initial admission also increases.

There are several factors that can contribute to unplanned readmissions. These factors can be divided into following categories (Holloway & Thomas, 1989):

- Medical factors: the data is accessible from secondary sources (e.g., discharge abstracts and claims forms) whenever needed, such as sex, age, diagnosis, and practices performed.
- Other medical factors: e.g., self-reported global and functional health status.
- Nonmedical factors: e.g., living arrangements, marital status, care accessibility, social factors, cultural factors and insurance coverage.

Preventable factors i.e. factors which was under the control of the hospital, include: surgical complications, errors related to medication and poor discharge measures which do not properly involves patients, their relatives, general practitioners or aged-care worker (Maali et al., 2018).

In this project, a general structure for hospital-specific and all-cause model for risk of hospital readmission prediction was developed and investigated. The real-world hospital data structure has the advantage that it can take all attributes the hospital has collected for its patient's population. Five machine learning algorithms were applied, and all five algorithms achieved the prediction accuracy rate around 65%. However, considering the real-world data is highly imbalance in nature, hence, choosing the best model on their accuracy rate would be misleading. Therefore, the best model was selected on the bases of discrimination score i.e. ROC area. Most of the studies included in the literature review, also selected discrimination score as their performance metrics.

RHR-30 model was built on WDHB's data for calculating risk of hospital readmission within 30 days of discharge. RHR-30 out-performs all the other machine learning algorithms applied in this project and also out-performs LACE model, by having the highest ROC area (discrimination score) and performs much better compare to the well-known pre-existing risk of readmission model i.e. LACE. The model yielded a ROC area of 0.72 – a modest value for a clinical predictive rule.

The LACE Index model is based on length of stay, acuity of admission, comorbidity index score and emergency department visits in previous six months. It was observed that RHR-30 model's top attributes are quite similar to LACE Index model's attributes, as the top four attributes on which RHR-30 is dependent are emergency department visits in previous six months, length of stay, age and acuity of admission. The emergency department (ED) visits counts dominates the most compare to all other attributes. Other than ED visits, the two attributes: the length of stay and the admission acuity, holds the most important role in building a model for calculating risk of hospital readmissions.

It was also found that applying a mixed-method approach was valuable and extra efforts are needed during selection of risk factors/features that are of high-quality data, certainly accessible, and can be generalized across multiple populations. To create a good accurate predictive model which is multidimensional and responsible is dependent on several factors, including, however not limited to, the quality and accessibility of data, the power to reproduce the findings beyond training dataset, and the balance between a tight and comprehensive prediction model.

However, we cannot blindly rely on 30-days hospital readmissions model for several reasons (Joynt & Jha, 2012; Purdy, 2010; Shulan et al., 2013).

1. Some proportion of readmissions at 30 days after the initial discharge are avoidable. Community-level factors which are outside the hospital's control are also responsible for hospital readmissions rates.
2. It is a debate whether readmissions always suggest poor hospital quality, as hospital readmission rates can also be the result of low mortality rates or good access to hospital care.
3. It is worth improving care coordination, implementing personalized health care programs, structured discharge planning and focusing on making more aimed and effective policies for achieving the goal.
4. Applying models from one healthcare system to another healthcare system could result in misspecification as the same model will shows different results for two completely different data. Though in our project when the LACE Index was generated on Canada's data the discrimination score was 0.68 whereas when the same LACE index model was applied on New Zealand's data the score was 0.64, which is remarkably close however still not the same. Also, RHR-30 shows much better results on New Zealand's data compare to LACE model.

6.2 Future Work

An interesting investigation that can be done with the information that we have within this project is to check the day of discharge and the day of readmission and observe the probability of discharging a patient on Friday and readmitting the same patient on the following Monday. Also, doing a primary diagnosis by spotting top five discharge diagnoses common within the readmissions, to discover and get familiar with common traits in those specific conditions' readmissions cases.

The next plan in continuation to the extend the project is to increase this structure to other risk modeling such as death, hospital-acquired infection, etc. It would be interesting to have more informative dataset which includes attributes like insurance data, mental health data, social determinacies like smoking,

drinking, etc. in the data and observe their significance in hospital readmissions. Furthermore, to note all the discharges who had a behavior health condition and who didn't and comprehend if there is an important difference.

Another stage that we can include in future is to understand, according to the healthcare when is the best time to activate the model; whether it's at discharge or after discharge or prior to discharge. If healthcare prefers to active the model prior to discharge, it should be considered that data for some comorbidity factors will not be available, hence have to plan a strategy to deal with that limitation.

The best practice to use predictive modelling to calculate risk of hospital readmission is to refresh the analysis and fill in the gaps periodically.

References:

Baillie, C. A., VanZandbergen, C., Tait, G., Hanish, A., Leas, B., French, B., . . . Umscheid, C. A. (2013). The readmission risk flag: Using the electronic health record to automatically identify patients at risk for 30-day readmission. Journal of hospital medicine, 8(12), 689-695.

Berry, J. G., Gay, J. C., Maddox, K. J., Coleman, E. A., Bucholz, E. M., O'Neill, M. R., . . . Hall, M. (2018). Age trends in 30 day hospital readmissions: US national retrospective analysis. bmj, 360, k497.

Blunt, I., Bardsley, M., Grove, A., & Clarke, A. (2014). Classifying emergency 30-day readmissions in England using routine hospital data 2004–2010: what is the scope for reduction? Emerg Med J, emermed-2013-202531.

Bühlmann, P., & Yu, B. (2002). Analyzing bagging. The Annals of Statistics, 30(4), 927-961.

Chassin, M. R., Loeb, J. M., Schmaltz, S. P., & Wachter, R. M. (2010). Accountability measures—using measurement to promote quality improvement: Mass Medical Soc.

Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook (pp. 875-886): Springer.

Choudhry, S. A., Li, J., Davis, D., Erdmann, C., Sikka, R., & Sutariya, B. (2013). A public-private partnership develops and externally validates a 30-day hospital readmission risk prediction model. Online journal of public health informatics, 5(2), 219.

Domingos, P. M. (1997). Why Does Bagging Work? A Bayesian Account and its ImplicationsCiteseer. Symposium conducted at the meeting of the KDD

Escobar, G. J., Ragins, A., Scheirer, P., Liu, V., Robles, J., & Kipnis, P. (2015). Nonelective rehospitalizations and postdischarge mortality: predictive models suitable for use in real time. Medical care, 53(11), 916.

Glass, D., Lisk, C., & Stensland, J. (2012). Refining the hospital readmissions reduction program. Washington, DC: Medicare Payment Advisory Commission.

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian journal of internal medicine, 4(2), 627.

Health, M. o. (2016). District Health Board Sector Financial Performance for year to date 30 September 2016.

Holloway, J. J., & Thomas, J. W. (1989). Factors influencing read mission risk: Implications for quality monitoring. Health care financing review, 11(2), 19.

Institute, N. C. (2016). NCI dictionary of cancer terms.

Joynt, K. E., & Jha, A. K. (2012). Thirty-day readmissions—truth and consequences. New England Journal of Medicine, 366(15), 1366-1369.

Joynt, K. E., & Jha, A. K. (2013). A path forward on Medicare readmissions. New England Journal of Medicine, 368(13), 1175-1177.

Khan, A., Malone, M. L., Pagel, P., Vollbrecht, M., & Baumgardner, D. (2012). An electronic medical record-derived real-time assessment scale for hospital readmission in the elderly. WMJ, 111(3), 119-123.

Krumholz, H. M., Lin, Z., Drye, E. E., Desai, M. M., Han, L. F., Rapp, M. T., . . . Normand, S.-L. T. (2011). An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. Circulation: Cardiovascular Quality and Outcomes, 4(2), 243-252.

Lakshmi Devasena, C. (2014). Comparative analysis of random forest, REP tree and J48 classifiers for credit risk prediction Symposium conducted at the meeting of the International Journal of Computer Applications (0975-8887), International Conference on Communication, Computing and Information Technology (ICCCMIT-2014)

Lee, E. W. (2012). Selecting the best prediction model for readmission. Journal of Preventive Medicine and Public Health, 45(4), 259.

Maali, Y., Perez-Concha, O., Coiera, E., Roffe, D., Day, R. O., & Gallego, B. (2018). Predicting 7-day, 30-day and 60-day all-cause unplanned readmission: a case study of a Sydney hospital. BMC medical informatics and decision making, 18(1), 1.

Mesarić, J., & Šebalj, D. (2016). Decision trees for predicting the academic success of students. Croatian Operational Research Review, 7(2), 367-388.

Narkhede, S. Understanding AUC-ROC Curve. Retrieved from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
Niu, B., Cai, Y.-D., Lu, W.-C., Li, G.-Z., & Chou, K.-C. (2006). Predicting protein structural class with AdaBoost learner. Protein and peptide letters, 13(5), 489-492.
Niu, B., Jin, Y.-H., Feng, K.-Y., Lu, W.-C., Cai, Y.-D., & Li, G.-Z. (2008). Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. Molecular diversity, 12(1), 41.
Purdy, S. (2010). Avoiding hospital admissions. What does the research evidence say, 7-8.
Shulan, M., Gao, K., & Moore, C. D. (2013). Predicting 30-day all-cause hospital readmissions. Health care management science, 16(2), 167-175.
Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., . . . Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology (Cambridge, Mass.), 21(1), 128.
Stiglic, G., Wang, F., Davey, A., & Obradovic, Z. (2014). Pediatric readmission classification using stacked regularized logistic regression modelsAmerican Medical Informatics Association. Symposium conducted at the meeting of the AMIA Annual Symposium Proceedings
Swinscow, T., & Campbell, M. (1997). Study design and choosing a statistical test. Statistics at Square One. London: BMJ Publishing Group.
van Walraven, C., Dhalla, I. A., Bell, C., Etchells, E., Stiell, I. G., Zarnke, K., . . . Forster, A. J. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. Canadian Medical Association Journal, 182(6), 551-557.
van Walraven, C., Wong, J., & Forster, A. J. (2012a). Derivation and validation of a diagnostic score based on case-mix groups to predict 30-day death or urgent readmission. Open Medicine, 6(3), e90.
van Walraven, C., Wong, J., & Forster, A. J. (2012b). LACE+ index: extension of a validated index to predict early death or urgent readmission after hospital discharge using administrative data. Open Medicine, 6(3), e80.
van Walraven, C., Wong, J., Forster, A. J., & Hawken, S. (2013). Predicting post-discharge death or readmission: deterioration of model performance in population having multiple admissions per patient. Journal of evaluation in clinical practice, 19(6), 1012-1018.
Wakefield, D. S., & Mehr, D. R. (2013). Risk factors for all-cause hospital readmission within 30 days of hospital discharge. JCOM, 20(5).
Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann.
Yu, S., Farooq, F., Van Esbroeck, A., Fung, G., Anand, V., & Krishnapuram, B. (2015). Predicting readmission risk with institution-specific prediction models. Artificial intelligence in medicine, 65(2), 89-96.
Zhou, H., Della, P. R., Roberts, P., Goh, L., & Dhaliwal, S. S. (2016). Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. BMJ open, 6(6), e011060.

Appendix:

Appendix 1:

Since the clinical_codes raw dataset has a high number of attributes, therefore the dataset has been divided into two keeping two attributes: patient identifier and encounter identifier (highlighted in yellow) common in both the screenshots of sample of raw clinical_codes dataset. The screenshot of the sample of raw clinical_codes can be seen in Appendix 1.1 and Appendix 1.2

Appendix 2:

Since the clinical_codes processed dataset has a high number of attributes, therefore the dataset has been divided into two keeping one attributes: encounter identifier (highlighted in yellow) common in both the screenshots of sample of processed clinical_codes dataset. The screenshot of the sample of processed clinical_codes can be seen in in Appendix 2.1 and Appendix 2.2

Patient_Identifier	PIMS5Encounter	ICDCode	ICD_Description	CCI_MI	CCI_CHF	CCI_PVD	CCI_Cerebrovascular_Dis	CCI_Dementia	CCI_Chronic_Pulmonary_Dis	CCI_Connective_Tissue_Dis	CCI_Peptic_Ulcer_Dis	CCI_MiMd_Liver_Dis	CCI_Chronic_Comp_Dis	Chronic_Comp_Dis	CCI_Paraplegia_Hemiplegia	CCI_Renal_Disease	CCI_Cancer	CCI_Mod_or_Severe_Liver_Dis	CCI_Metastatic_Carcinoma	CCI_HIV_AIDS
10000008616	H1000325917	F101	Mental and behavi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10000008616	H1000325917	E1172	Type 2 diabetes mi	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
10002256757	H1000333101	F101	Mental and behavi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10002129009	H1000393106	E119	Type 2 diabetes mi	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
10009919534	H1001816556	F102	Mental and behavi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10019803380	H1401399864	F101	Mental and behavi	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10003969227	H1401463791	J069	Acute upper respir	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10000728008	H1401485390	E1172	Type 2 diabetes mi	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0

Appendix 1.1 Sample of raw clinical_codes data – Part 1

Patient_Identifier	HVMSEncounter	sequencenumber	DRG	DRG_FirstLetter	Primary_Dx	Primary_Dx_Desversion_c	Condition_Group	IschaemicScore	CerebrovascularScore	StrokeScore	COPDScore	AsthmaScore	RespiratoryScore	DiabetesScore	ArthritisScore	DevelopmentalDisScore	PeripheralVascScore	RenalFailureScore	SickleCellScore	AlcoholismScore
1000008616	H1000329917	3	U61B	U	F209	10_8_ Schizophrenia, uns	Alcoholism	0	0	0	0	0	0	0	0	0	0	0	0	1
1000008616	H1000329917	2	U61B	U	F209	10_8_ Schizophrenia, uns	Diabetes	0	0	0	0	0	0	1	0	0	0	0	0	0
1000296757	H1000533101	2	U61B	U	F201	10_8_ Hebephrenic schiz	Alcoholism	0	0	0	0	0	0	0	0	0	0	0	0	1
1001219009	H1000793106	2	U61B	U	F205	10_8_ Residual schizoph	Diabetes	0	0	0	0	0	0	1	0	0	0	0	0	0
1000919534	H1001181656	3	U61B	U	F200	10_8_ Paranoid schizoph	Alcoholism	0	0	0	0	0	0	0	0	0	0	0	0	1
1001980380	H1401399864	3	B81B	B	F798	10_8_ Unspecified menti	Alcoholism	0	0	0	0	0	0	0	0	0	0	0	0	1
1000396927	H1401463791	4	U61A	U	F200	10_8_ Paranoid schizoph	Other Respiratory	0	0	0	0	0	0	1	0	0	0	0	0	0
1000772008	H1401485390	3	U61B	U	F200	10_8_ Paranoid schizoph	Diabetes	0	0	0	0	0	0	1	0	0	0	0	0	0

Appendix 1.2 Sample of raw clinical_codes data – Part 2

encounter	IschaemicScore	CerebrovascularScore	StrokeScore	COPDScore	AsthmaScore	RespiratoryScore	DiabetesScore	ArthritisScore	DevelopmentalDisScore	PeripheralVascScore	RenalFailureScore	SickleCellScore	AlcoholismScore
H1000325917	0	0	0	0	0	0	1	0	0	0	0	0	1
H1000533101	0	0	0	0	0	0	0	0	0	0	0	0	1
H1000793106	0	0	0	0	0	0	1	0	0	0	0	0	0
H1001181656	0	0	0	0	0	0	0	0	0	0	0	0	1
H1401399864	0	0	0	0	0	0	0	0	0	0	0	0	1
H1401463791	0	0	0	0	0	1	0	0	0	0	0	0	0
H1401485390	0	0	0	0	0	0	1	0	0	0	0	0	0
H1401531227	0	0	0	0	0	0	0	0	0	0	0	0	1

Appendix 2.1 Sample of processed clinical_codes data – Part 1

encounter	CCI_MI	CCI_CHF	CCI_PVD	CCI_Cerebrovascular_Dis	CCI_Dementia	CCI_Chronic_Pulmonary_Dis	CCI_Connective_Tissue_Dis	CCI_Peptic_Ulcer_Dis	CCI_Mild_Liver_Dis	CCI_Diabetes_wo_Chronic_Comp_Dis	CCI_Diabetes_w_Chronic_Comp_Dis	CCI_Paraplegia_Hemiplegia	CCI_Renal_Disease	CCI_Cancer	CCI_Mod_or_Severe_Liver_Dis	CCI_Metastatic_Carcinoma	CCI_HIV_AIDS
H1000325917	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
H1000533101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H1000793106	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
H1001181656	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H1401399864	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H1401463791	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H1401485390	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
H1401531227	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Appendix 2.2 Sample of processed clinical_codes data – Part 2