DATA VISUALIZATION AND PREDICTIVE MODELING FOR IDENTIFYING COMORBIDITIES IN DIABETIC PATIENTS

by

Giridhar Krishnan

Bachelor of Engineering (B.E.), MNM Jain Engineering College
Anna University, Chennai, 2009

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER SCIENCE

UNIVERSITY OF NORTHERN BRITISH COLUMBIA

November 2020

Abstract

Diabetes is one of the most common chronic diseases in the world. Diabetic patients are also more susceptible to develop additional comorbidities over time even causing death. This makes it essential to identify the risk of developing comorbidities as early as possible for effective diabetes management and to reduce the burden on healthcare system. Large volumes of clinical data which has been collected over the years has potential to be translated into meaningful information to enable healthcare professionals gain insights into diabetic patient comorbidities. This research has two key contributions. First, an interactive diabetes dashboard is developed in which the data is integrated and shown in the form of visually appealing charts, graphs and tables. The dashboard displays aggregated results with drilldown capabilities to allow navigation at finer granularities of various metrics. Second, predictive models are built to forecast the likelihood of one of the three common comorbidities for diabetic patients – Benign Hypertension, Congestive Heart Failure, and Acute Renal Failure. The models use advanced data mining algorithms such as Logistic Regression, Neural Network, CHAID, Bayesian Network, Random Forest and Ensemble. Results from these models are also incorporated into an interactive assessment tool that has the ability to take user input and predict the likelihood of one of these comorbidities. Northern Health (NH) dataset consisting exclusively of diabetic patients is used for this research.

Acknowledgement

There were a lot of people who have supported me in this journey at University of Northern British Columbia and I would like to express my appreciation and sincere gratitude to all of them.

Firstly, I would like to acknowledge the traditional and unceded territory of Lheidli Tenneh on which UNBC stands. I am thankful to have had the opportunity to pursue my research here.

I would like to express my sincere gratitude to my supervisor Dr. Waqar Haque as without his guidance, effort, motivation and unconditional support, this research would not have been possible. Also, I would like to thank my committee members Dr. Fan Jiang and Dr. Pranesh Kumar who have been very kind with their support and guidance for my research. I would also like to thank every faculty member who has helped me in my time as a graduate student.

Then I would like to thank my family and friends in Canada and India for supporting me through the time as a graduate student. I would like to give a special mention to every team member of BIRG lab I have worked with during this time and I am thankful for their motivation and support which meant a lot to me. Also, a special mention to all my family and friends in Prince George, Vancouver, Ontario and India who have all been very kind and supportive during my time as a graduate student.

Additionally, I would also like to thank Northern Health for providing me access to the data for my research.

Giridhar Krishnan

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Diabetes, or Diabetes Mellitus, is a chronic disease in which the body cannot either produce or utilize insulin. Insulin is a hormone which controls the amount of glucose (sugar) in blood. Elevated blood sugar levels may lead to damage of vital organs and can be fatal. There are three main types of Diabetes Mellitus [1]:

1. Type 1 diabetes (T1D) - occurs when body does not produce enough insulin (the cause is unknown)
2. Type 2 diabetes (T2D) - starts with insulin resistance and can progress to a lack of insulin (primary causes are lack of physical activity and obesity)
3. Gestational diabetes – occurs in pregnant women with no history of diabetes

According to the Public Health Agency of Canada (PHAC), 5 to 10% of diabetes patients have T1D and the remainder have T2D. Four percent of all pregnant women are affected by gestational diabetes which puts both the baby and mother at risk [2]. The cause for T1D and gestational diabetes has not yet been discovered by scientists. However, the list of risk factors for T2D are known to include [3]:

- Being overweight or obese
- Prediabetes (a condition that may occur before developing T2D)
- Advanced age
- Physical inactivity

- Having high blood pressure and/or high cholesterol
- Having a family history of diabetes
- Belonging to certain high-risk ethnic populations (e.g. Aboriginal, African, Hispanic, Asian)
- Having a history of gestational diabetes
- Having other conditions which may include vascular disease, polycystic ovary syndrome, and schizophrenia

In prediabetes, the blood sugar levels are higher than normal but lower than the threshold which defines T2D. Prediabetes and T2D can be prevented by maintaining a healthy lifestyle, eating a balanced diet, and ensuring regular physical activity [4]. Undiagnosed T2D in Canadian adults was found to be 1.13% contributing to 20% of total T2D patients [5]. Diabetes also leads to other comorbidities and puts a great burden on patients as well as the healthcare system. This disease can impact the entire human body from head to toe, causing blindness, stroke, heart attack, kidney failure and even non-traumatic amputations (Figure 1). Early detection of prediabetes can help prevent diabetes, and early diagnosis of T2D can help physicians recommend guidelines to ensure a healthy post-diabetes lifestyle to lessen the chances of developing related comorbidities.

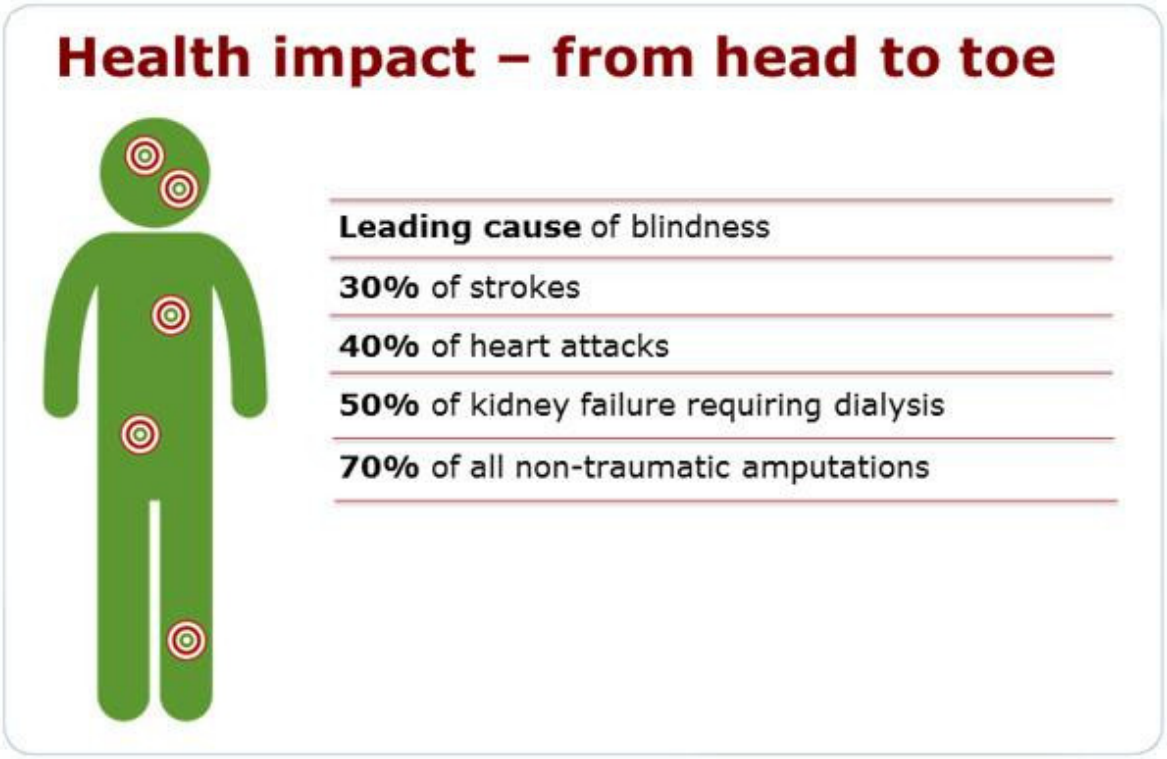


Figure 1 Impact of Diabetes on Human Body [2]

Diabetes is a disease that needs to be monitored constantly. Even the slightest changes in the health of diabetic patients can have adverse effects on their wellbeing and in some cases even lead to death. Diabetes is often considered a modern society disease which can lead to other complications listed earlier. The Canadian Community Health Survey (CCHS) has been collecting information related to health status, healthcare utilization and health determinants for the Canadian population (Figure 2). It produces an annual micro data file which can be used to extract information related to diabetes as well as other health related data [6].

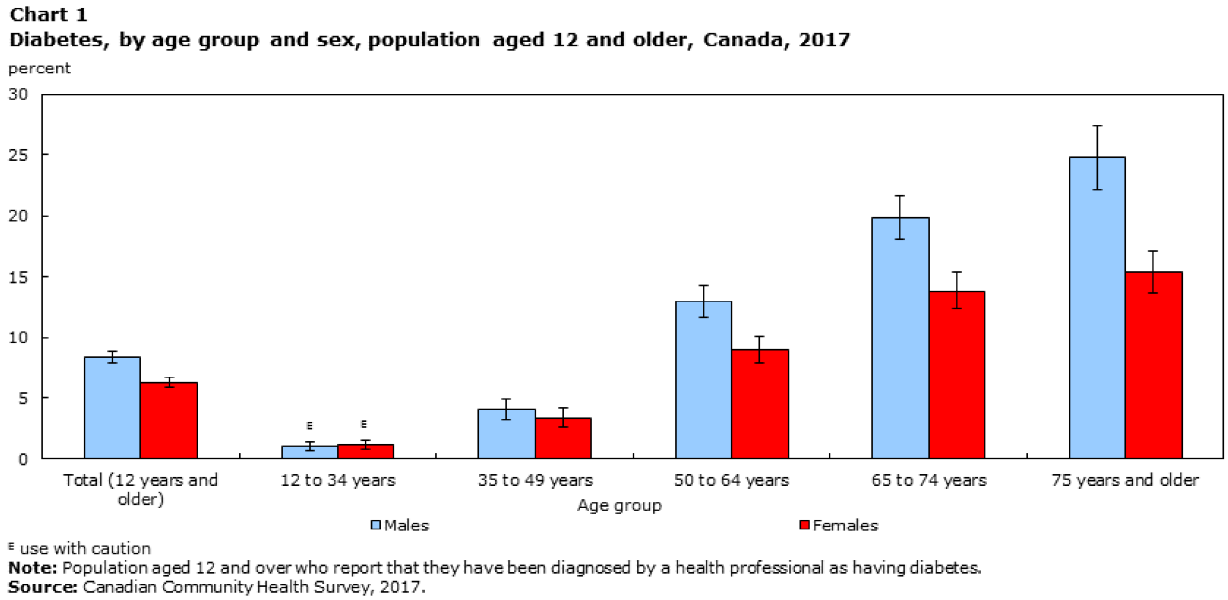


Figure 2 CCHS 2017 Diabetes Chart [2]

According to Diabetes Canada, 29% of Canadians are affected by diabetes. One million Canadians have diabetes but are yet to be diagnosed, and 3.9 million Canadians have been diagnosed with diabetes. Statistics for prediabetes are also a great concern with an alarming number of 5.7 million Canadians. Cumulatively, out of 37 million, more than 10 million people have diabetes or prediabetes (*Figure 3*). This number is expected to reach 33% by 2025 *[7]*.

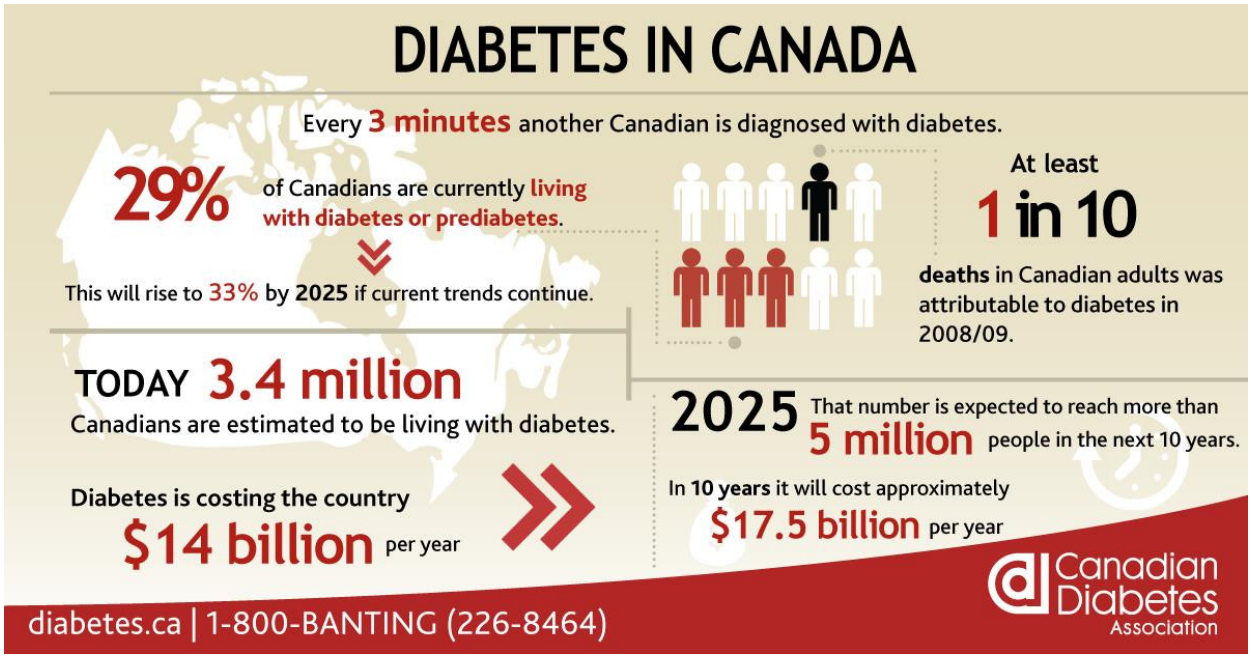


Figure 3 Canadian Diabetes Association Infographic [7]

While lack of physical activity, overweight and obesity makes one more vulnerable to diabetes, an additional observation was that clinically depressed people have 40%-60% increased risk of being diagnosed with T2D. The Diabetes Canada Backgrounder published in February 2020 has observed that 45.4% adults and 44.5% youth are physically inactive, 23.7% youth are either overweight or obese and 26.8% adults are living with obesity. The estimated mortality rate for Canadians with diabetes was twice in comparison with those without diabetes. Diabetes also has a significant cost impact with majority of patients in Canada paying more than 3% of their income for the treatments (Table 1). This cost is estimated to grow to $4.9 billion in 2030 [8].

Table 1 Estimated prevalence and cost of Diabetes [8]

Estimated Prevalence and Cost of Diabetes

Prevalence (1)	2020	2030
Diabetes (type 1 and type 2 diagnosed)	3,772,000 / 10%	4,891,000 / 12%
Diabetes (type 1)	5-10% of diabetes prevalence	
Diabetes (type 1 + type 2 diagnosed + type 2 undiagnosed) and prediabetes combined	11,232,000 / 29%	13,559,000 / 32%
Increase in diabetes (type 1 and type 2 diagnosed), 2020-2030	30%	
Direct cost to the health care system	$3.8 billion	$4.9 billion
Out-of-pocket cost per year (2)		
Type 1 diabetes on multiple daily insulin injections	$1,100-$2,600	
Type 1 diabetes on insulin pump therapy	$1,400-$4,900	
Type 2 diabetes on oral medication	$1,200-$1,900	

1.1 Knowledge Discovery in Databases (KDD) and Diabetes

Healthcare is one of the fields where foreseeing future outcomes and possibilities can be utilized effectively. Diabetes is one such disease, where early detection and management is vital to address the related health concerns. Foreseeing the possibility of a patient having diabetes and related comorbidities would be highly beneficial and this can be accomplished using predictive modeling which analyzes patterns and correlations in historical data. The entire process, methods, theories and techniques involved to make sense of available data is called Knowledge Discovery in Databases (KDD). Figure 4 illustrates the basic steps involved in KDD [9].

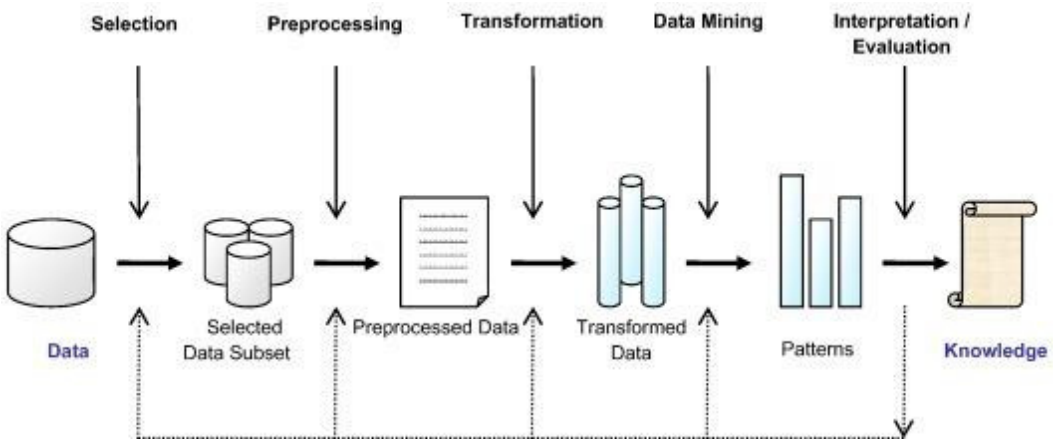


Figure 4 KDD Steps [9]

Data mining is a vital step of the knowledge discovery process and involves cleansing, integrating, mining, selecting, modeling, pattern recognition and knowledge representation of massive amounts of data (Figure 5). This process discovers unknown patterns that provide useful results and plays a valuable role in healthcare research to improve quality of life of patients diagnosed with health conditions [10]. Data mining can also be interfaced with statistics, machine learning, neural networks and inductive logic programming to play an important and decisive role in diabetes research [11]. Machine Learning is the process in which machines learn and adapt from experience by repeating a task for *n* number of times which in turn improves the performance. It is imperative to note that machine learning and data mining are two terms that are closely related with the latter being more generic. Thus, in literature, machine learning methods are also sometimes referred to as data-mining methods [9].

In healthcare, interfacing data mining with data warehousing and using Online Analytical Processing (OLAP) can enable efficient decision making. Data warehouses contain consolidated data which facilitates complex analyses and visualization through OLAP

[12]. OLAP has the capability to perform various operations such as rollup, drill-down, slice and dice, and pivot on the data warehouse. Rollup and drilldown increases and decreases the level of aggregation, respectively; slice and dice is used to select specific dimensions, and pivot re-orients the multidimensional view of the data warehouse.

Machine learning and data mining can be utilized to extract knowledge from huge volumes of diabetes–related data. Data mining algorithms are used to identify correlations between different variables in the data source and build predictive models. These models have insightful information of diabetic patients, comorbidities and other demographics for the purposes of clinical administration, diagnosis as well as management of diabetes.

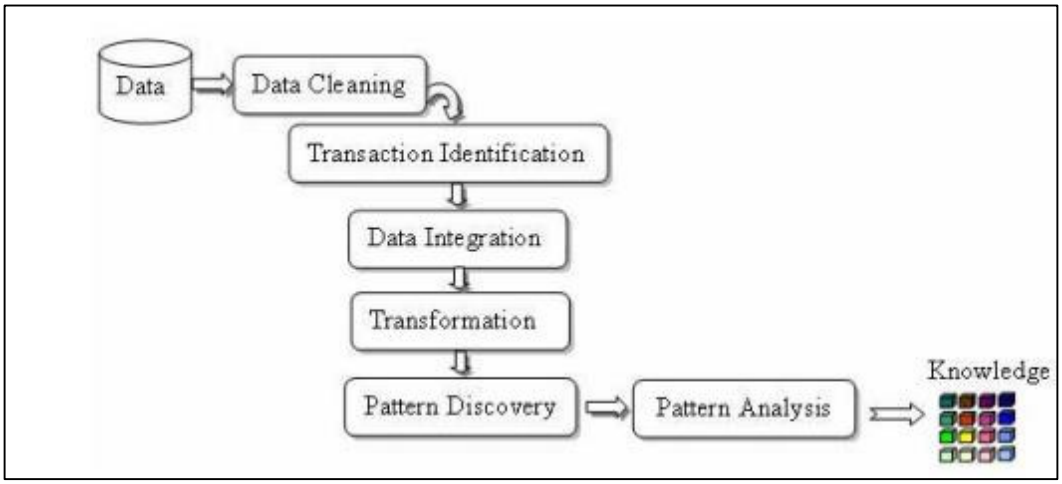


Figure 5 Data Mining Process [11]

1.2 Data Visualization

Interpreting data can be a complicated and tedious task. Complex statistics and equations may not be understood by all, but when visualized it can be made more relevant to the end user. Interactive data visualizations can help users to quickly identify patterns and trends which can enable effective decisions. An effective way to represent data is through dashboards which can translate key performance of organizations into visual displays. Dashboards allow visualization of huge amounts of data in an intuitive manner using charts, graphs, gauges and more. Interactive dashboards with color-coded visualizations are more appealing to the end user and enhances their experience.

In healthcare, time is vital. Professionals have to make decisions rapidly to ensure optimal care for the well-being of patients and manage resources efficiently. Research shows that dashboards significantly reduce time when compared with the conventional approach of using electronic health records (EHR) for analysis and management [13]. For instance, a study compared the time for ten physicians to access ten common variables for two diabetic patients with similar volumes of clinical data using conventional EHR and a diabetes dashboard [13]. The mean time taken to access the ten variables for two diabetic patients was 1.9 minutes using the dashboard and 6.3 minutes with the conventional approach, showing that dashboards can significantly help reduce time spent by physicians and help optimize patient management. The research further established that usability analysis tools like dashboards can be an insightful asset for health care information technology [13].

In 2017, an electronic diabetes dashboard, iScreen, designed by the Canadian Diabetes Association was introduced [14]. For this research, T1D patients between 14-18 years

were assessed for other comorbidities. Fifty charts were used for review, 25 using iScreen and 25 without iScreen. The results showed an increase in appropriate initial screening and decrease in under- as well as over-screening of patients for nephropathy and retinopathy after using iScreen electronic diabetes dashboard (Figure 6). This study concluded that dashboards have potential to impact clinical outcomes and healthcare costs.

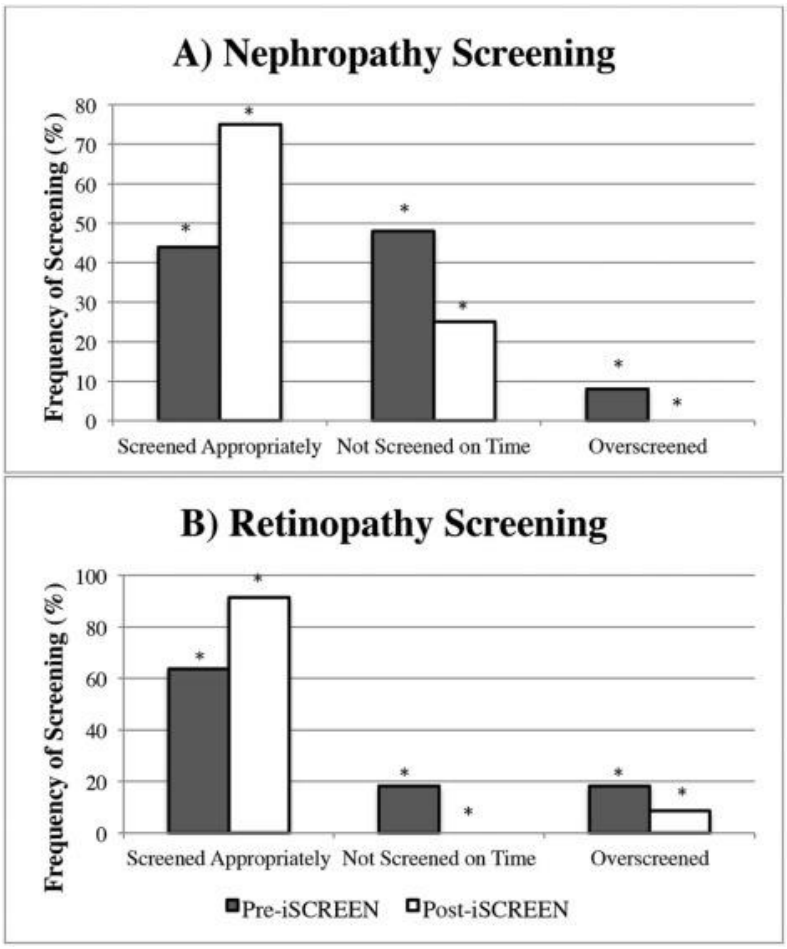


Figure 6 Screening of T1D patients [14]

In sum, an interactive diabetes dashboard listing the risk factors and comorbidities alongside different demographics has a potential to save time as well as enable effective decision making to optimize clinical outcomes and costs of treatment.

1.3 Current State & Motivation

Diabetes is one of the major chronic diseases worldwide and efforts are being made on a global scale to deal with it in the most efficient manner. Since early detection of diabetes is one of the major factors to prevent it, there has been considerable research done on developing diabetes risk calculators. For instance, in 2009, a simple tool [15] for detecting undiagnosed diabetes and prediabetes was proposed and data was used from the Third National Health and Nutrition Examination Survey [NHANES]. The models were built using two methods – classification tree analysis and logistic regression. In 2011, the Public Health Agency of Canada (PHAC) developed a non-laboratory based screening questionnaire to identify diabetes and prediabetes among middle-aged adults. The questionnaire was built on the basis of Finnish Diabetes Risk Score (FINDRISC) which led to the development of The Canadian Diabetes Risk Questionnaire (CANRISK) [15] which is a diabetes screening tool for Canadians aged over 40. There has also been significant research done on using data mining algorithms in this area [16] [17]. In 2013, three data mining models, namely, artificial neural networks (ANN), C5.0 decision tree and logistic regression were compared to predict diabetes and prediabetes [16]. The C5.0 decision tree model demonstrated the highest accuracy for the dataset used in this research; this dataset was based on information collected from a questionnaire [16].

Machine learning algorithms for detecting undiagnosed diabetic patients have also been compared and used to create best models using ANN and logistic regression [9]. A more comprehensive literature review is provided in Chapter 2.

Existing studies have focused more on building diabetes calculators (FINDRISC, CANRISK) but there is a lack of research when it comes to tools developed for identifying diabetes comorbidities [15] [18] [19]. Another limitation prevalent in existing work [16] [17] is that the risk factors identified for predictive modeling in diabetes uses survey data and interviews. While such databases can be useful to get an overall picture of the disease, the authenticity of the underlying data is highly questionable. Specifically, self-reported data has a high possibility of containing unreliable information [20]. Predictive modeling uses historical data as its base to build models and if this data is inaccurate, the model's accuracy becomes questionable. One way to eliminate this issue is to use clinical data recorded by healthcare professionals. Clinical data is authentic as patient diagnosis has been confirmed by qualified physicians. However, obtaining clinical data for research can be quite challenging as it is seldom available in the public domain due to privacy concerns.

Considering these factors, this research proposes building predictive models for diagnosed diabetic patients using a clinical dataset obtained from Northern Health. These models will enable users to identify hidden patterns in historical data and predict the likelihood of comorbidities resulting in effective diabetes management. The models have been integrated with an interactive diabetes dashboard for visual analytics.

1.4 Problem Statement

Predictive models that accurately forecast the likelihood of various diabetes comorbidities could be an efficient tool for healthcare providers as well as patients. This could facilitate early diagnosis and interventions with the possibility to prevent other comorbidities as well as effective management of diabetic patients reducing the cost quotient on the healthcare system. The results from these predictive models should be user friendly and beneficial for end users; this leads to the first research question.

1.4.1 How to enhance diabetes management using intuitive visualization techniques?

Chronic conditions such as diabetes require frequent follow-ups and monitoring of patients for effective management. This can be a tedious task considering the large number of patients to track, and the required tests for related comorbidities. Summarized information of individual patients can enable healthcare professionals to take useful and timely decisions. Aggregated data visualization of multiple patient records can give an overview of the entire dataset while drill-downs can let users navigate to finer granularities focusing on individual patients. This can be accomplished with an interactive diabetes dashboard which identifies patients with their associated clinical visits and treatment plans.

A previous study described earlier has shown that conventional approach using EHR took 6.3 minutes to identify all the associated variables for diabetic patients compared to 1.9

minutes using a diabetes dashboard. The mean number of mouse clicks were 60 for EHR and significantly reduced to 3 using the dashboard [13].

Taking these factors into account, an interactive diabetes dashboard is built with the following features:

- Color coded visualization of existing and predicted data in the form of charts and graphs with available demographics (i.e. patient Local Health Area (LHA), patient community, age and comorbidities)
- Results of models represented in comparative charts
- Aggregated data with drill down capability to view information at finer granularity

This interactive diabetes dashboard will help decision makers to identify patterns and understand the relationship of different variables specific to comorbidities and patients which is complicated and time consuming using EHRs. These relationships also help identify associated risk factors using historical and predicted data. This leads to the next research question:

1.4.2 What are the vital risk factors for diabetes comorbidities?

Diabetes is associated with a number of comorbidities affecting the entire human body. Analyzing and applying data mining algorithms to existing clinical patient data can help identify the risk factors specific to comorbidities. This research focuses on three common comorbidities which are acknowledged by Diabetes Canada - hypertension, congestive heart failure and renal failure [3]. The risk factors associated with each of these comorbidities and the prominent common risk factors are presented using a diabetes

dashboard. Another interesting observation would be to associate the comorbidities themselves with the help of a predictive model; this leads to the final research question:

1.4.3 What is the likelihood of a patient to be diagnosed with other comorbidities?

Diabetes Canada has observed that diabetic patients lifespan can reduce by five to 15 years; further, diabetes has been attributed as the reason for death of one in every ten Canadian adults in 2008-2009 [8]. These patients are also more likely to be hospitalized with cardiovascular disease and twelve times more susceptible to end-stage renal disease compared to the general population. These observations emphasize the importance of being able to predict the likelihood of comorbidities so that early detection and efficient diabetes management can be achieved. Taking these factors into account, predictive models to find the likelihood of diabetic patients with the following three comorbidities are proposed:

- Hypertension
- Congestive Heart Failure
- Acute Renal Failure

The results from these models can act as a useful guideline to identify patients vulnerable to specific comorbidities and recommend appropriate management to prevent escalation to further comorbidities. Effective diabetes management would ensure optimal patient care as well as reduced costs on the healthcare system.

1.4.4 Methodology

This research focuses on building predictive models using existing diabetes related clinical data. The model predicts the probability of diabetic patients to develop related comorbidities. To make the results of the model easily accessible, a simple user-friendly assessment tool is developed which predicts the probability of the three comorbidities to the users. In addition, an interactive dashboard is designed to visualize insightful information representing several years of diabetes data including the identified risk factors related to various comorbidities.

The raw clinical data is integrated into a database using SQL Server Management Studio (SSMS) 15.0 [21]. The integrated database is used to build the predictive models with IBM SPSS modeler 18.2 [22] which predicts the likelihood of the three comorbidities (congestive heart failure, hypertension and renal failure). The models help to identify the prominent risk factors for these comorbidities as well as the significance of variables for the predictions. The results produced by each model and the identified risk factors are analyzed in detail.

An interactive diabetes dashboard is built using SQL Server Reporting Services 15.0 (SSRS) which is a component of Microsoft Business Intelligence tool stack [23]. SSRS is an effective reporting platform which includes various data visualization tools such as charts, graphs, and gauges to represent data; capability to integrate maps and embed images is also included. The aggregated results are presented in a visually pleasing format with the option of drilling down to reports at finer granularities.

Finally, a user friendly tool is developed using IBM SPSS modeler. This tool allows diabetic patients and healthcare professionals to view results generated by the models predicting likelihood of one of the three comorbidities.

The study methodology is described in detail in Chapter 3.

1.5 Contributions

This research has two major contributions. Firstly, predictive models for Canadian diabetic patients which forecasts the likelihood of related comorbidities have been developed. These models could be used by healthcare professionals as a guideline to identify patients who are at higher risk of developing predicted comorbidities and ensure effective management of diabetes. Clinical data of diabetic patients who accessed Northern Health (NH) facilities between April 1 2012 and March 31 2018 has been used to train and test the models for accuracy.

The second contribution is the design and development of an interactive diabetes dashboard. Data visualization in the form of charts and graphs can enable healthcare professionals to have a better and deeper understanding of the variables associated with the disease. The dashboard provides individual as well as aggregated data at facility and community levels with drill down and drill through reporting. This information can be useful to identify the gaps in healthcare and enhance related services by making informed decisions in a timely manner. These contributions are described in detail below:

Identifying diabetic-patient comorbidities: Predictive models built specific to diabetic patients predicting the likelihood of comorbidities - benign hypertension, congestive heart

failure and acute renal failure. These results can be used as a guideline by healthcare professionals to identify and treat patients who are at a higher risk for developing these comorbidities.

Enhanced patient-care: Early detection of patients who are at high risk for developing one or more comorbidities could help prevent further complications to their health and timely treatment ensuring overall well-being of patients.

Reduce healthcare costs: Identified high risk patients who receive timely care are less susceptible to other complications; this in turn benefits the patients as well as the healthcare system with elimination of complex treatments reducing the burden of cost.

Holistic healthcare approach: Interactive diabetes dashboard provides an overview of the current state of diabetes and diabetic patients in the selected dataset. The embedded drill downs allow filtering of results by various demographics such as Health Service Delivery Area (HSDA), LHA and comorbidities. The historical information is presented in a way which facilitates analysis and assists decision makers in identifying gaps in provision of healthcare. The stakeholders can use this information to develop plans for improving related services.

Diabetes comorbidity prediction tool: The results from the model are incorporated into a user-friendly web form which predicts the possibility of one of the three comorbidities for diabetic patients. The interactive web form asks for a user to enter input for the selected variables and predicts the value of the target variable using the underlying models. This tool could be used by the healthcare professionals to enhance treatment and management of diabetes.

Chapter 2

Related Work

Diabetes being a worldwide chronic illness has no shortage of research especially with focus on early diagnosis and detection of the disease. One reason for such extensive research in this direction is the cost and toll of diabetes management on healthcare systems. The research community has explored various data mining techniques to detect and diagnose diabetes. The research has encompassed not only diabetes, but all associated comorbidities including hypertension, renal failure and cardiovascular diseases. Diabetes research thus branches into various fields including but not limited to healthcare, data mining and data visualization. The literature review presented in current chapter was done to align with the focus of this research, that is, building predictive models for diabetes comorbidities and designing an interactive diabetes dashboard.

This chapter is divided in three main sections. Firstly, representative studies on diabetes and data mining are presented. These include analysis of various data mining algorithms and techniques used in the study of diabetes and related comorbidities. A review of the application of research work to develop user-friendly tools such as diabetes calculators

is then presented. Finally, the use of data visualization for enhanced and cost effective healthcare is explored. Limitations of existing literature are provided to identify research gaps that are addressed by the work presented in this thesis.

2.1 Diabetes and Data mining

Data mining plays a huge role in the healthcare sector with algorithms that have the ability to analyze, detect and predict the presence of diseases in patients. Early detection of diseases can help in timely and efficient decision making by healthcare professionals. Diabetes is one such disease where data mining can be a vital part of developing tools that can facilitate enhanced healthcare service. In this section, research with respect to data mining and diabetes is explored and techniques for implementing predictive modeling specific to diabetes are analyzed.

In 2012, a study was done to predict T2D using data mining. The aim of the research was to apply artificial metaplasticity on multilayer perceptron (AMMLP) as a data mining (DM) technique for diabetes and compare results with the decision tree, Bayesian classifier and other algorithms [24] .The comparisons were done using classification accuracy, analysis of sensitivity and specificity and confusion matrix. The results showed an accuracy of 89.5% for AMMLP which was superior to decision tree and Bayesian classifier algorithms. The dataset used for this research was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset comprised of a specific group of Pima Indian women tested for diabetes. The sample used eight variables:

- number of times pregnant
- plasma glucose concentration
- glucose tolerance test
- diastolic blood pressure
- triceps skin fold thickness
- serum insulin
- body mass index
- diabetes onset within five years

This research [24] had a sample size of 768 which was further reduced to 763 after elimination of records with missing data. It is also to be noted that out of 768 patients only 268 had diabetes. T2D is common among both men and women but this dataset did not include men and it was also focused on a specific group making the relevancy of results for other groups questionable. It was concluded that AMMLP performed with better accuracy but it was compared only with two other classifier algorithms.

In 2013, a research project proposed automated detection of diabetes mellitus using neural networks without patients undergoing clinical tests [17]. The neural network had a total of 27 nodes (13 input, 13 hidden and 1 output) [17]. The input nodes are the variables with shared historical data, the hidden nodes are where the computation occurs and the output node is the result for the given input. The neural network was built using the backpropagation[1] algorithm and out of 20 datasets tested, 18 produced accurate

[1] "backward propagation of errors" calculating gradient of error function

results with an overall accuracy of 92.8%. For this research, a survey was done using 100 datasets. Each data set included people of various ages, genders and lifestyles to get an unbiased result. Eighty datasets were used for training and twenty were used for testing the system. The following variables were used for building the model:

- Age
- Gender
- Weight
- Height
- Weight loss
- Thirst increase
- Hunger increase
- Appetite increase
- Nausea
- Fatigue
- Vomiting
- Bladder, skin infections

Considering that the data was collected using surveys and was self-reported, the authenticity of the diagnostics becomes questionable. This research [17] also mentions another study using ANN and feature extraction which achieved an accuracy of 94.6% for classifying patients as diabetic and non-diabetic.

Another study [16] was conducted to compare three data mining models (ANN, decision tree and logistic regression) to predict diabetes or prediabetes by various risk factors [16]. A questionnaire to obtain information on demographics, family diabetes history, anthropometric measurements and lifestyle risk factors was given to 1,457 participants, 735 of whom had diabetes. The following twelve input variables were used:

- Age
- Family history of diabetes
- Marital status
- Education level
- Work stress
- Duration of sleep
- Physical activity
- Preference for salty food
- Gender
- Eating fish
- Drinking coffee
- Body mass index

The output variable was a flag variable with possible values of 0 and 1, where 1 indicated if the person had diabetes or prediabetes. Results from the three predictive models are shown in Table 2.

Table 2 Comparison of Data Mining Models

	Logistic regression	ANN	Decision tree
Sensitivity	79.40%	79.40%	78.11
Specificity	73.54%	65.47%	75.78%
Accuracy	76.54%	72.59%	76.97%

In conclusion, the decision tree had the highest classification accuracy, followed by logistic regression and ANN. Classification accuracy is the percentage of correct predictions in a model and is considered to be a vital performance indicator. A limitation of this study [16] was that the sample population chosen was only from two communities in Guangzhou, China and cannot be considered an appropriate representation of the entire population. Also, some of the individuals who participated in the study provided self-reported data which make the results less reliable.

In a recent study published in 2020, data was collected from over 230,000 participants during the years 2006-2017 to develop a T2D risk prediction model using machine learning algorithms [25]. This research excluded all diabetic participants as well as any participants taking medication for diabetes. The collected medical, behavioral, demographic and incidence data was used to predict T2D in participants at 3, 5, 7 and 10 years. The participants selected in the research were followed up for the entire time period thus making it a longitudinal dataset. Three machine learning algorithms, random forest, multilayer feedforward artificial neural network implementing a deep-learning approach, and a gradient boosting machine approach, were compared with conventional logistic regression model. The AUC (Area under Curve) in machine learning models was higher

than the conventional regression model. AUC is a statistical performance measurement which is used to validate the model. A higher AUC implies better prediction capabilities of the model. The highest accuracy was recorded by gradient boosting algorithm with an AUC of 79% in 3-year prediction and 75% in 10-year prediction. The machine learning models also predicted BMI as the vital risk factor contributing to T2D. It was also noted that diabetes incidence was recorded higher among men than women over the ten-year period. Limitations of this research were that it used self-reported data and the exclusion of participants was done by use of diabetes related medication instead of a clinical diagnosis.

The studies presented above focused on predicting diabetes [16] [17] [25]. The next few research works [24] [26] explore the use of data mining techniques to improve management and treatment plans for diagnosed diabetic patients. Management of diabetes is a critical challenge for healthcare professionals as well as for the patients themselves. Diabetic patients have higher risk of being diagnosed with multiple comorbidities which, in turn, increases the complexity of treatment and care. Hence, it would be ideal to predict comorbidities using data mining techniques. Existing literature shows a few studies that have focused on this topic.

A comorbidity study [24] done on diabetic patients identified that hypertension plays a critical role in its association with other comorbidities. Hypertension was identified as a critical factor for T2D patients having stroke as well as dyslipidemia [27]. For this research, 20,314 patients with T2D were chosen from Keimyung University Dongsan Medical Center. Apriori algorithm was used to find the association between T2D and various

comorbidities. Hypertension had the highest association followed by gastritis and senile cataract. Apriori algorithm was implemented through a proprietary tool, Dx Analyze, which aided in the process of data cleansing and construction of data marts as well. A limitation of the study was that the data represented only one medical facility and the Dx Analyze tool needs to be applied on data from multiple facilities to check for relevancy of the results. The authors also acknowledge the limitations of Apriori to determine causality of disease and recommend further research considering chronology of diseases in patients.

In another interesting research, mortality of diabetic patients in ICU was predicted [26]. The MIMIC-III database which records ICU admissions was used for this study. There were a total of 10,318 diabetic patients in this database; this number came down to 4,111 after exclusion of missing values for blood glucose. Existing algorithms to predict mortality were used for the models - Charlson Comorbidity Index (CCI), Elixhauser Comorbidity Index and Diabetes Complications Severity Index (DCSI). CCI and Elixhauser calculate risk-scores based on ICD-9 diagnosis codes for each patient while the DCSI is an alternative risk score designed specific to diabetic patients. The results showed AUC values to be 0.694, 0.682 and 0.656 for DCSI, Elixhauser and CCI, respectively. The AUC improved to 0.785 when all three metrics were combined using logistic regression. In addition, the random forest model achieved an accuracy of 0.787. This research focused on five variables:

- Age
- Gender

- Ethnicity
- Insurance
- Admission Type

A limitation of this research was that it used random sampling of 70/30 for analysis which resulted in an imbalance of less than 10% of positive cases. Also, it did not consider patients directly admitted for diabetes related care because it was complicated to identify with the different diagnostic codes recorded for a patient. Length of stay is a variable which was not analyzed and is recommended to be explored by the authors. This research also recommends exploring other machine learning algorithms such as random forest and ANN for better predictions.

Data mining algorithms can be effectively used and adopted in healthcare to build predictive models with patient-specific information to predict diseases such as diabetes. Predictive models for T2D comorbidities could contribute to associating the relation between risk factors and identify onset of specific comorbidities [28]. The models can also be used to develop tools to aid in informed decision making for optimized treatment of diabetic patients. User-friendly electronic tools to identify patients with diabetes can be highly beneficial for efficient treatment and management of diabetes. The next section covers literature focusing on existing calculators for diabetes.

2.2 Diabetes Calculator

Early identification of diabetes is ideal for well-being and treatment of patients and diabetes calculators are an effective tool to accomplish this. These calculators can act as a guideline for patients to analyze their risk of being diagnosed with diabetes; higher the potential risk, more advisable and essential to contact a physician. Over the years, there has been a lot of research done globally on diabetes calculators, some of which is presented in this section.

In 2003, a tool to predict T2D (Diabetes Risk Score) was developed to identify individuals at risk without undergoing laboratory tests [18]. The risk factors taken into account were:

- Age
- Body Mass Index (BMI)
- Waist circumference
- History of antihypertensive drug treatment and high blood glucose
- Physical activity
- Daily consumption of fruits, berries, or vegetables

For this study [18], a random population sample between ages 35-64 was selected and followed for 10 years. Each category was assigned a score using multivariate logistic regression model coefficients. The cumulative sum of all scores was calculated as the Diabetes Risk Score (DRS). The research identified 182 cases of diabetes incidence in 4,435 subjects. DRS has been implemented in Finland as one of the tools in their diabetes prevention program. The SAS (version 8.2; SAS Institute, Cary, NC) software was used

for analysis. This research has several limitations. First, the risk factors do not include family history of diabetes which is an important factor contributing to an increased risk of acquiring the disease [29]. The researchers recommend addition of this factor in future work. The individuals with high glucose levels were not excluded at the baseline under the assumption that no biochemical tests were performed at that stage. In addition, the data used to build the model was obtained from surveys and the national population register.

In 2005, Indian Diabetes Risk Score (IDRS) was proposed for screening undiagnosed diabetic patients [19]. Indian Diabetes Risk Score used four risk factors: age, abdominal obesity, family history of diabetes and physical activity. Multiple logistic regression analysis was applied using undiagnosed diabetes as the dependent variable. When risk score was greater than or equal to 60, the IDRS had an accuracy of 61.3% with a positive predictive value of 17.0% and a negative predictive value of 95.1%. Receiver Operating Characteristic (ROC) curves showed that area under ROC curve was 0.698 with a confidence interval of 95%. Indian Diabetes Risk Score, which categorizes risk factors based on their severity, can be a cost effective tool for mass screening in developing countries like India where a large number of cases are undiagnosed. The risk score for this research was derived from Chennai Urban Rural Epidemiology Study (CURES). The response rate for this study was 90.4% and the results were subject to internal validation. The sample size for this research was 2,350 patients. This research [19] did not take dietary consumption into account which is one of the recommended risk factors by the American Diabetes Association. In addition, anti-hypertensive medication was

excluded as one of the variables considering that a lot of people do not take medication. The other shortcoming of this research is that it is a cross-sectional study and the authors recommend validating this study with prospective studies. A cross-sectional study collects data from various sects of the population at a given time opposed to collecting data over time. For medical research involving predictions, prospective studies are preferred as they are longitudinal and the results obtained can have a better relevance.

A simple tool for detecting undiagnosed diabetes and prediabetes was proposed using data from the Third National Health and Nutrition Examination Survey [NHANES] [30]. The models were built using two methods – classification tree analysis and logistic regression. The diabetic risk calculator tool used the following risk factors:

- Age
- Waist circumference
- Gestational diabetes
- Height
- Race/Ethnicity
- Hypertension
- Family History
- Exercise

The classification tree model was used based on its ease of use and the results obtained are shown in Table 3:

Table 3 Diabetes Risk Calculator Results for the United States [30]

	Undiagnosed Diabetes	Prediabetes
Sensitivity	88%	75%
Specificity	75%	65%
Positive Predictive value	14%	49%
Negative Predictive value	99.3%	85%
ROC	85%	75%

ROC area under the curve for undiagnosed diabetes was 0.85 and for prediabetes was 0.75. ROC is used to evaluate the performance of models where the true positive rate is represented by sensitivity and false positive is represented by specificity. With ROC analysis, optimal models for predictions can be evaluated. This research [30] eliminated the variables for body mass index (BMI) in favour of height and weight, and the cholesterol variables were eliminated due to missing fields and low predictor value. Another important variable eliminated was diabetes in any blood relative. There were 18 variables chosen but not all of them were used in the final model. Finally, the tool is yet to be developed into a patient friendly electronic version for broader use.

In 2011, the Public Health Agency of Canada (PHAC) came up with a strategy for preventive intervention [15]. Before such an intervention can be applied in Canada, it is important to have an early detection strategy to be successfully implemented. The PHAC developed a non-laboratory-based screening questionnaire to identify diabetes and prediabetes among middle-aged adults. The questionnaire was built on the basis of Finnish Diabetes Risk Score (FINDRISC) which led to the development of The Canadian Diabetes Risk Questionnaire (CANRISK) [15].

CANRISK asks 13 questions that categorizes people as low risk, moderate risk and high risk. The low risk has a score of less than 21, and the high risk has a maximum score of 86. The moderate risk scores can vary from 21 to 32. The 13 questions focus on the various risk factors such as age, gender, height & weight (to calculate BMI), blood pressure and blood pressure during pregnancy (gestational diabetes). CANRISK also has questions related to family history of diabetes as well as ethnicity and education. Each of these variables contribute to the total diabetes risk score. In case of moderate risk, CANRISK recommends consulting a healthcare practitioner whereas in the case of high risk blood sugar test is recommended. CANRISK has been implemented and translated into different languages [15]. Limitations of this work are that some ethnic groups are under-represented in the sample and CANRISK is yet to be evaluated as a screening tool for high risk patients.

2.3 Data Visualization and Diabetes

Healthcare is one of the areas where abundant data is stored in various disparate formats. Integrating and organizing such data is an ongoing challenge faced by healthcare providers. Critical data can be challenging to be retrieved from electronic health records. Data visualization can help solve this challenge and lead to enhanced patient care and optimized diabetes management. Research has shown that management of diabetes improves when patients are provided with information and knowledge about their health condition.

In a study, patients were assessed by a diabetologist and given access to a web portal which had information regarding diabetes, their personal health status as well as the ability to contact the diabetologist [31] . The primary goal of this research was to monitor the blood glucose levels (A1C) and to observe differences between users who had access to the web portal and those who did not. This study observed that the web portal users had lower levels of A1C compared to the non-users. Further, it confirmed the usefulness of a web based tool to enhance patient management and cut costs in the long term. This research used only 8% of the original patients (157/1957) for the final analysis as only 157 patients had covariate data and did a follow-up visit. This study also did not explore the demographic factors that would influence the usage of the web portal, and did not distinguish between patients with T1D and T2D.

There has been work done towards building dashboard for diabetes. As mentioned in Chapter 1, the iScreen electronic diabetes dashboard [14] observed that evaluation of decision support tools facilitate complicated screening for diabetes care. However, iScreen included only T1D patients and had a small sample size of only fifty patients.

Mosaic is a project funded by the European Union (EU), specifically to explore predictive models and decision support system for T2D care and management; a clinical decision support system (CDSS) dashboard was built for this project [32]. The dashboard explored diabetic patient data and risk of complications; it consisted of three sections consolidating metabolic control, frequent temporal patterns and drug purchase patterns. An outcome assessment and research support system (ORSS) was designed for clinicians

[32]. Figure 7 shows an example of CDSS where patients are grouped by complication categories and details.

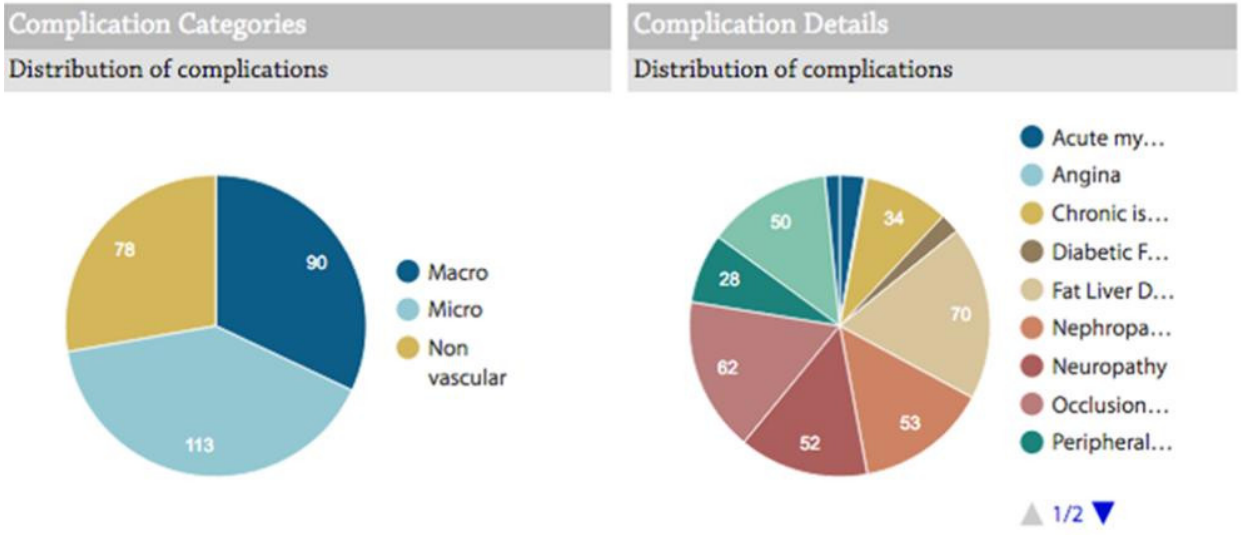


Figure 7 CDSS Diabetic Patients Complications

Upon evaluating the CDSS for nine clinicians, it was observed that T2D patients who had access to CDSS recorded shorter durations with their clinical visits and screening for complications increased in the visits indicating optimized patient care. The researchers observed that the dashboard can be improved by implementing a more detailed human-computer interaction study. The dashboard was evaluated for patient management but not for any clinical outcomes. There were a limited number of clinicians involved and they all were from the same facility.

Nevertheless, chronic diseases such as diabetes can be aided with the help of data visualization tools such as dashboards to support clinical decision-making, including diagnosis, treatment plans, and effective management if used in a coordinated fashion to improve the overall well-being of patients complementing the healthcare system.

2.4 Summary

Research in data mining and diabetes has identified risk factors which have a strong relationship to diabetes, such as age, BMI and dietary consumption. Predictive models play an important role in forecasting future health outcome of patients. Unfortunately, majority of existing research work in prediction focuses on comparing different data mining algorithms to determine the most efficient algorithm to build the model. Undoubtedly, there is a lack of research to identify risk factors leading to comorbidities such as cardiovascular disease, renal failure, and hypertension for diagnosed diabetic patients. Several tools have been developed for early diagnosis and management of the disease. One of the major issues with the current risk calculators is that the majority of them are paper-based questionnaires as opposed to online tools [33]. The identified research gaps discovered in current literature are summarized below:

Self-reported data: Majority of the published work is based on survey data and questionnaires which makes the data quality highly questionable. This, in turn, impacts the reliability of the predictive models which are built on top of this data.

Low Count of Diabetic Patients: Many datasets had a low count of diabetic patients, and there is a lack of research using datasets which exclusively represent diagnosed diabetic patients.

Domain-Specific Datasets: The datasets used in majority of the researches were specific to an ethnic group or to a particular facility which makes results applicable and relevant only to the group associated with the dataset.

Location Demographics: The datasets used lack information on demographics such as the community and facilities accessed by the patients. Location demographics can contain useful information specific to a community or facility which can represent insightful information. Also, there is limited research for predictive models based on data for diabetes patients in Canada.

Adding New Input Variables:

Majority of the studies have included age, gender, ethnicity and BMI as input variables for building the models. Variables such as length of stay, discharge date, and availability of family physician have not been explored to evaluate their impact on the models.

Diabetes Comorbidity Assessment Tool:

Several calculators for diagnosing diabetes have been developed over the years, including CANRISK, but tools available for identifying comorbidity or multi-morbidities in diabetic patients are scarce.

To overcome these limitations, predictive models for diabetes comorbidities have been built using NH clinical data. Clinical data eliminates the issue of self-reporting as only diagnosed patients are part of the dataset. This dataset has exclusive information of diabetic patients who are diagnosed with either T1D or T2D, also this dataset is specific to Canada and has information of all patients who have accessed NH facilities from 2012 to 2018. This includes patients from different communities accessing various facilities which makes it a generic dataset rather than specific to a domain. The models included the lesser explored variables such as length of stay, access to family physicians and

facilities. The importance of each of these variables with respect to the different comorbidities was analyzed for building the predictive models. The results of the diabetes comorbidity models were integrated with a user-friendly tool to predict the risk of hypertension, cardiovascular disease and renal failure in diabetic patients. In addition to this, a dashboard has been developed for visualization of existing clinical diabetes data to give the users insightful and useful information about diabetes. This enables users to interact and analyze anonymized patient data for effective decision making and improved healthcare outcomes. The existing research has served as a guideline to choose relevant variables and algorithms for developing the models.

Chapter 3

Methodology

This research has three interrelated components. First, a model for predicting diabetes comorbidities is proposed. Second, an interactive dashboard has been developed to provide insights about diabetes using visual analytics. In the process, hidden data patterns are uncovered and the newly discovered knowledge is imparted via this interface. Finally, a user-friendly assessment tool allows users to benefit from the model results for their specific cases. These components are explained in detail in this chapter.

3.1 Proposed Model

The key components of the model are shown in Figure 8.This model was studied for three representative comorbidities using several data mining algorithms and the NH clinical dataset of diagnosed diabetic patients. IBM SPSS modeler was used to identify and apply the most efficient data-mining algorithms for prediction of these comorbidities. Relationships between different comorbidities and demographics were also identified. The visual analytics dashboard was built using SSRS as the underlying platform. The front end for the assessment tool consists of a simple web form wherein the user enters information such as age, diagnosis code and health service delivery area. This information is processed based on a Microsoft SQL server database back end and the

predictive models to determine the possibility of diabetes comorbidities in future. To recap, after importing the Excel csv data file into the SQL Server database, the entire process can be grouped into three distinct phases, namely, predictive modeling, dashboard design and assessment tool. The steps within each phase are listed below.

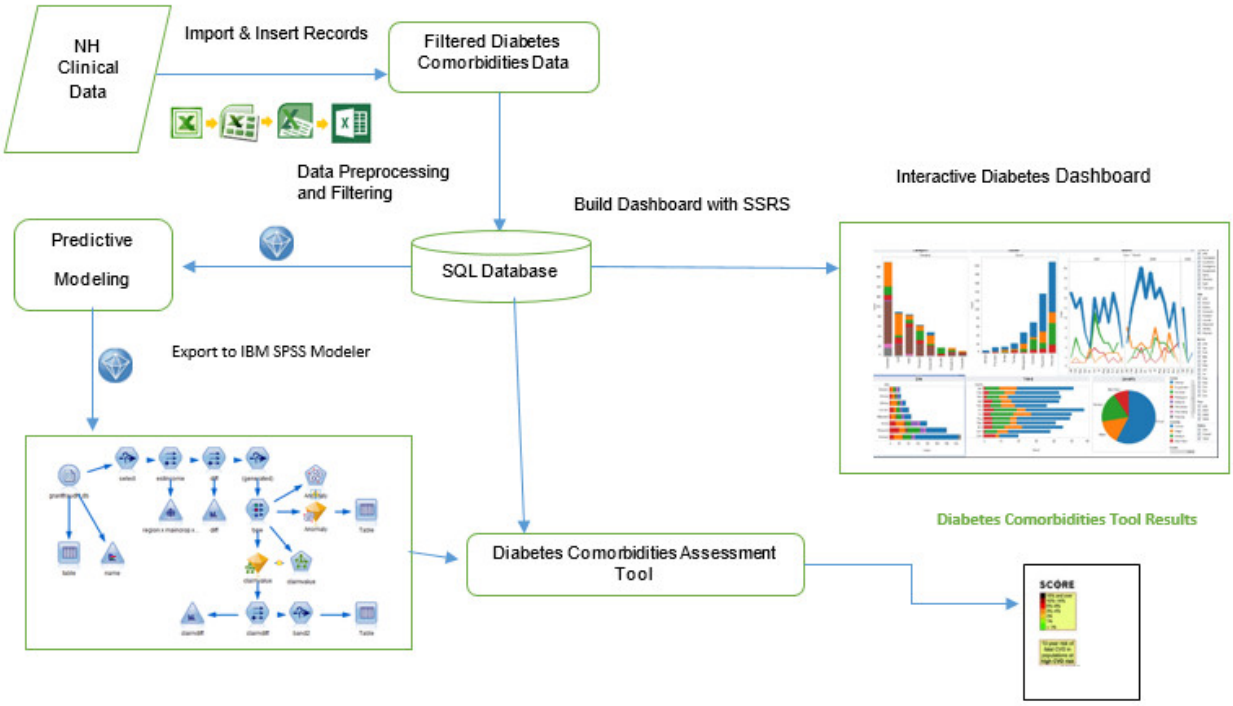


Figure 8 Components for Predictive Modeling and Data Visualization

Phase 1: Data preprocessing and modeling

- Data was cleansed and prepared for the model.
- After data preprocessing, testing and training tables for three diabetes comorbidities were created.
- Relationships were established within the database to associate demographic and diagnostic data from different tables for data visualization.
- Three predictor variables were chosen from the top twenty diagnostic codes. A separate model was built for each of these variables. The remainder diagnostic codes together with demographic data then became the input variables.
- Various data mining algorithms such as logistic regression, decision tree and artificial neural network together with their ensembles were compared for relevance and accuracy.
- To evaluate the model, the data was divided in two sets, one for training and the other for testing. A larger dataset was used for training which improved accuracy of the algorithm. The test data, which was a smaller dataset, was then used to evaluate performance.

Phase 2: Dashboard

- A dashboard was designed to allow users to analyze/compare various key performance indicators (KPIs) and filter results by the selected parameters. The drilldown capabilities of the dashboard allow filtering by specific demographics and understand KPIs at a finer granularity.

- The performance of various data mining algorithms is also shown in the dashboard.

Phase 3: Diabetes Comorbidity Assessment tool

The results from predictive models were integrated with a web-based, user-friendly assessment tool to predict likelihood of comorbidities for individual patients. This tool displays the risk score for diabetes comorbidities.

3.2 Data Source

A clinical dataset obtained from Northern Health (NH) has been used for this research. This dataset consists of patients who have accessed one of the eighteen NH facilities in three Health Service Delivery Areas (Northeast, Northern Interior, Northwest). The NH dataset exclusively consists of diagnosed diabetic patients who were admitted to these facilities for either acute care or day surgery. All patients were diagnosed with at least one of 4,592 unique diagnostic codes. The dataset used for this research consists of a total of 141,900 records representing 34,824 unique admissions for the period from April 1, 2012 to March 31, 2018. It is to be noted that these timelines were specified in fiscal years (2012/13-2017/18) and there were no cases of gestational diabetes. The variables included in this dataset are:

- Patient Code
- Stay Code (this code is unique to a particular acute/daycare stay (visit))
- Diagnosis Code Order of Entry (identifies the order in which the diagnosis codes were abstracted)

- Health Service Delivery Area
- Facility Name
- Diagnosis Code (ICD-10-CA code that describes the diagnoses, conditions, problems, or circumstances of the patient during the length of stay in the health care facility)
- Diagnosis Code Long Description
- Age Units
- Average Total Length of Stay (the summation of both the acute care length of stay and the ALC length of stay)
- Physician Code (has family doctor or not)

This dataset is a reliable source for building predictive models and analyzing data because it only consists of diagnosed diabetic patients with associated diagnostic codes.

In addition, it is also to be noted that this dataset has been anonymized by Northern Health to protect the privacy of patients. No personal information of patients was included in the dataset.

3.3 Data Preprocessing

Data preprocessing is the technique of cleaning and processing the data to ensure efficient and accurate adaptation by different data mining algorithms. The performance of predictive models not only depends on the data mining/machine learning techniques, but

is also highly dependent on the data quality. Hence, it is imperative to ensure that negative factors such as noise, missing values and inconsistencies are addressed through data preprocessing methods [34]. Figure 9 shows the common data preparation and data reduction techniques involved in data preprocessing [35].

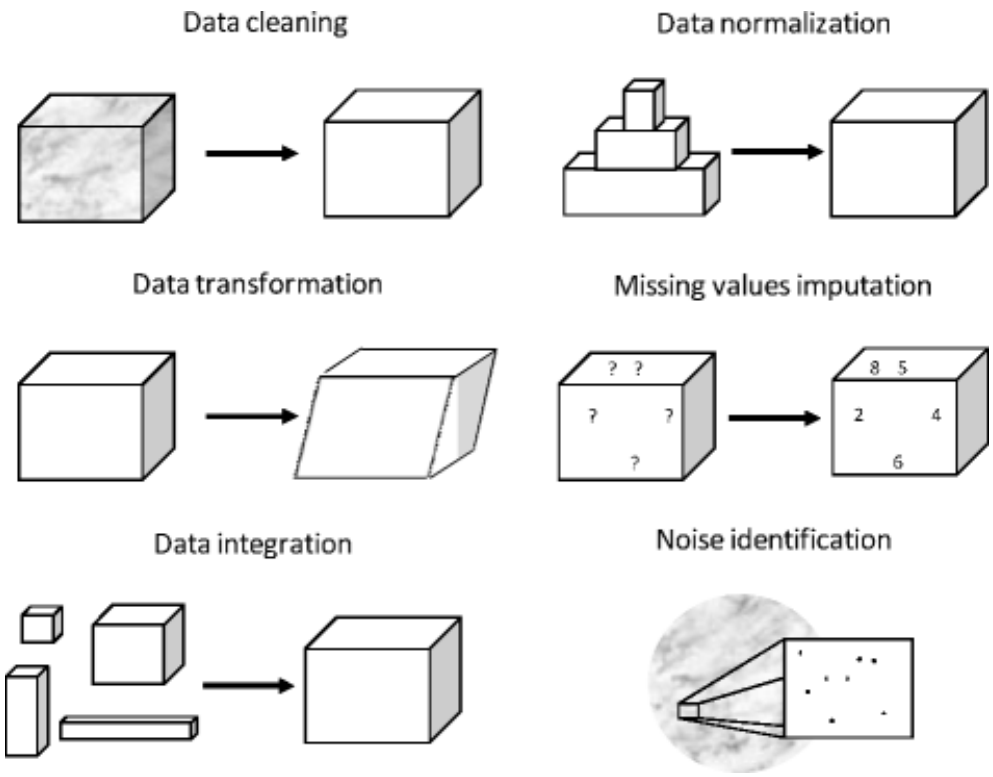


Figure 9 Tasks in Data Preprocessing [35]

For this research, data preprocessing has been done to obtain a specific set of relevant variables through extensive analysis and filtering.

Dataset Preprocessing: The dataset provided by NH was an Excel sheet in a csv format. This sheet was converted into a database. There were a total of 141,900 records with

stay codes repeating multiple times. To resolve this, pivot queries were used to get a table with the 34,824 unique admissions. This was verified with the original Excel sheet. This pivoting required that diagnosis codes for comorbidities be added as columns in the table for each unique admission. This was deemed to be an unnecessary overhead because 73% of these codes occurred in less than ten cases in the original dataset. The process implemented for exclusion of comorbidities is explained in detail in the next section. Considering all these factors, the top twenty diagnosis codes with maximum counts were chosen to build the model. These codes together with their descriptions and counts are listed in Table 4 below:

Table 4 Top Twenty Diagnostic Codes by Count

	Diagnosis Code	Diagnosis Description	Count
1	E119	Type 2 diabetes mellitus without (mention of) complications	13,268
2	E1152	Type 2 diabetes mellitus with certain circulatory complications	6,516
3	I100	Benign hypertension	3,598
4	E149	Unspecified diabetes mellitus without (mention of) complication	2,526
5	E1164	Type 2 diabetes mellitus with poor control, so described	2,452
6	I500	Congestive heart failure	2,303
7	E1123	Type 2 diabetes mellitus with established or advanced kidney disease	2,262
8	N179	Acute renal failure, unspecified	1,714
9	N390	Urinary tract infection, site not specified	1,673
10	N0839	Unspecified glomerular disorders in diabetes mellitus	1,429
11	E1138	Type 2 diabetes mellitus with other specified ophthalmic complication not elsewhere classified	1,337
12	H251	Senile nuclear cataract	1,297
13	E1128	Type 2 diabetes mellitus with other specified kidney complication not elsewhere classified	1,293
14	Z22300	Carrier of drug-resistant staphylococcus	1,255
15	J189	Pneumonia, unspecified	1,102

16	E109	Type 1 diabetes mellitus without (mention of) complication	1,058
17	Z22302	Carrier of drug-resistant enterococcus	996
18	U980	Place of occurrence, home	967
19	Z515	Palliative care	958
20	E1178	Type 2 diabetes mellitus with multiple other complications	939

Based on literature and Table 4, three diagnosis codes were selected as predictor or target variables:

1. I500 (Congestive Heart Failure)
2. I100 (Benign Hypertension)
3. N179 (Acute Renal Failure)

For each of these diagnosis codes, training and testing datasets were initially created with a ratio of 70:30. This ratio was later adjusted to study the efficiency of the models.

It is to be noted that the final dataset used to build the predictive models aggregated patient admissions which resulted in each patient to have only one record to be consistent with the total number of unique patients (14,016) after exclusions. This is due to the reason that some patients had recorded a comorbidity in one of their admissions but in the subsequent admissions, these comorbidities were missing which lead to data inconsistencies. To handle this particular issue, it was assumed that if a patient had been diagnosed with one of the twenty comorbidities in any one of their admissions, then they were recorded with that particular comorbidity. This particular issue has also been explained in the challenges section of this chapter.

3.4 Inclusion and exclusion

For relevance of data, inclusion/exclusion was done at two stages. First, NH ensured that the data consisted of only diabetic patients. Second, irrelevant /redundant data was excluded and only relevant variables based on literature were retained.

NH Inclusion/Exclusions: Discharges were included if at least one of the following diagnosis codes was found on the record: E11*, E12* E13*, E14* and/or E232.

- Type 1 Diabetes codes begin with E10
- Type 2 Diabetes codes begin with E11
- Type other codes begin with E13 and include Diabetes Insipidus E232 Diabetes
- Type unspecified codes begin with E14

3.5 Predictive Modeling Inclusions/Exclusions:

Large datasets face curse of the dimensionality problem which impedes operations of data mining algorithms raising the computational costs [36]. One solution to handle this issue is to use the Feature Selection (FS) algorithm. FS eliminates irrelevant and redundant variables. The NH dataset includes demographic and diagnostic data for each patient with every admission. The variables including the twenty diagnostic codes identified in data preprocessing stage were evaluated for importance using the FS algorithm [37].

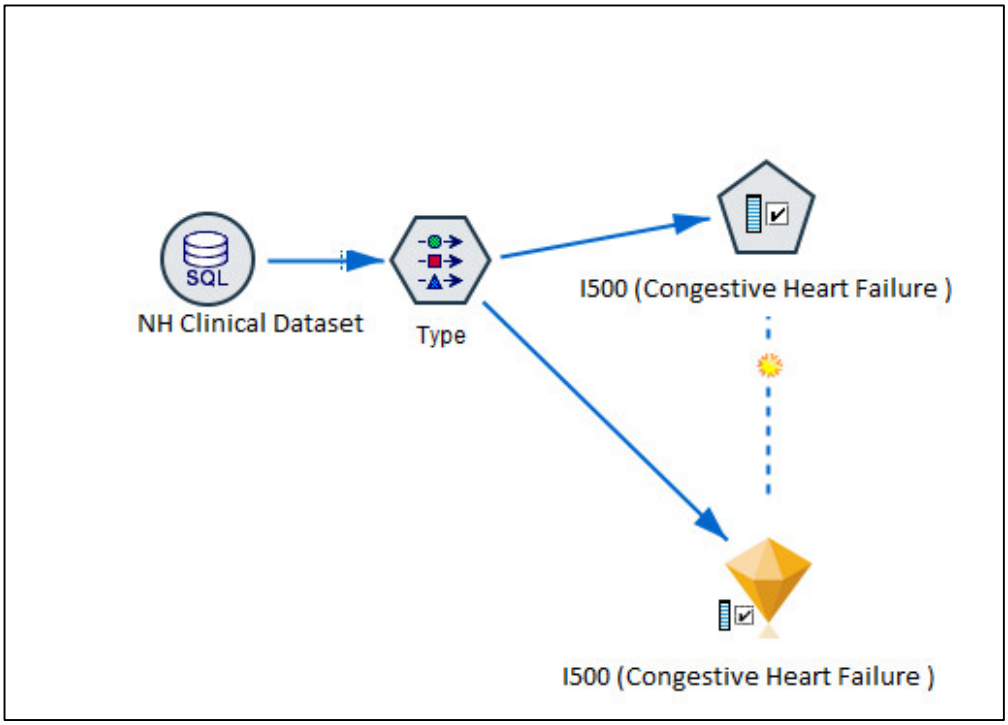


Figure 10 Feature Selection Model

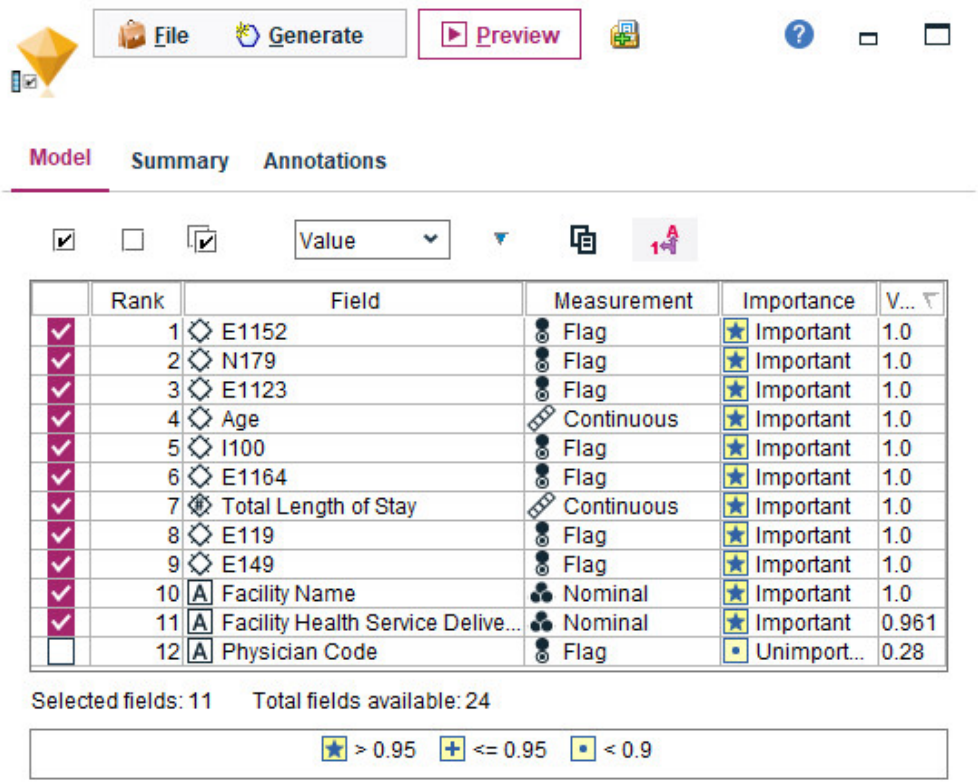


Figure 11 FS Model Results

Figure 10 shows the FS model used on the NH clinical dataset. In this Figure, the SQL data source represents the NH database, and Type identifies the data types of variables; Type is also used to select the target predictor variable I500 (Congestive Heart Failure). The golden model nugget contains the results of the FS model. Figure 11 shows the variables evaluated and ranked by order of importance by the FS model. The Field column provides the names of input variables which were described earlier in the data preprocessing section. The Measurement column describes the variable type as identified by the SPSS modeler either as continuous, nominal, ordinal or flag. Continuous is used for integers, real numbers and date/time; Nominal can be used for numeric/string/date or time; Flag is used for two distinct values such as true/false or binary values 0 or 1. It is

to be noted that out of the twenty diagnosis codes included, only seven were identified as important by FS. This is due to the reason that remainder of the diagnosis codes had a majority of values which were '0' which prevented FS from ranking them; instead, these codes were categorized as "Single category too large". Since these variables were not evaluated as unimportant by FS, they were included in the final dataset as input. The finalized list of variables including diagnostic and demographic information was explored further for data analysis and visualization. The data mining algorithms which were used include Logistic Regression, CHAID, Neural Network, Random Forest, Bayesian Network and Ensemble. Some of these algorithms are described later in this chapter.

The following variables were also excluded:

- Stay Code - This code is unique to a particular acute/daycare stay (visit)
- Fiscal Year - A fiscal year ranges from April 1 of the current year to March 31 of the following year
- Fiscal Period - Periods within the fiscal year (The days in period 1 and 13 will vary, the remaining periods 2-12 will always be 28 days)
- Discharge Date - The date the patient was discharged from the hospital
- Institution Type - Identifies whether the hospital stay was an acute care stay or a daycare visit
- Age Code - Age code is either Year (Y), Months (M), or Days (D)
- Acute Length of Stay - The length of stay in days associated with the acute care portion of the stay
- ALC Length of Stay - ALC length of stay is the number of days a patient is classified as alternate level of care

- Patient Community - Community of patient residence
- Patient LHA - Local Health Area of patient residence
- Patient Province - Province of patient residence

Stay code, Fiscal Year, Fiscal Period, Discharge Date were excluded because patient data from different admissions was aggregated into one record for each patient. Age Code specified the units in which the age was recorded (year, months or days). This only resulted in removal of five patients. 'Acute Length of Stay' and 'ALC Length of Stay' were excluded as the variable Total Length of Stay captured this information by default. Since multiple admissions for patients and their diagnoses were aggregated, the average of Total Length of Stay was calculated for all patients. Patient Community and LHA were excluded due to patient migration across communities resulting in data inconsistency. Patient Province was also excluded as majority of patients were from British Columbia and this information was redundant. These eleven exclusions reduced the dataset to twenty-six variables. Out of these variables, Patient Code (unique identifier) and the target variable are not considered to be input variables thereby leaving twenty-four variables which are listed below:

1. E119 - T2D without complications
2. E1152 - T2D with certain circulatory complications
3. E149 - Unspecified diabetes mellitus without (mention of) complication
4. E1164 - T2D with poor control
5. I100 - Benign hypertension

6. E1123 - T2D with established or advanced kidney disease
7. N179 - Acute Renal Failure
8. N390 - Urinary tract infection
9. N0839 - Unspecified glomerular disorders in diabetes mellitus
10. E1138 - T2D with other specified ophthalmic complication
11. H251 - Senile nuclear cataract
12. E1128 - T2D with other specified kidney complication
13. Z22300 - Carrier of drug-resistant staphylococcus
14. J189 - Pneumonia, Unspecified
15. E109 - T1D without complication
16. Z22302 - Carrier of drug-resistant enterococcus
17. U980 - Place of occurrence, home
18. Z515 - Palliative care
19. E1178 - T2D with multiple other complications
20. Facility HSDA - Facility Health Service Delivery Area
21. Facility Name - Specifies facility in which patient is admitted
22. Age - Specifies age of patient in years
23. Average Length of Stay - Average of total length of stay
24. Physician Code - Specifies if a patient has family physician or not

It is to be noted that, the above input variables are for predicting I500 which is the reason for it not to be included as an input. An analogous process was followed for predicting I100 and N179.

Out of 14,021 patients, there were only five patients who had their age units recorded either as month or days; these were excluded from the study thus reducing the dataset to 14,016 records.

3.6 Predictive Modeling

Predictive modeling is a statistical data mining technique normally used to predict future behaviour. Predictive models analyze historical and current data to predict future outcomes. Data mining algorithms, such as logistic regression, decision tree and neural networks, have been used for building predictive models for early detection of diabetes [11]. There are two types of learning used by data mining algorithms - supervised learning and unsupervised learning [43]. Supervised learning uses labeled data whereas unsupervised learning uses unlabeled data. Labeled data refers to data accompanied with metadata, while unlabeled data lacks this information. Unsupervised learning has been primarily used to solve association, clustering and anomaly detection problems. In contrast, supervised learning are more suited to solve classification problems. With the exclusive use of labeled data, this research makes use of supervised learning methods in order to generate predictive models for classification. Some of the data mining algorithms which use classification are briefly explained below:

Artificial Neural Networks

Artificial Neural Networks (ANNs) are biologically inspired models which have recently found their applications in the field of healthcare. ANNs are based on the brain structure and can be used to model extremely complex nonlinear functions. ANNs can be used in sophisticated predictive applications such as multilayer perceptron (MLP) and radial basis function (RBF) networks [38].

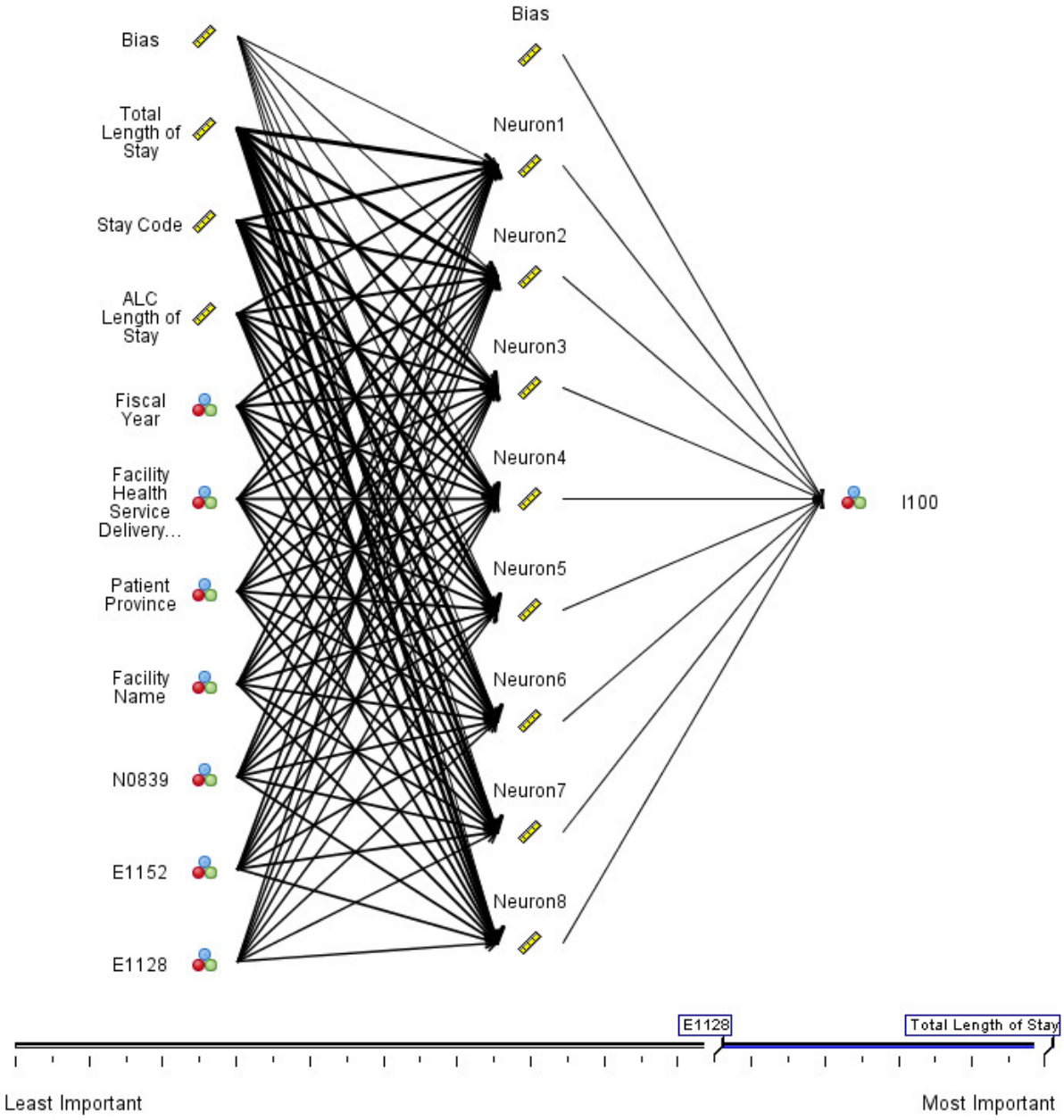


Figure 12 Neural Network Mapping

As seen in Chapter 2, ANNs have been effectively used for building predictive models for diabetes. An example of neural network mapping shows how input data is matched to the predictor variable which in this case was I100 (Hypertension) (Figure 12). Each neuron output is calculated by the sum of inputs and activation functions.

Some advantages of ANNs are given below [39]:

- Information is stored on a network which ensures that it can function even with missing values; models can be trained to produce results even with incomplete data.
- ANNs lend well to parallel processing.
- ANNs provide better fault tolerance; if one or more cells is corrupted, it still generates results.

A notable disadvantage of ANNs is that solutions probed are unexplained in some cases which reduces trust on the network.

Logistic Regression

Logistic regression has been used typically in the analysis of binary outcomes. It is a statistical method for prediction of probability of occurrence of an event which is represented by 1 and a non-event by 0. Predictor variable in logistic regression can be either qualitative or quantitative [38].

Figure 13 shows the importance of the different input variables identified by the logistic regression model for predicting I100 (Hypertension).

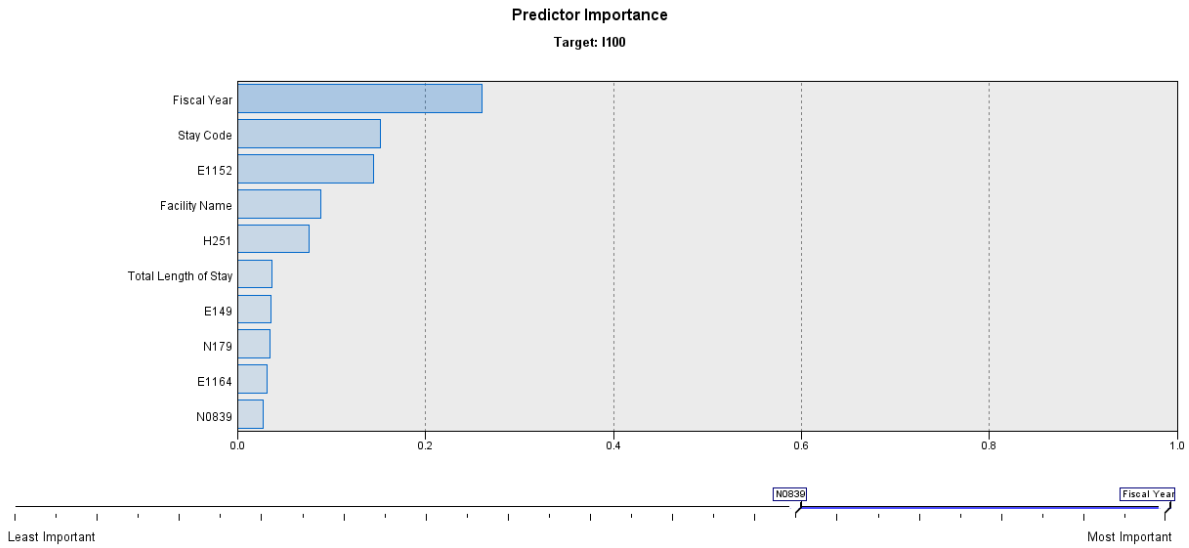


Figure 13 Logistic Regression Predictor Importance

Decision Tree

A decision tree assigns probability to each of the possible choices based on the context of the decision and acts as a decision-making device**.** Decision tree has attribute nodes that can be linked to two or more subtrees. Brieman's classification and regression tree (CART) is one of the popular decision tree algorithms [16].

Random Forest

Random Forest is an algorithm where subsets of a given dataset are chosen and multiple classification trees are generated. A forest is then created from the ensemble of these trees. Random Forest has been used in diabetes research and is recommended as an efficient algorithm for building predictive models [40].

Bayesian Network

Bayesian Network algorithm is based on probabilistic theory and represents a set of variables and their dependencies in the form of an acyclic graph [41]. This could be used to identify relationships between a disease and its associated symptoms. Bayesian Network has been used in the past to predict T2D patients [42].

CHAID

CHAID (Chi Square Automatic Interaction Detection) is a decision tree technique [43]. CHAID can be used to find the relationship between input and target variables. It is to be noted that CHAID can have limitations due to the sample size of predictor variable. CHAID has also been used to build predictive models for diabetic patients [44].

The performance of a predictive model depends on various factors such as data quality, structure of data and variable selection [36]. In 2013, a study was done comparing the performance of logistic regression and artificial neural network (ANN) models for identifying risk factors for diabetes mellitus using IBM SPSS Modeler [38]. Figure 14 shows the overall data modeling process, and Figure 15 shows the process specifically for partitioned data. The dataset consisted of 229 diabetes patients, 69.9% of whom had uncontrolled blood glucose level. Results revealed that ANN model had a higher classification accuracy of 72.5% in comparison to logistic regression which had an accuracy of 69.9%. Similar results were recorded for partitioned data with ANN model having an accuracy of 72.5% while logistic regression had an accuracy of 71.35%.

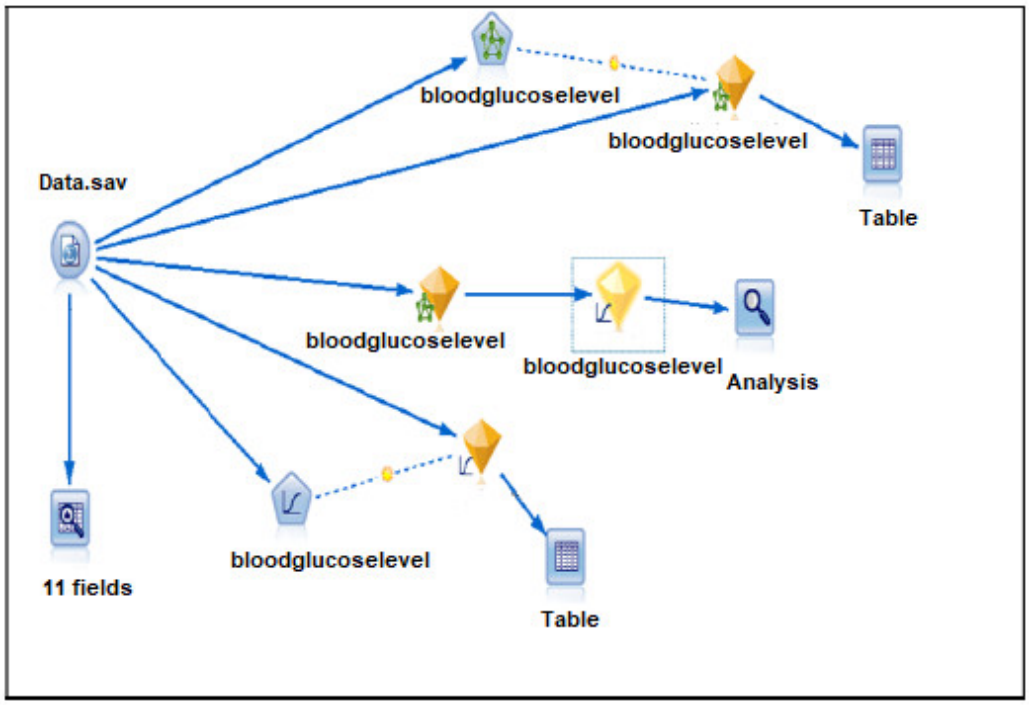


Figure 14 Data Mining Process for Entire Sample Data [38]

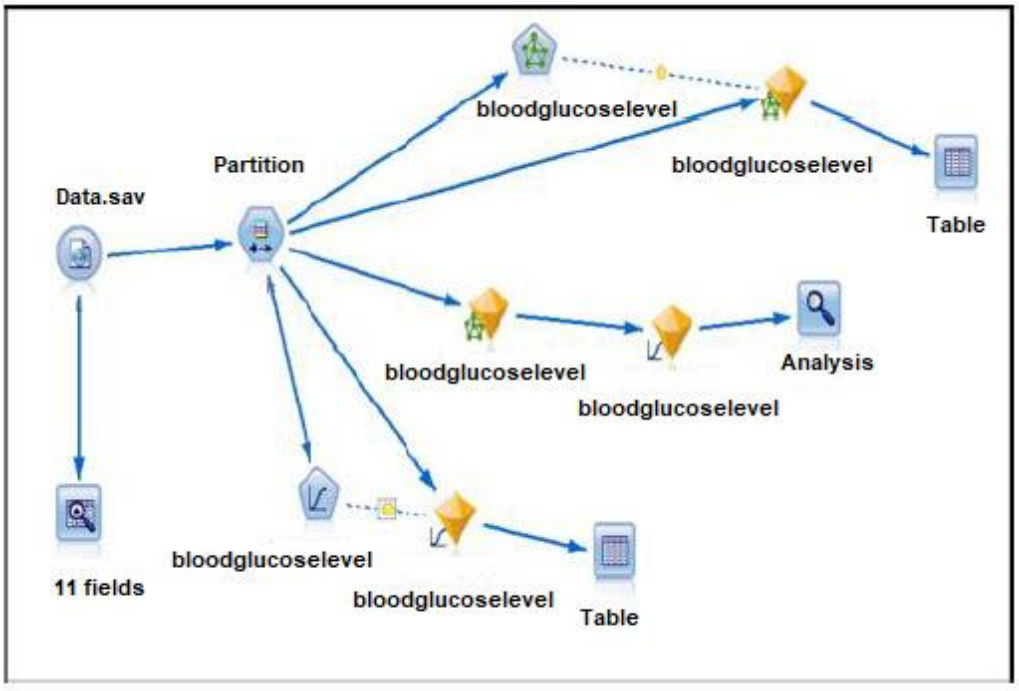


Figure 15 Data Mining Process for Partitioned Sample Data [38]

3.7 IBM SPSS Modeler

A number of predictive modeling tools such as R, Weka, Orange, Rapid Miner, GraphLab Create, Octave and IBM SPSS are available. The IBM SPSS Modeler is a graphical data science and predictive analytics platform which assists in providing insights to improve decision making. For this research, IBM SPSS Modeler v18.1 [22] was chosen due to its features listed below:

- Advanced statistical analysis
- Easy to use and flexible
- Supports multiple data sources including Microsoft SQL database
- Graduate licensing packages for students.
- Supports all phases of data mining including model development, deployment and refreshing
- Ability to merge data from multiple sources
- Options to choose advanced data mining algorithms to build predictive models
- FS algorithm which helps identify important variables
- Supports use of smaller datasets
- Scalable platform makes it accessible to users of different skill levels
- Automated data preparation and modeling
- Visual analytics
- Multiple deployment methods

3.8 Challenges

Data

The major challenge was the selection of appropriate variables as there was a lot of diagnostic data with minimal records. Data inconsistency was another issue as there were patients who had missing diagnosis codes on readmission. Both these challenges are elaborated in detail below. Each patient had been recorded with at least one or more diagnosis codes for every admission. The Table 5 below shows the distribution of diagnosis for patients.

Table 5 Diagnosis/Patient Distribution

No of Diagnosis	No of Patients
2-5	7,140
6-10	3,382
11-15	1,490
16-20	756
>20	1,044

To filter and select the diagnosis codes which were relevant to this research was a tedious task. This has been elaborated in the data preprocessing section. Even after filtering and selecting prominent twenty diagnostic codes, there were inconsistencies found with the data. For instance, there were patients who had been recorded with T2D the first time

they came in and on readmission, the same patient was recorded with Type 1 diabetes. An example of this case scenario is shown for a patient in Figure 16.

Diagnosis Code	Diagnosis Long Description	Stay Code	Discharge Date
M6651	Spontaneous rupture of unspecified tendon, shoulder region	45495	2013-10-31 00:00:00.000
M779	Enthesopathy, unspecified site	45495	2013-10-31 00:00:00.000
M2571	Osteophyte, shoulder region	45495	2013-10-31 00:00:00.000
M751	Rotator cuff syndrome	45495	2013-10-31 00:00:00.000
E119	Type 2 diabetes mellitus without (mention of) complications	45495	2013-10-31 00:00:00.000
M171	Other primary gonarthrosis	26534	2015-02-19 00:00:00.000
E109	Type 1 diabetes mellitus without (mention of) complication	26534	2015-02-19 00:00:00.000

Figure 16 Data Inconsistency Example

The data inconsistency issue raises the question that what happened to the patient's previous diagnosis of T2D on the first admission.

Another challenge was that not all of the recorded diagnosis codes were repeating on readmission. For instance, diagnosis code I100 (hypertension) was recorded for a patient on the initial admission but this diagnosis code was not recorded upon readmission (Figure 17). This data inconsistency made it extremely challenging to build the predictive models with longitudinal data. To resolve this issue, the dataset used for building the predictive models retained all diagnosis codes for a patient if it was recorded in any of their admissions.

Diagnosis Code	Diagnosis Long Description	Stay Code	Discharge Date
K8001	Calculus of gallbladder with acute cholecystitis with ob...	20254	2013-01-29 00:00:00.000
I100	Benign hypertension	20254	2013-01-29 00:00:00.000
E119	Type 2 diabetes mellitus without (mention of) complicat...	20254	2013-01-29 00:00:00.000
Y443	Anticoagulant antagonists, vitamin K and other coagul...	20254	2013-01-29 00:00:00.000
T811	Shock during or resulting from a procedure, not elsewh...	20254	2013-01-29 00:00:00.000
Y839	Surgical procedure, unspecified, as the cause of abno...	20254	2013-01-29 00:00:00.000
R000	Tachycardia, unspecified	20254	2013-01-29 00:00:00.000
T810	Haemorrhage and haematoma complicating a procedu...	23730	2013-02-09 00:00:00.000
T796	Traumatic ischaemia of muscle	23730	2013-02-09 00:00:00.000
T811	Shock during or resulting from a procedure, not elsewh...	23730	2013-02-09 00:00:00.000
N179	Acute renal failure, unspecified	23730	2013-02-09 00:00:00.000
Y836	Removal of other organ (partial) (total) as the cause of ...	23730	2013-02-09 00:00:00.000
I460	Cardiac arrest with successful resuscitation	23730	2013-02-09 00:00:00.000
Y848	Other medical procedures as the cause of abnormal re...	23730	2013-02-09 00:00:00.000
T813	Disruption of operation wound, not elsewhere classified	23730	2013-02-09 00:00:00.000
Y838	Other surgical procedures as the cause of abnormal re...	23730	2013-02-09 00:00:00.000
E1152	Type 2 diabetes mellitus with certain circulatory compli...	23730	2013-02-09 00:00:00.000
H110	Pterygium	42620	2013-12-17 00:00:00.000
E119	Type 2 diabetes mellitus without (mention of) complicat...	42620	2013-12-17 00:00:00.000
C793	Secondary malignant neoplasm of brain and cerebral ...	21666	2017-11-27 00:00:00.000
G941	Hydrocephalus in neoplastic disease	21666	2017-11-27 00:00:00.000
G328	Other specified degenerative disorders of nervous syst...	21666	2017-11-27 00:00:00.000
C3410	Malignant neoplasm of upper lobe, right bronchus or lu...	21666	2017-11-27 00:00:00.000
E119	Type 2 diabetes mellitus without (mention of) complicat...	21666	2017-11-27 00:00:00.000

Figure 17 Data Inconsistency I100 (hypertension)

To illustrate this, the diagnosis code I100 (hypertension) was retained for the patient even though it was not recorded in the recurring admissions (Figure 17). For patient 110 (Figure 16) diagnosis codes for both T1D and T2D were retained. This ensured that no diagnosis code was missed for building the predictive models. The corresponding dashboard is consistent with this logic as well.

As mentioned earlier, all patients in this dataset were diabetic with diagnosis codes recorded for T1D, T2D, other types of diabetes, diabetes insipidus and unspecified diabetes. In total, there were 100 diagnosis codes for different types of diabetes with zero or more associated comorbidities. These 100 diagnosis codes included diabetes comorbidities such as E1123 (Type 2 diabetes mellitus with established or advanced kidney disease) and E1128 (Type 2 diabetes mellitus with other specified kidney complication not elsewhere classified). For kidney/renal comorbidities, there were a total of 84 diagnosis codes. The word 'kidney' was found in 45 diagnosis codes and 39 diagnosis codes contained the word 'renal'. Some of the patients who were recorded with diagnosis codes such as E1123 were also recorded with other diagnosis codes such as N179 (Acute renal failure) (Figure 18). The American Urological Association quotes renal as a synonym for kidney [45], this was an interesting observation which could lead to the possibility of combining diagnosis codes with the word renal/kidney under one umbrella. However, there were 84 diagnosis codes for words kidney/renal alone and they included different types of diabetes as well as other comorbidities. To combine these meaningfully, a considerable medical background would be required. Thus, it was not explored further in this research. All of these challenges made it tedious to finalize the predictor variables filtering the different diagnosis codes. Thus diagnosis codes with more prominence were chosen as the target variables.

In Figure 18, it can also be seen that a patient was recorded with T2D on their first visit, but the same patient was recorded with unspecified diabetes in a subsequent admission. Considering these anomalies in the dataset, the predictive models were built for all diabetic patients instead of a specific diabetic type.

Diagnosis Code	Diagnosis Long Description	Stay Code	Discharge Date
J441	Chronic obstructive pulmonary disease with acute exacerbation, unspecified	22798	2012-06-05 00:00:00.000
N179	Acute renal failure, unspecified	22798	2012-06-05 00:00:00.000
E1123	Type 2 diabetes mellitus with established or advanced kidney disease	22798	2012-06-05 00:00:00.000
N0839	Unspecified glomerular disorders in diabetes mellitus	22798	2012-06-05 00:00:00.000
R074	Chest pain, unspecified	22798	2012-06-05 00:00:00.000
I100	Benign hypertension	22798	2012-06-05 00:00:00.000
J441	Chronic obstructive pulmonary disease with acute exacerbation, unspecified	22905	2012-07-11 00:00:00.000
B029	Zoster without complication	22905	2012-07-11 00:00:00.000
I100	Benign hypertension	22905	2012-07-11 00:00:00.000
N179	Acute renal failure, unspecified	22905	2012-07-11 00:00:00.000
E1428	Unspecified diabetes mellitus with other specified kidney complication not elsewhere classified	22905	2012-07-11 00:00:00.000

Figure 18 Diabetic Patient with Kidney Disease

Please note that the stay code is unique for each patient visit and no patient information is included in Figure 16, Figure 17 and Figure 18 to ensure patient privacy.

Integration

Software from different vendors such as Microsoft SQL [46] for database, Microsoft SSRS [47] for dashboard and IBM SPSS Modeler [22] for building predictive models required integration packages to be built. For instance, to connect the Microsoft SQL data source [46] with IBM modeler [22], an ODBC (Open Database Connectivity) data source had to be created. Similarly, the results from SPSS modeler had to be exported to MS SQL database for analysis using SSRS [47].

3.9 Data Visualization

3.9.1 Dashboard

Dashboards are an effective tool which allow visualization of large amounts of data in an intuitive manner. The key performance indicators are integrated into visual displays which can be further drilled down for finer granularity. These do not only provide insight into data, but can be effective for quickly finding information. For instance, it was demonstrated that the mean time to find all data elements was 6.3 minutes using conventional approach compared to 1.9 minutes using a diabetes dashboard. The research further established that analysis tools like dashboards can be an insightful asset for healthcare information technology [27]. As seen in Chapter 2, a similar research [28] by the Canadian Diabetes Association established that health professionals were more effective in treating diabetes patients when using a dashboard which provided knowledge of other risk factors and associated guidelines [28]. Similarly, the patients who are presented with a dashboard listing the risk factors tend to benefit from the knowledge contained therein.

To build a dashboard, it is essential to choose an appropriate visualization tool. A number of tools are now available for this purpose. These include Tableau [48], QlikView [49], Datawrapper [50], Fusioncharts [51] and SSRS [47]. In this research, SSRS was selected

due to its simplicity, capability to produce interactive visualizations and ability to adjust with fast changing datasets.

3.10 SSRS

SSRS is a business intelligence module which enables users to create visually appealing reports via charts, maps and dashboards [47]. SSRS provides features including:

- Compatibility with different data sources ranging from simple Excel sheets to databases
- Interactive sorting capabilities
- Drilldown/Drillthrough reporting
- Security via access controls
- Intuitive Visualization
- Export features – reports can be exported to various formats including Word, Excel, PowerPoint, pdf, TIFF, MHTML, CSV, and XML

SSRS requires a backend database and a wrapper for rendering reports in a browser. For this research, the Microsoft SQL server database was used.

SSRS reporting has been effectively used in healthcare to maximize profits, minimize risks, reduce costs and enhance patient experience. In addition, it has also been used for presenting data in a user-friendly form (Figure 19).

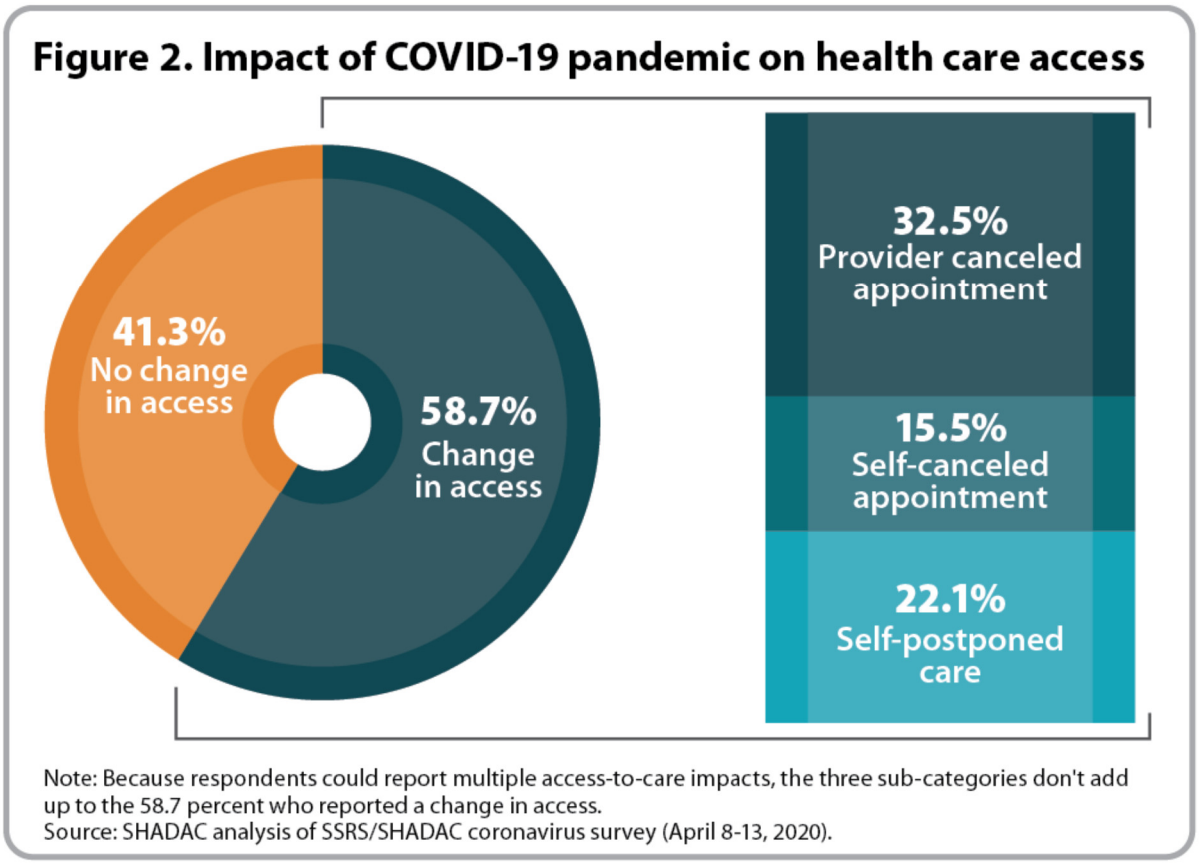


Figure 19 COVID Prevalence in the World [29]

3.11 Summary

The proposed models are based on several data mining algorithms and produce data-driven results which can assist physicians in developing effective treatment plans. The dashboard presents insightful information such as obesity rates based on geographical locations, food habits and the overall trend for diabetes over the years. Such information can educate the users about diabetes and its impact on health. The assessment tool, which uses results from predictive models, would facilitate users to be better informed about diabetes comorbidities. The existing diabetes risk score calculators are mostly

based on limited paper based questionnaires and not easily accessible. In summary, the study methodology consists of the following steps:

- NH dataset was imported into SQL database
- Data preprocessing techniques were used to eliminate redundant information.
- Data analysis allowed selection of the input and three predictor variables for diabetes comorbidities.
- The database served as a data source for the IBM SPSS Modeler and was used to evaluate relative performance of various data mining algorithms.
- The accuracy of each algorithm was determined using training and testing datasets.
- The results from the models were displayed using SSRS via an interactive dashboard.
- A user-friendly tool was developed to calculate the risk of developing comorbidities for individual patients.
- The dashboard was integrated with the diabetes assessment tool.

The information provided by the predictive model will be both helpful and insightful for diabetes patients as well as non-diabetic users to have a better understanding of their health and act as an effective indicator to further discuss with a healthcare practitioner.

Chapter 4

Experiments and Results

In this chapter, the experiments and results of this research are discussed. This chapter is split into two parts. First, the diabetes dashboard is explained, and then results from the predictive modeling for three target variables are presented. For both these parts, the NH clinical dataset was used which included only diabetic patients who accessed one or more of the NH facilities between the period 2012-2018.

The Diabetes Dashboard consists of three main reports with drilldown capabilities. The first report shows overall aggregated statistics for the NH diabetes dataset; the second report is the Diabetes Types and Comorbidities dashboard which shows the prominent comorbidities for patients with different types of diabetes and also includes the prominent Primary Diagnosis Codes. The third report is the HSDA comparison which shows aggregated patients and admission statistics across the three Health Service Delivery Areas - Northwest, Northeast and Northern Interior.

Predictive Modeling is done using five base classification algorithms together with their ensemble for three comorbidities (Hypertension, Congestive Heart Failure and Acute Renal Failure) using IBM SPSS Modeler. The results for each target variable is shown with the corresponding explanation and analysis. The relationships found between the input and target variables using the FS algorithm are also explained followed by a summary of the analysis.

4.1 Diabetes Dashboard

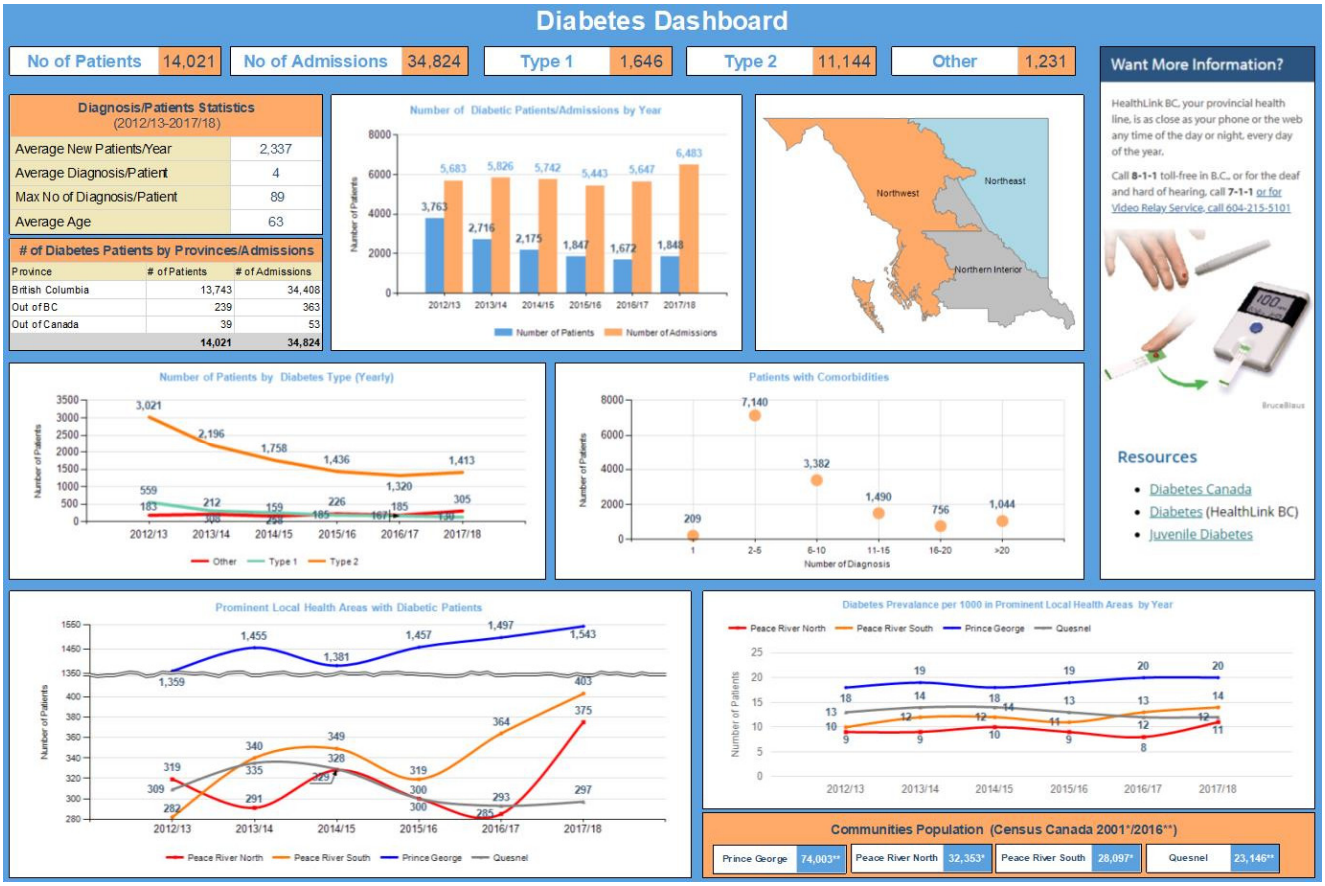


Figure 20 Diabetes Dashboard

Figure 20 shows the main diabetes dashboard which displays the clinical data sliced along various dimensions including population, diagnosis codes, diabetes types, admissions and comorbidities for patients admitted in NH facilities over the years. The dashboard also allows navigation to reports at a finer granularity via drilldowns. Each of the charts/tables included in this dashboard is further explained below. The image on the top-right was obtained from Diabetes Canada [8].

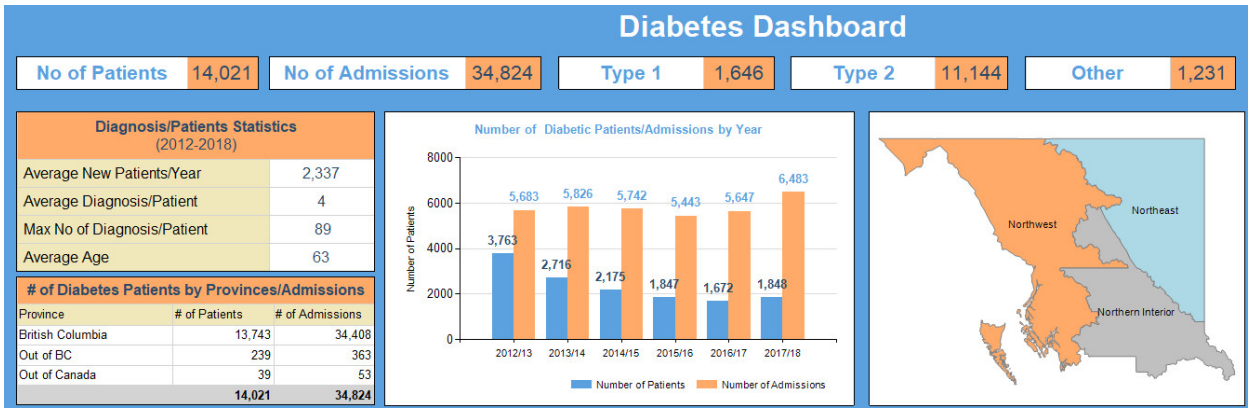


Figure 21 Diabetes Dashboard Overall Statistics

Figure 21 shows an overview of aggregated statistics obtained from the NH clinical dataset. The top row shows that there were a total of 14,021 patients with 34,824 admissions which averages out to approximately three admissions per patient. Out of these patients, 12% with T1D, 80% were diagnosed with T2D, and remaining 8% with other types of diabetes (includes diabetes insipidus, other and unspecified types). The Diagnosis/Patient Statistics table shows that the average age of all patients over the years was 63 and an average of 2,337 patients were admitted each year. Another observation was that the admitted patients recorded an average of four diagnoses from the possible 4,592 diagnosis codes. The maximum number of diagnosis codes recorded for a patient was 89, there were four patients who recorded more than 80 comorbidities and fifty patients who recorded between 50-80 comorbidities. A detailed breakdown of comorbidities is shown in Figure 25. The number of diabetic patients and admissions by province is also shown in Figure 21. Drilldown from this chart shows these numbers for each LHA specific to British Columbia (*Figure 22*). It should be noted that the higher number of patients in the drilldown is due to the patient migration which records the patient

more than once. However, this anomaly does not impact the model. LHAs with fewer than ten patients and those recorded as ‘Unknown‘ were grouped in a single category labeled as ‘Other’.

# of Diabetes Patients/Admissions by LHA		
Patient LHA	Patient	Admissions
Prince George	5,072	12,728
Peace River South	1,269	2,946
Quesnel	1,199	2,659
Peace River North	1,197	2,750
Terrace	964	2,286
Nechako	797	2,105
Prince Rupert	725	1,934
Kitimat	656	1,750
Smithers	588	1,487
Burns Lake	376	1,106
Queen Charlotte	242	709
Upper Skeena	226	709
Fort Nelson	203	525
Other	157	196
Nisga'a	98	214
Cariboo-Chilcotin	84	113
Snow Country	21	26
Vernon	19	22
Midtown	18	24
Telegraph Creek	16	31
Central Okanagan	15	16
Surrey	13	15
Kamloops	12	17
100 Mile House	12	13
Stikine	11	26
	13,990	**34,407**

Figure 22 Diabetes Dashboard - Patients/Admissions Drilldown

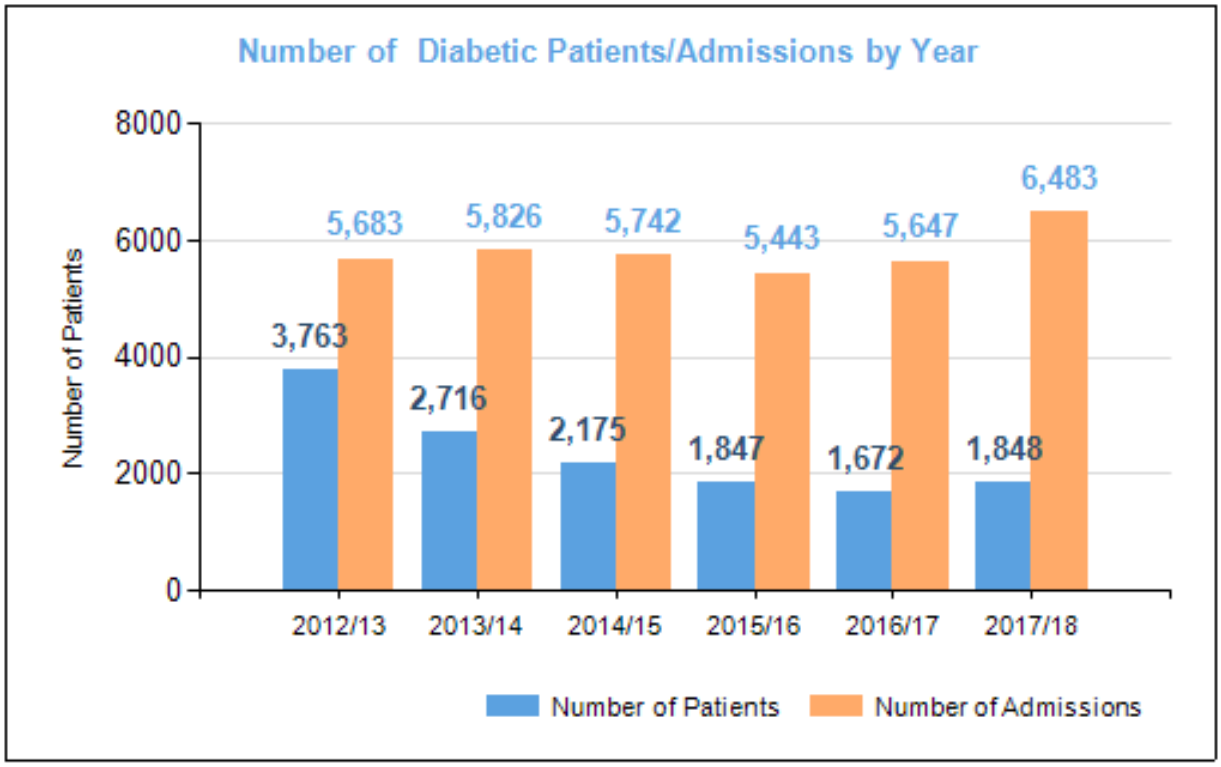


Figure 23 Diabetes Dashboard - Patients/Admissions by Year

Figure 23 shows the number of patients and admissions for each year from 2012/13 to 2017/18. The maximum number of patients were recorded for the year 2012/13 (3,763) and the minimum was in 2016/17 (1,672), also the patients were consistently decreasing till 2017/18 followed by a slight increase in 2017/18 (1,848). This trend is consistent with Statistics Canada numbers. In 2012/13, 5.7% of British Columbia residents were diagnosed with diabetes which was lower than national average of 6.5%. In the following years (2013/14, 2014/15) this number dropped to 5.5% compared with the national average of 6.6% (2013/14) and 6.5% (2014/15). In 2017/18, the national average went up to 7.3% and British Columbia recorded a corresponding increase to 5.9%. On the other hand, the admissions trend is not consistent with the trend observed for number of patients. In 2015/16, the number of admissions (5,443) was lowest and 2017/18 recorded

highest number of admissions (6,483) even though the number of patients was almost identical. It should be noted that the admissions include patients from previous years and also readmissions of the same patient. For instance, the year 2017/18 recorded 6,483 admissions for the cumulative number of patients (14,021) and not the new patients (1,848) only.

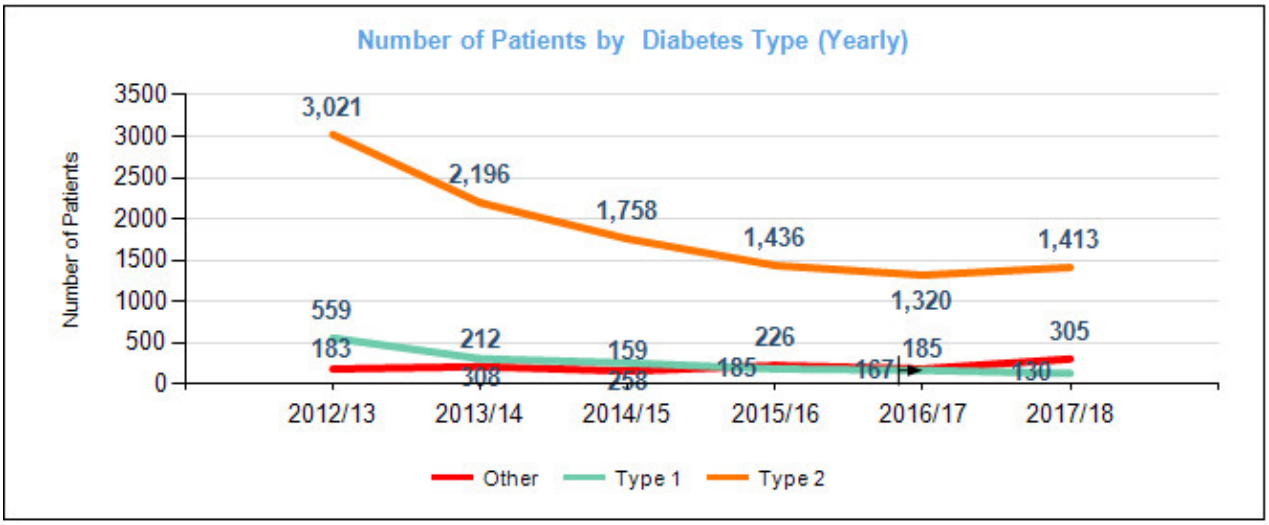


Figure 24 Diabetes Dashboard – Patients by Diabetes Type (Yearly)

Figure 24 shows the number of patients by diabetes types (T1D, T2D and Other). The number of patients consistently decrease for T2D until 2016/17 (1,320) from 2012/13 (3,021) and then increases slightly in 2017/18 (1,413). A similar pattern was observed for T1D patients. These observations are consistent with Figure 23 which showed an increase of patients in 2017/18. For other types of diabetic patients, a different trend was observed which recorded the lowest number of patients in 2017/18 (130) and the highest number in 2013/14 (308). Since this group represents only 8% of the total number of patients, the impact on the overall trend is relatively insignificant.

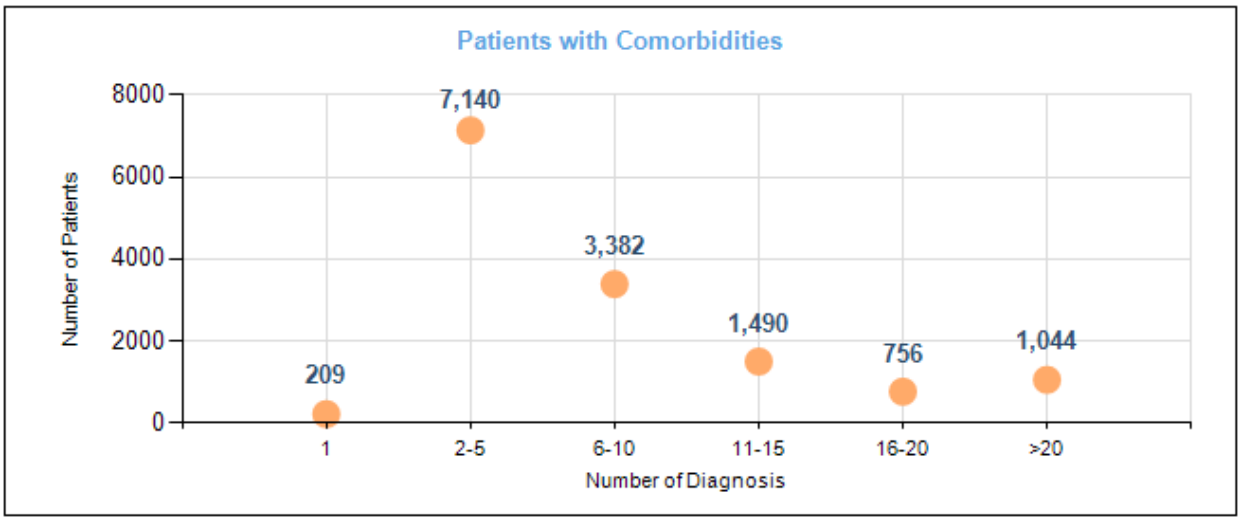


Figure 25 Patients with Comorbidities

Figure 25 shows the comorbidities per patient. The average number of diagnosis for a patient was observed as four with the majority of patients (7,140) having two to five comorbidities. There were fewer patients with higher number of comorbidities. The lowest number of patients (756) was recorded for 16-20 comorbidities and then an increase was observed for 20+ comorbidities. On further breakdown for patients with greater than twenty comorbidities, it was observed that 990 patients recorded 20-40 comorbidities, thirty-three patients recorded 50-70 comorbidities, seventeen patients had 60-80 comorbidities and only four patients recorded over 80 comorbidities. As mentioned earlier, all patients in the NH dataset had at least one type of diabetes (T1D, T2D or other). The 209 patients shown in Figure 25 are those who had only one diabetes diagnosis code recorded.

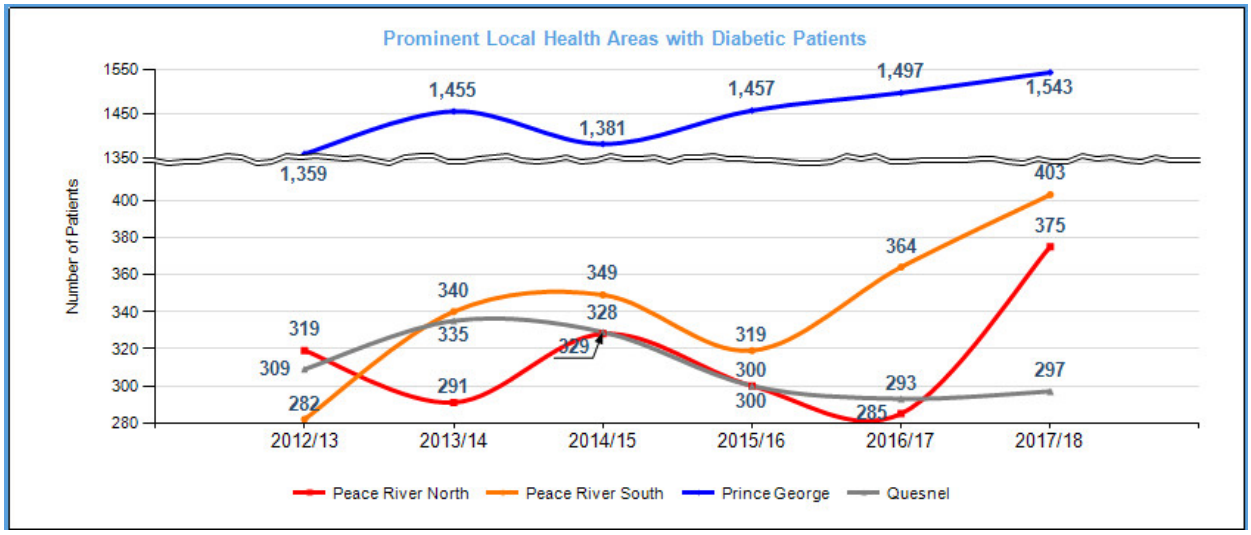


Figure 26 Diabetes Dashboard – Prominent LHAs with Diabetic Patients

Figure 26 shows the prominent communities which had the highest number of diabetic patients. The dataset consisted of 305 communities and seventy LHAs of which Prince George recorded the maximum number of patients consistently over the years. The University Hospital of Northern British Columbia in Prince George accounted for 53% of the total patients and 47% of overall admissions. The GR Baker Memorial Hospital in Quesnel accounted for 7% of the total patients and 13% of the overall admissions. It was also observed that all communities showed an increase of patients in the year 2017/18 from the previous year making it consistent with the trends noted earlier (Figure 23 and Figure 24). Since Prince George and Quesnel are categorized as both LHA as well as communities, these names will refer to one or the other depending on the context. Peace River South and Peace River North consists of thirteen and sixteen communities, respectively. Quesnel has three communities and Prince George has fifteen communities

including itself. It is also to be noted that the LHAs are specific to the patient and not to the facilities. For instance, a patient can have their community recorded as Quesnel and still be admitted to a facility in Prince George. The top ten LHAs with the maximum number of patients were Prince George, Peace River South, Quesnel, Peace River North, Terrace, Nechako, Prince Rupert, Kitimat, Smithers and Burns Lake.

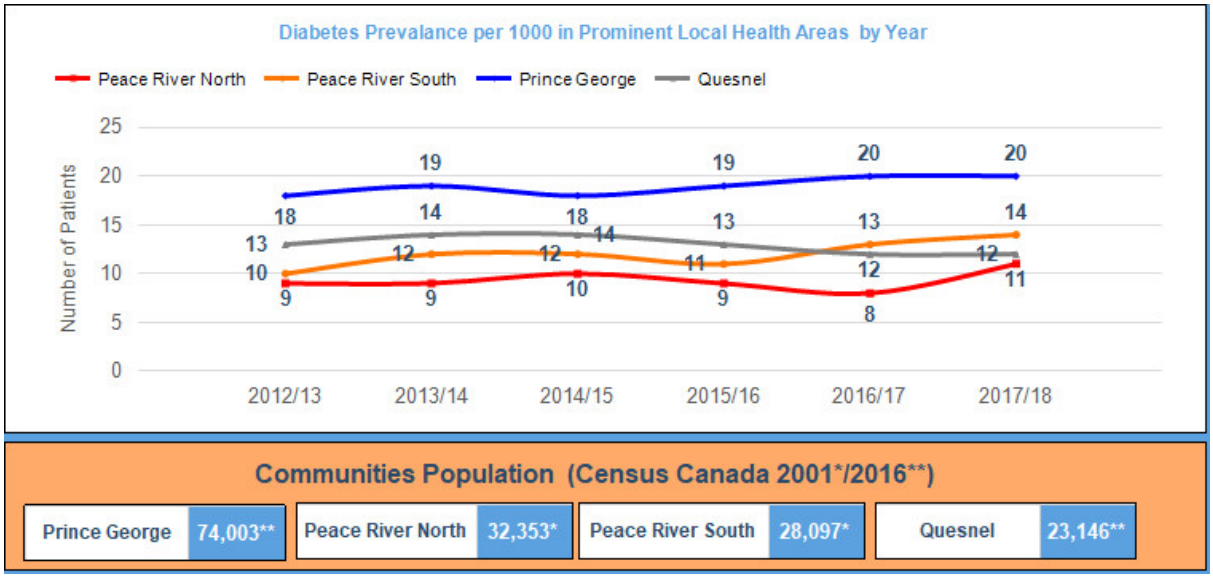


Figure 27 Diabetes Dashboard - Prevalence of Diabetes by LHAs

Figure 27 shows prevalence of diabetes per thousand of the population. The population figures were obtained from Census Canada (Peace River North and Peace River South - 2001; Prince George and Quesnel - 2016).

Similar to

Figure 26, Prince George recorded the maximum prevalence per thousand residents over the study period. An interesting observation is that while Prince George and Quesnel did

not show any change between 2016/17 and 2017/18, both Peace River North and Peace River South showed a slight increase over the same period. This is consistent with Figure 26 where a spike in the number of patients was observed for both Peace River North (32%) and Peace River (11%) South during this period.

4.1.1 Diabetes Types and Comorbidities

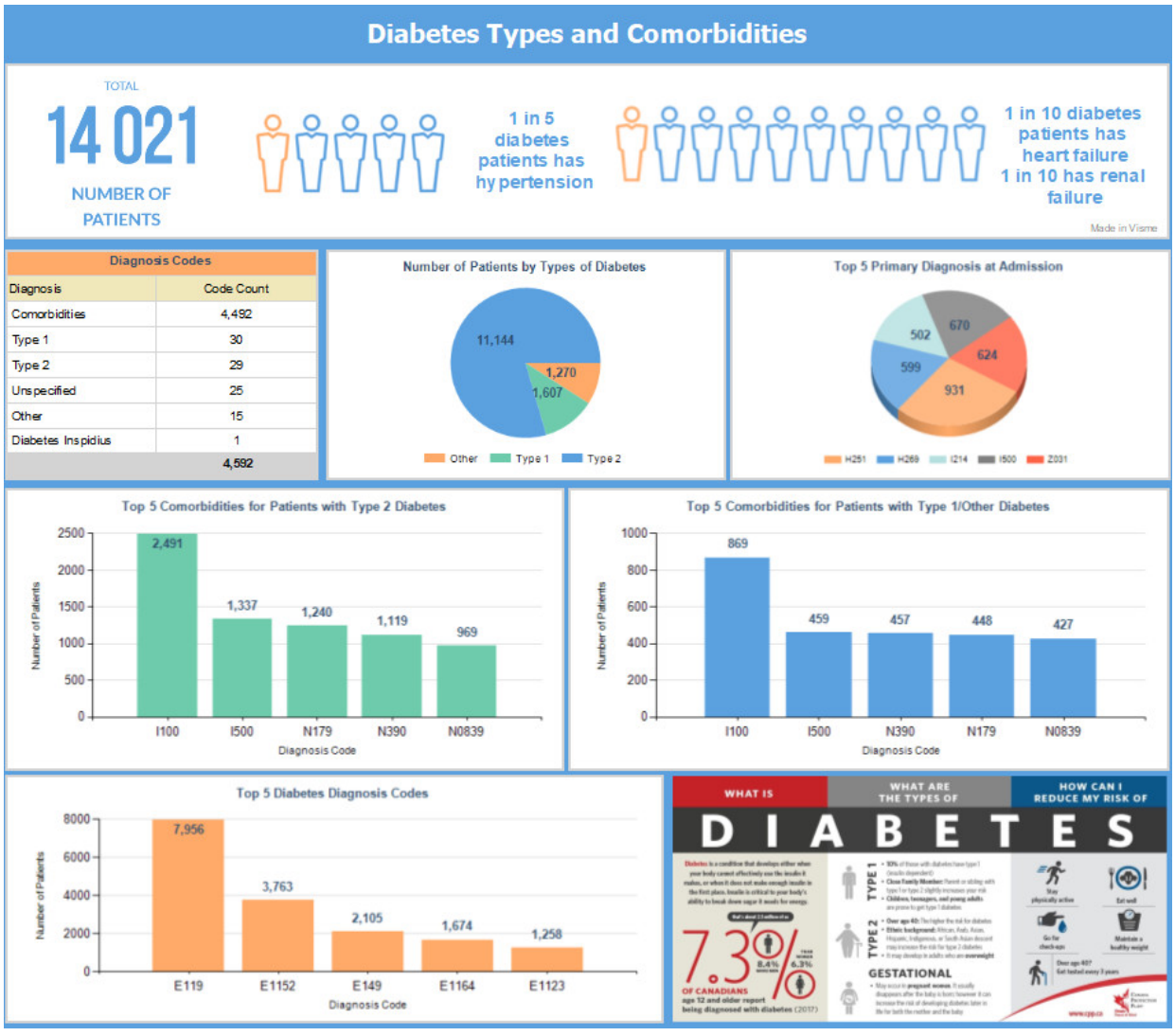


Figure 28 Diabetes Types/Comorbidities Dashboard

Figure 28 shows the overall aggregated statistics broken down by diagnosis codes specific to the types of diabetes and comorbidities. Using charts and tables, the clinical data from NH has been sliced along various patient groups (T1D, T2D and other types of diabetes) and diagnosis. Each of these charts is explained below along with the drilldowns, where applicable. The image on the bottom right has been taken from Diabetes Canada [8].

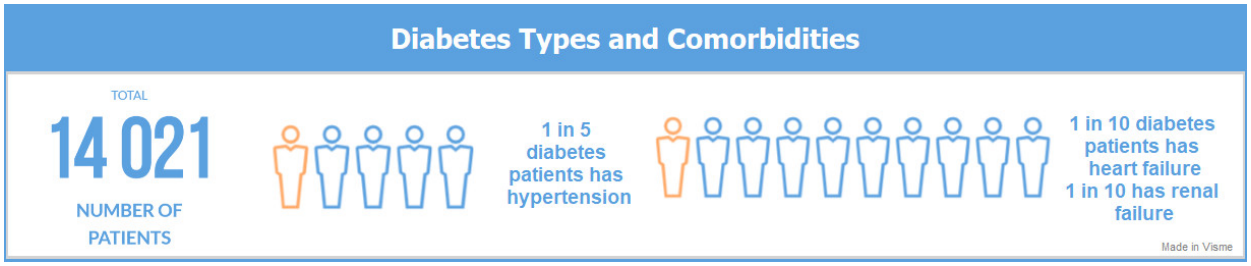


Figure 29 Diabetes Types/Comorbidities Dashboard Statistics

Figure 29 shows vital statistics related to comorbidities. Out of 14,021 patients, it was observed that one in five had hypertension and one in ten had heart/renal failure. These three comorbidities accounted for 39% of the total patients and 22% of the total admissions. This observation also became the basis of selection of the three target variables identified in chapter 3.

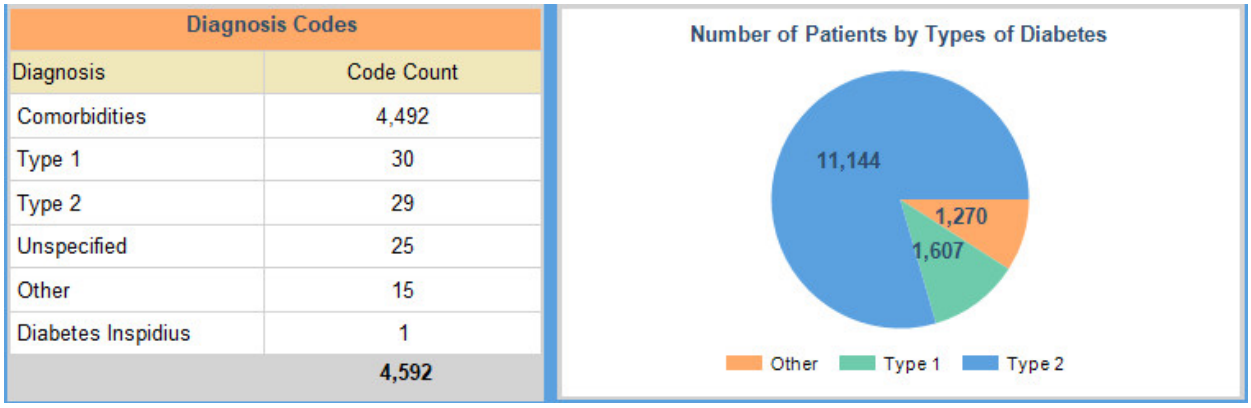
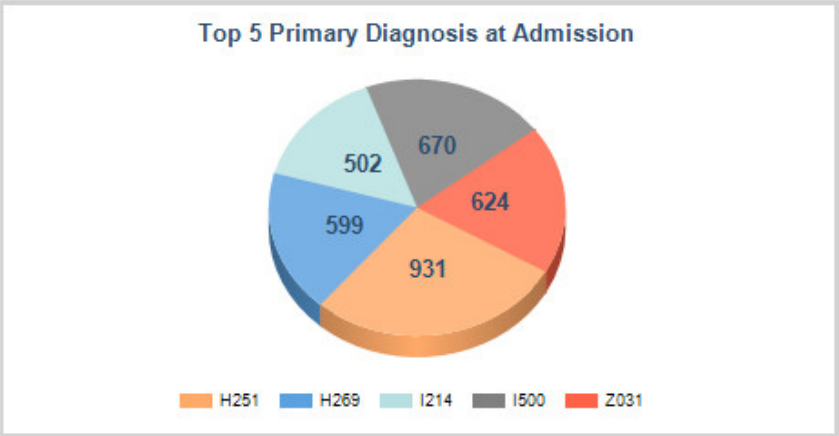


Figure 30 Diabetes Types/Comorbidities Dashboard - Diagnosis Codes/ Diabetes Types

The diagnosis codes table in Figure 30 are grouped by different types of diabetes and other comorbidities. It was observed that comorbidities accounted for 98% of the total diagnosis codes.

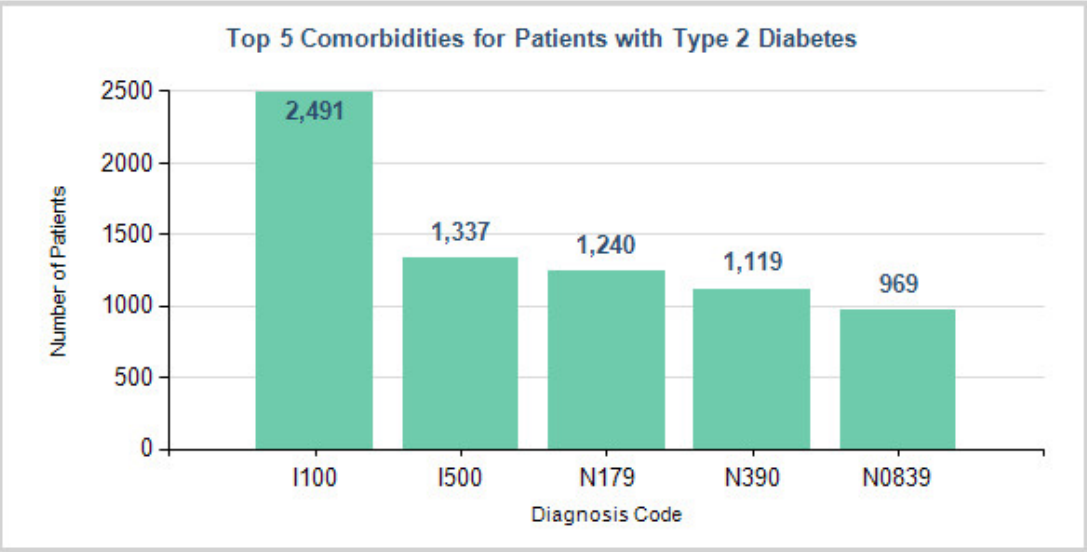
Figure 30 also shows the number of patients with different types of diabetes. 'Other' type of diabetes includes diabetes insipidus and unspecified diabetes types. It is observed that 80% of the total patients had T2D and the remaining 20% had T1D or Other types of diabetes. The comorbidities specific to diabetes types are shown in Figure 32 and Figure 33.



Diagnosis Code	Comorbidities Description
H251	Senile nuclear cataract
H269	Cataract, unspecified
I214	Acute subendocardial myocardial infarction
I500	Congestive heart failure
Z031	Observation for suspected malignant neoplasm

Figure 31 Diabetes Types/Comorbidities Dashboard - Diagnosis Codes/ Diabetes Types

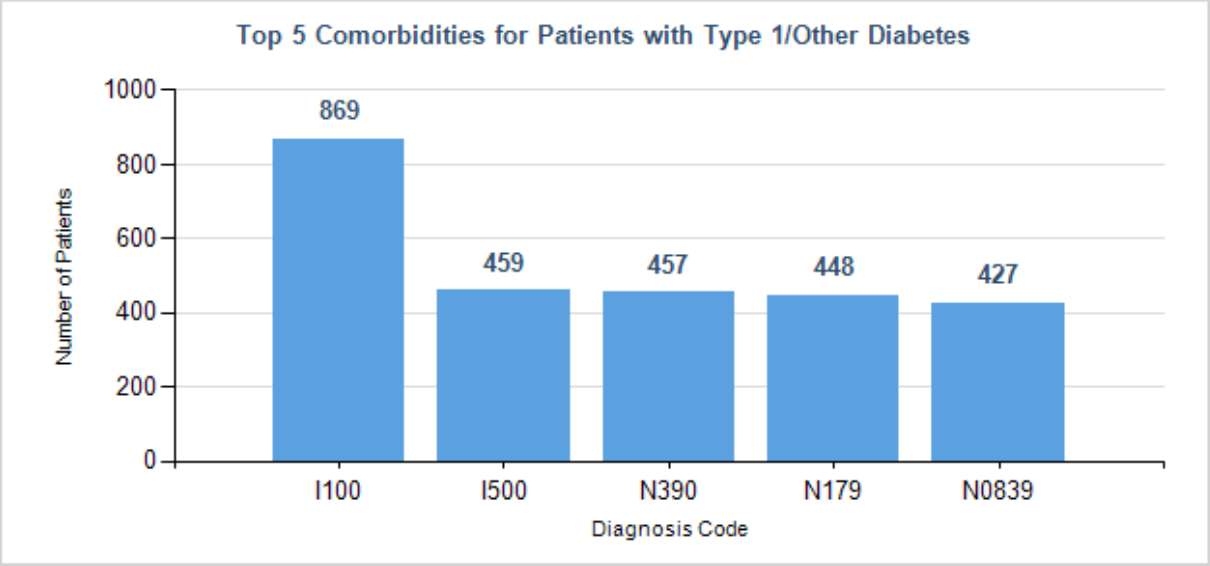
Upon admission, multiple diagnosis codes are normally entered, one of which becomes the primary 'most responsible' code. Figure 31 shows the top five primary diagnosis codes which account for 23% of total patients and 15% of total admissions. In this figure, while H251 is showing the maximum number of patients' primary diagnosis, it is not the case when all diagnosis types are included. For example, H251 accounted only for 4% of the total admissions and 7% of the total patients. Thus, it was not identified as a target variable when building the model. It is observed that when H251 and I500 were included in the diagnosis set for the patient, they were recorded as primary diagnosis in 98% and 48% of the cases, respectively. The description for the diagnosis codes is shown in the table in Figure 31.



Diagnosis Code	Comorbidities Description
I100	Benign hypertension
I500	Congestive heart failure
N0839	Unspecified glomerular disorders in diabetes mellitus
N179	Acute renal failure, unspecified
N390	Urinary tract infection, site not specified

Figure 32 Diabetes Comorbidities Dashboard- T2D Comorbidities

Figure 32 shows the top five comorbidities for patients with T2D together with their corresponding description. It was observed that 65% of the patients with T2D were diagnosed with one or more of these comorbidities, 48% were diagnosed with one or more of the top three comorbidities (I100, I500, N179). These three comorbidities were selected as target variables for the predictive model.

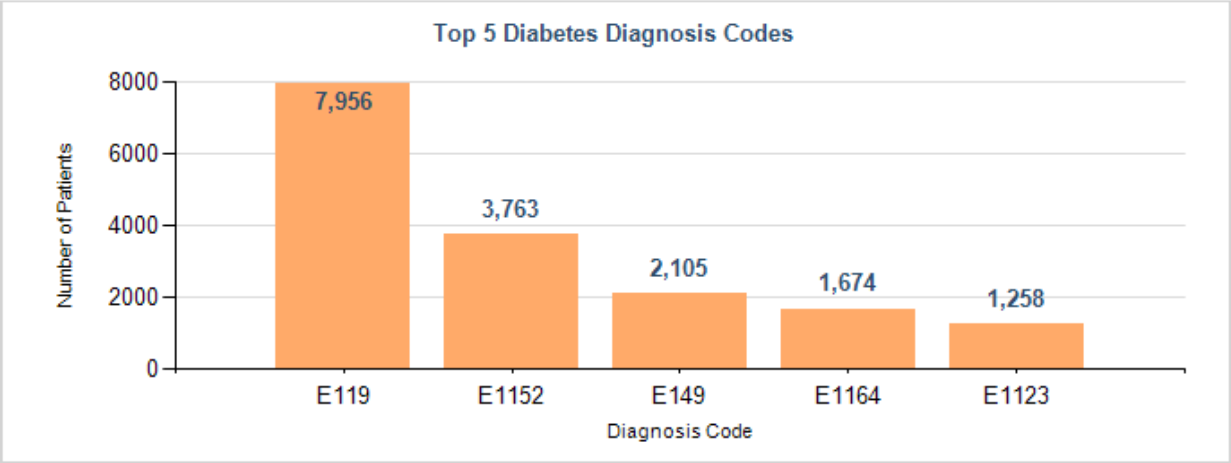


Diagnosis Code	Comorbidities Description
I100	Benign hypertension
I500	Congestive heart failure
N0839	Unspecified glomerular disorders in diabetes mellitus
N179	Acute renal failure, unspecified
N390	Urinary tract infection, site not specified

Figure 33 Diabetes Comorbidities Dashboard- T1D/Other Diabetes Comorbidities

Figure 33 shows the top five comorbidities diagnosed for patients with T1D or any other types of diabetes excluding T2D. These comorbidities represented 95% of the total patients in this group. The three target variables (I100, I500, N179) selected for building the models accounted for 63% of patients. The top two comorbidities (hypertension and congestive heart failure) are the same in both sets (Figure 32 and Figure 33). However, the third and fourth comorbidities (N390 and N179) are reversed in the two sets. N179 was selected as the target variable because of its high cumulative impact. Figure 34

shows the top five diagnosis codes embedded with different types of diabetes. Four of these codes (starting with ‘E11’) represent T2D patients which can be attributed to the fact that majority of the patients in this dataset have been diagnosed with T2D.



Diagnosis Code	Comorbidities Description
E1123	Type 2 diabetes mellitus with established or advanced kidney disease
E1152	Type 2 diabetes mellitus with certain circulatory complications
E1164	Type 2 diabetes mellitus with poor control, so described
E119	Type 2 diabetes mellitus without (mention of) complications
E149	Unspecified diabetes mellitus without (mention of) complication

Figure 34 Comorbidities Dashboard- Diabetes Specific Diagnosis Codes

Top 20 Diabetes Diagnosis	
Code Description	**Number of Patients**
Type 2 diabetes mellitus without (mention of) complications	7,956
Type 2 diabetes mellitus with certain circulatory complications	3,763
Unspecified diabetes mellitus without (mention of) complication	2,105
Type 2 diabetes mellitus with poor control, so described	1,674
Type 2 diabetes mellitus with established or advanced kidney disease	1,258
Type 2 diabetes mellitus with other specified kidney complication not elsewhere classified	1,046
Type 2 diabetes mellitus with other specified ophthalmic complication not elsewhere classified	969
Type 1 diabetes mellitus without (mention of) complication	809
Type 2 diabetes mellitus with multiple other complications	658
Unspecified diabetes mellitus with certain circulatory complications	476
Type 1 diabetes mellitus with poor control, so described	349
Type 2 diabetes mellitus with other specified complication, not elsewhere classified	303
Type 1 diabetes mellitus with ketoacidosis	296
Type 2 diabetes mellitus with foot ulcer (angiopathic)(neuropathic)	285
Type 2 diabetes mellitus with hypoglycaemia	269
Type 1 diabetes mellitus with certain circulatory complications	266
Type 2 diabetes mellitus with ketoacidosis	205
Unspecified diabetes mellitus with other specified ophthalmic complication not elsewhere classified	200
Type 2 diabetes mellitus with foot ulcer (angiopathic) (neuropathic) with gangrene	199
Unspecified diabetes mellitus with poor control, so described	187

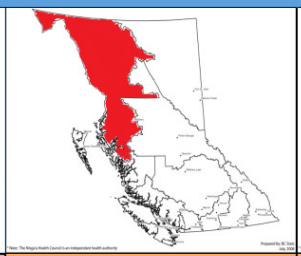
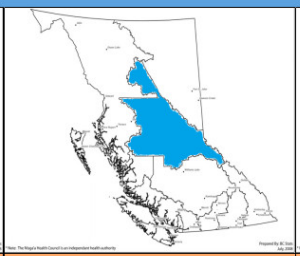
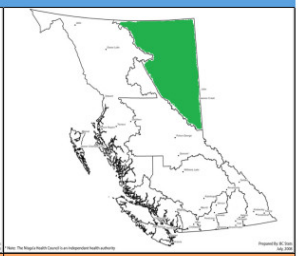
Figure 35 Diabetes Types/Comorbidities Dashboard- Diabetes Diagnosis Codes Drilldown

Cumulatively, the total number of patients represented by these codes exceed 14,021 patients because the same patient can be diagnosed with multiple codes. This issue was not obvious in earlier charts because the patients were either filtered by diabetes types

or by primary admissions. Figure 35 shows the drilldown report which lists the top twenty diabetes specific diagnosis for all patients.

4.1.2 HSDA Comparison

HSDA Comparison Report (2012/13-2017/18)

HSDA	Northwest	Northern Interior	Northeast
# of Patients	3,589	8,497	2,761
Total Patient Visits	8,670	20,254	5,900
Average # of Visits Annually	1,445	3,375	983
Total Average LOS (Days)	7	7	8
Average Acute Care LOS (Days)	5	6	5
ALC LOS (Days)	1	1	3
Total Hospital Day Stays	58,745	149,989	49,777
% of Patients with Family Physician	78%	80%	77%
% of Acute Care Patients	50% 58% 66% 74% 82% 90%	50% 58% 66% 74% 82% 90%	50% 58% 66% 74% 82% 90%
% of Out-of-Province Patients	1%	0.5%	4%
Patients with 2-5 Commorbidities	1903	4224	1620
Patients with 6+ Commorbiditites	1627	4186	1065
% of Type 2 Diabetes Patients	2962	7623	1990
# of Facilities Visited	7	7	4

Figure 36 Diabetes HSDA Dashboard

Figure 36 shows a comparison of aggregated statistics for each of the three HSDAs – Northwest (NW), Northern Interior (NI), Northeast (NE) - which recorded 24%, 57%, and 19% of the total patients, respectively. An interesting observation was that 6% of the

patients migrated to other communities and were thus counted more than once. This, however, does not impact the number of visits because those are recorded independent of the patient's community. On average, approximately two admissions per patient were recorded across all HSDAs, including patients from outside of BC. Even though NI recorded majority of the patients as well as admissions, the average length of stay (LOS) was very similar across all HSDAs. A similar pattern was also observed for patients who had family physicians. In BC, there were a total of nineteen communities which recorded over 100 patients for the years 2012/13 to 2017/18. Among these, Fort Nelson had 85% of patients without a family doctor followed by Fort St. James, Houston, Queen Charlotte and Burns Lake (73%, 42%, 40%, 33%). The five communities with the highest number of patients (Prince George, Quesnel, Fort St. John, Terrace and Dawson Creek) had 14%, 27%, 18%,14% and 11% patients with no family doctors, respectively. NE had the highest number of patients visiting from outside of BC. The facilities visited most by these patients were Dawson Creek District Hospital (69 patients), Fort St. John General Hospital (37 patients) and University Hospital of Northern British Columbia (28 patients). All of these patients were from Alberta.

The number of patients who were only recorded with only one diagnosis code was less than 2% in each of the HSDAs. Patients with two to five comorbidities represented 53%, 50% and 59% of the total number of patients in NW, NI and NE, respectively. Patients with six or more comorbidities were 45%, 49% and 39% for the same HSDAs, respectively.

While 85% of total patients had a diagnosis code related to T2D, the three HSDAs had a variation ranging from 72% (NE) to 90% (NI); NW was closer to the overall average (82%).

Of the eighteen facilities across all HSDAs, UHNBC admitted 50% of the total patients followed by Mills Memorial (11%) Hospital and Fort St. John General Hospital (10%). The lowest number of patients was admitted by McBride & District Hospital (0.5%).

Figure 37 shows the annual breakdown of cumulative visits and number of patients across all HSDAs. This drilldown is obtained by clicking on one of the HSDA maps in Figure 36.

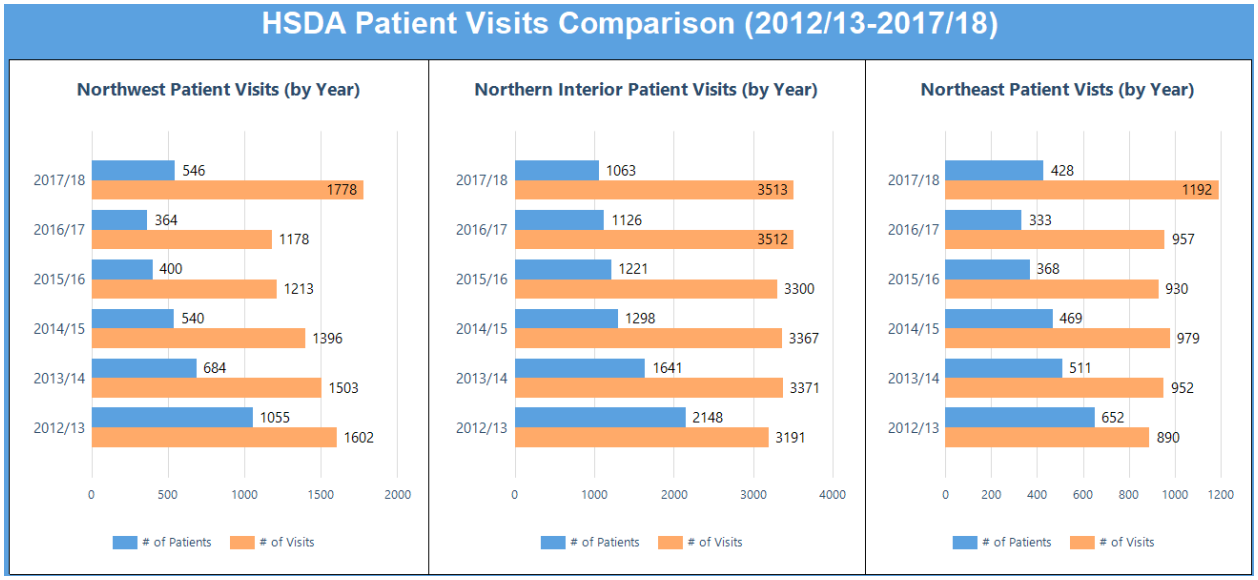


Figure 37 HSDA Dashboard - Patients/Visits Drilldown

4.1.3 Summary

An interactive diabetes dashboard was developed using dataset consisting of diabetic patients who had accessed Northern Health facilities from the years 2012 to 2018. The dashboard consisted of three main reports: 1) the diabetes dashboard which contained overall aggregated statistic of this dataset, 2) the diabetes types and comorbidities dashboard where the data was grouped by different types of diabetes and comorbidities of the patients, and 3) the HSDA dashboard which grouped the data by three HSDAs – NE, NW and NI.

The following are a few observations which were made from these reports:

- 80% patients were diagnosed with T2D
- Average age of patients was found to be sixty-three
- Average Number of diagnosis per patient was four
- Number of new patients were consistently decreasing till 2016/17 with a slight increase in 2017/18
- HSDA Northern Interior recorded 57% of the total number of patients where LHA Prince George had maximum number of Patients and Admissions.
- All three target variables (I100, I500, N179) were recorded as one of the top five comorbidities for T2D patients (excluding the diabetes diagnosis codes)

This dashboard also had drilldown capabilities to view reports at finer granularity by various parameters such as HSDA, LHA and patient comorbidities.

4.2 Predictive Modeling

Predictive modeling is the process of applying data mining algorithms on historical data to predict the likelihood of future outcomes. For the three target variables (I100, I500, N179), predictive models were built using six data mining algorithms. The corresponding results are explained in this section.

The six data mining algorithms chosen for building the predictive models were:

1. Bayesian Network
2. Neural Network
3. Random Forest
4. Logistic Regression
5. CHAID
6. Ensemble

4.2.1 Training Models

For each of the three target variables, the dataset was split such that the training component contained 70% of the patients diagnosed with the corresponding target variable. The remaining 30% was then used for testing. This resulted in a patient distribution as shown in Table 6. For instance, N179 had a total of 1,303 patients who were split into training (913) and testing (390) datasets, respectively. This number (1,303) represents 9.3% of the total number of patients. In order to maintain this 70:30 ratio, the desired number of patients in the dataset was then determined which in this case was

8,274 (59% of the total patient population). The remainder was used for testing. This method was consistently applied to all target variables. A higher percentage of records in the training dataset allowed the models to learn the underlying patterns better, which helped in making better predictions.

Table 6 Training/Testing Datasets

Target Variable	Training	Testing	Total Patients
I100	66%	34%	18.9% (2,656)
I500	65%	35%	9.9% (1,385)
N179	59%	41%	9.3% (1,303)

Figure 38 shows the training model for prediction of I100 (hypertension) using five data mining algorithms with different nodes. Each of these nodes is explained below:

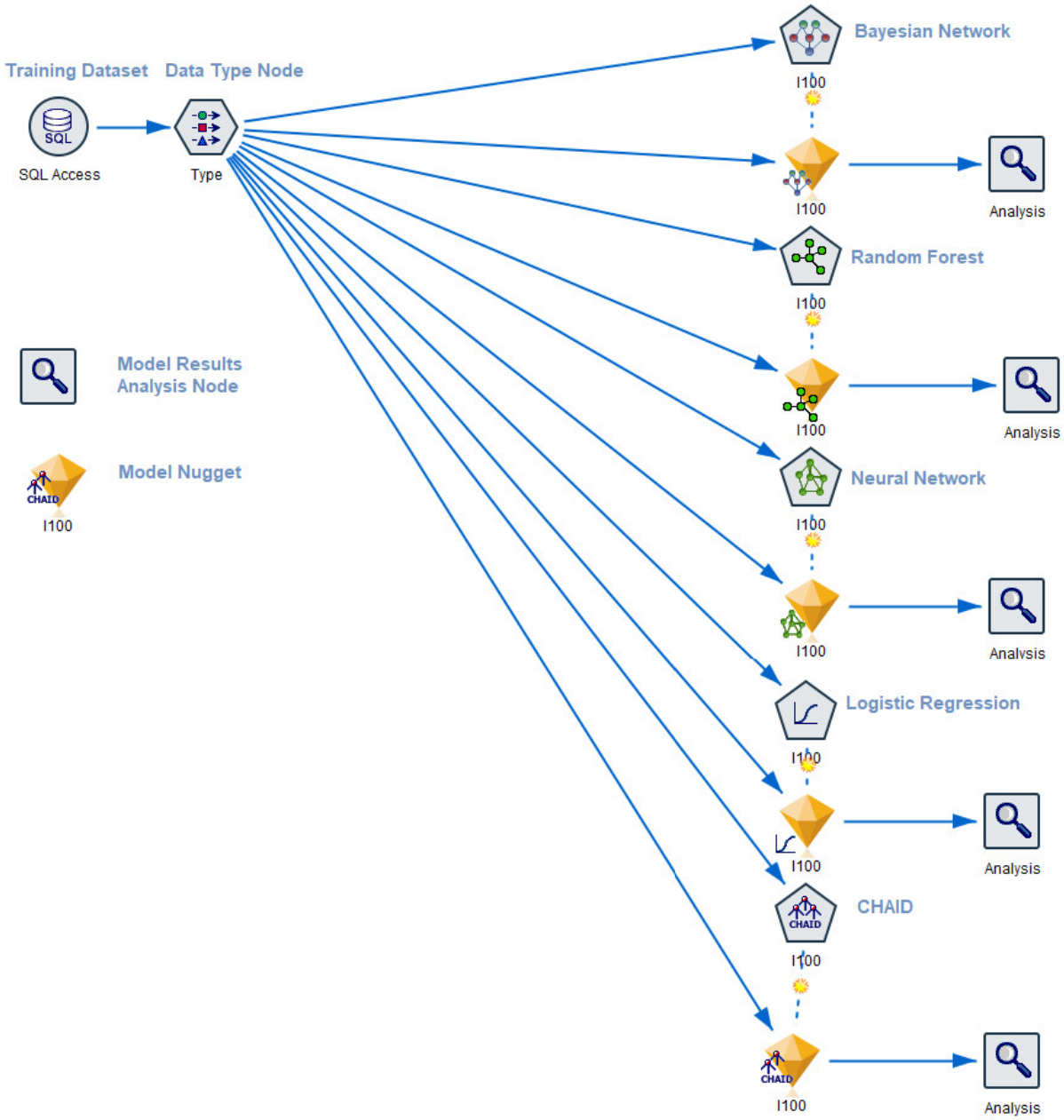


Figure 38 Predictive Modeling Training

SQL Access: This is the data source node which establishes a connection to diabetes database and extracts the dataset consisting of the finalized twenty-six variables for building the predictive models. Twenty-four of these variables were the input variables and the remaining two variables were excluded because they were either a unique identifier (Patient Code) or the target variable (I100).

Type: The Type node is used to specify the data type of the selected variables as either nominal, categorical, continuous, flag or ordinal. This node allows to specify whether a variable is input or target. Additionally, it also gives an option to specify one variable as the unique identifier (Patient Code).

The twenty diagnosis codes and Physician Code were all assigned as a *flag* including the target variable. The *flag* datatype is used for variables which have binary values, such as 0 or 1. Patient Code, Age, Average Length of Stay were assigned as *continuous* which is used to describe numeric values including decimals. Facility Health Service Delivery Area and Facility Name were assigned as *nominal* which is used for storing string values. For this predictive model, I100 was set as the target variable and the remaining variables were the input. Figure 39 shows the twenty-six variables with this information, where the Measurement column shows the data type, the Values column shows the sample values, the Missing column shows missing values in the dataset, the Check column specifies if a variable needs to be excluded, and the Role column specifies the variable as input, target or unique identifier.

Data Mining Model Node: The Type node is connected to the data mining model nodes each of which represent one of the five (Bayes Network, Neural Network, Random Forest, Logistic Regression, CHAID) algorithms. The Ensemble algorithm is not shown as it is explained later in this chapter. Executing these nodes generates the model nugget which contains the results of the trained model for the selected algorithm.

Field	Measurement	Values	Missing	Check	Role
Patient Code	Continuous	[100,9156]		None	Record ID
E119	Flag	1/0		None	Input
E1152	Flag	1/0		None	Input
I100	Flag	1/0		None	Input
E149	Flag	1/0		None	Input
E1164	Flag	1/0		None	Input
I500	Flag	1/0		None	Input
E1123	Flag	1/0		None	Input
N179	Flag	1/0		None	Target
N390	Flag	1/0		None	Input
N0839	Flag	1/0		None	Input
E1138	Flag	1/0		None	Input
H251	Flag	1/0		None	Input
E1128	Flag	1/0		None	Input
Z22300	Flag	1/0		None	Input
J189	Flag	1/0		None	Input
E109	Flag	1/0		None	Input
Z22302	Flag	1/0		None	Input
U980	Flag	1/0		None	Input
Z515	Flag	1/0		None	Input
E1178	Flag	1/0		None	Input
Facility Health Ser...	Nominal	Northeast,"Northern Interi...		None	Input
Facility Name	Nominal	"Bulkley Valley District Hos...		None	Input
Age	Continuous	[6,105]		None	Input
Average Length of ...	Continuous	[1.0,595.0]		None	Input
Physician Code	Flag	"No Family Doctor"/"Has F...		None	Input

Figure 39 Predictive Modeling Training - Type Node

Analysis Node: The results from the model nugget are connected to the analysis node which analyzes the prediction accuracy of the model. An example of analysis node for predicting I100 using neural network is shown in Figure 40 where approximately 81% of the total predictions were correct and 19% were wrong.

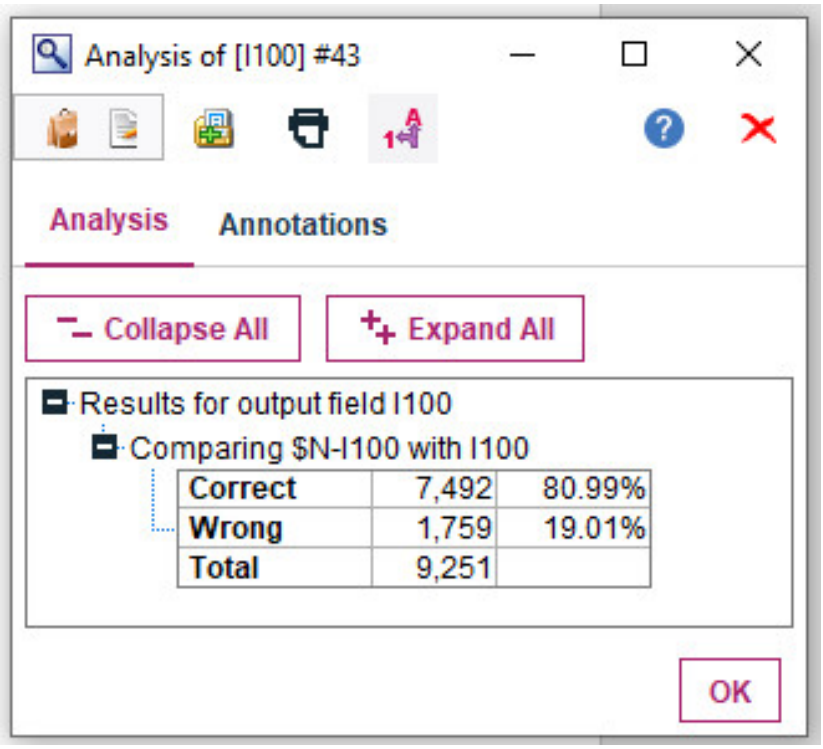


Figure 40 Predictive Modeling Training - Analysis Node

4.2.2 Testing Models

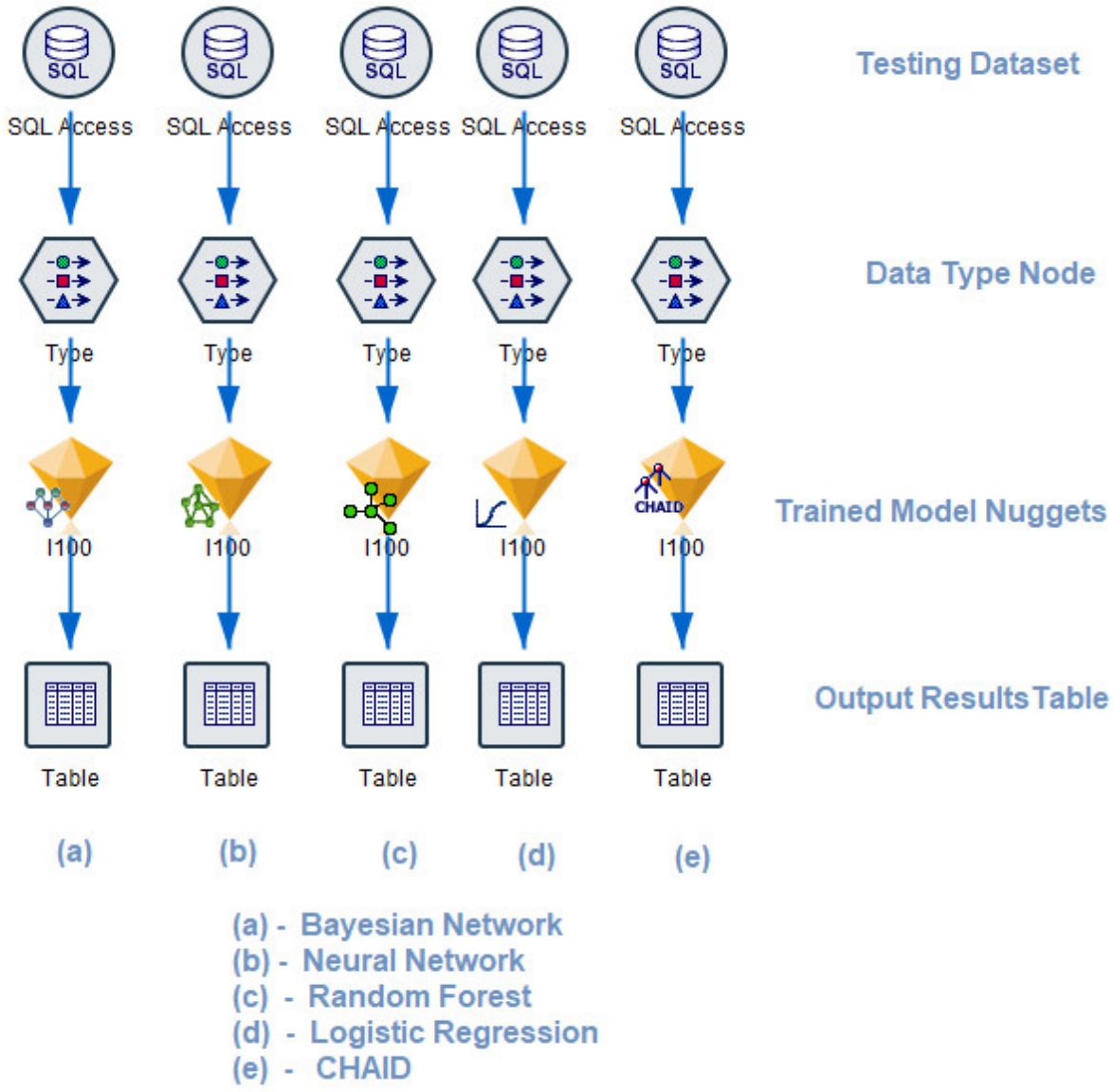


Figure 41 Predictive Modeling Testing

Figure 41 shows the testing model used for predicting one of the target variables (I100). The nodes shown are explained below:

SQL Access: The data sources represent the testing dataset with 30% of patients with I100. A major difference between the testing and training data source is that the former does not contain information of the corresponding target variable.

Type: The data type node used in testing is identical to the one used in training with the exception of target variable. It is necessary for the training and testing to have identical input variables with the specified data types for successful execution. An example of the type node used for testing is shown in Figure 42 where there is no target variable (I100) information being sent to the trained model nugget.

Trained Model Nuggets: These nuggets possess the required information to predict the target variable. Executing these trained models generate the results of one of the five corresponding data mining algorithms (Bayesian Network, Neural Network, Random Forest, Logistic Regression, CHAID). These results include the predicted values of the target variable (I100) which is pushed to an output table.

Output Table: This table contains the results of the executed training model nugget along with the other input variables. The predicted values of the five algorithms were evaluated for accuracy and are explained in the data analysis section.

Field	Measurement	Values	Missing	Check	Role
Patient Code	Continuous	[9351,14120]		None	Record ID
E119	Flag	1/0		None	Input
E1152	Flag	1/0		None	Input
E149	Flag	1/0		None	Input
E1164	Flag	1/0		None	Input
I500	Flag	1/0		None	Input
E1123	Flag	1/0		None	Input
N179	Flag	1/0		None	Input
N390	Flag	1/0		None	Input
N0839	Flag	1/0		None	Input
E1138	Flag	1/0		None	Input
H251	Flag	1/0		None	Input
E1128	Flag	1/0		None	Input
Z22300	Flag	1/0		None	Input
J189	Flag	1/0		None	Input
E109	Flag	1/0		None	Input
Z22302	Flag	1/0		None	Input
U980	Flag	1/0		None	Input
Z515	Flag	1/0		None	Input
E1178	Flag	1/0		None	Input
Facility Health Serv...	Nominal	Northeast,"...		None	Input
Facility Name	Nominal	"Bulkley Vall...		None	Input
Age	Continuous	[2,98]		None	Input
Average Length of ...	Continuous	[1.0,428.0]		None	Input
Physician Code	Flag	"No Family ...		None	Input

Figure 42 Predictive Modeling Testing - Type Node

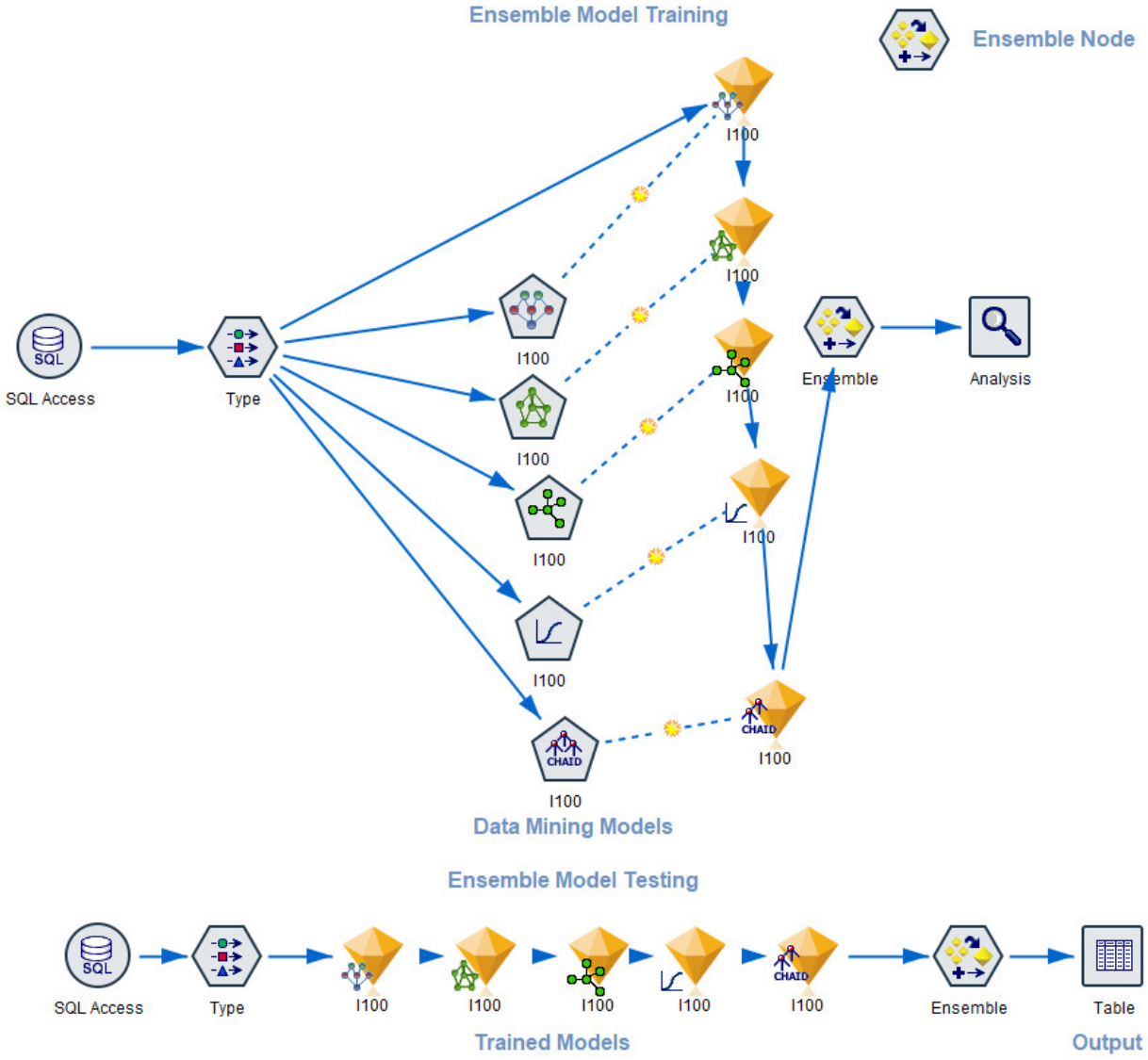


Figure 43 Predictive Modeling Ensemble Training/Testing

Figure 43 shows the training and testing models for the Ensemble algorithm. The SQL access node and the Type node are identical to the ones used in training (Figure 38) and testing (Figure 41), respectively. Ensemble training and testing is explained below:

4.2.3 Ensemble

Ensemble Model Training

The Ensemble node combines results of predictions for the target variable (I100) from the five trained models (Bayesian Network, Neural Network, Random Forest, Logistic Regression, CHAID) and generates a field containing the aggregated results. The Ensemble training results were observed by connecting the Ensemble node to the analysis node. It can be seen that the Type node is connected to only one model nugget (Bayesian Network). This is because the data types of the variables are fetched from the first model nugget (Bayesian Network) and then passed to the other four model nuggets followed by the Ensemble node. In Figure 38, the Type node was connected individually to the five data mining model nuggets, as each model fetched the data types of variables independently.

Ensemble Model Testing

The Ensemble model testing is very similar to that for the other five algorithms (Figure 41). The only difference is that instead of connecting individual model nuggets to the output table, the five model nuggets are connected to each other and then to the Ensemble node. This node is then connected to the Table node which generates the aggregated results. These results are evaluated for accuracy by comparing with existing data.

The process described above is also implemented for the other two target variables (I500, N179).

4.2.4 Analysis of Results

The results generated for the five base algorithms and Ensemble were evaluated for accuracy using the process described below:

The results from the output table for all testing models (Figure 41 and Figure 43) were pushed into the diabetes database. Since this table did not contain the target variable, it was added using a SQL query. The predicted column and the existing target variable information was compared for each row and the statistical accuracy of predictions was computed as follows:

$$Model\ Accuracy = \frac{Number\ of\ Accurate\ Predictions}{Total\ Number\ of\ Values\ in\ Dataset}$$

For instance, the number of accurate predictions for I100 (testing dataset) using Bayesian Network was 3,976. The total number of values in the testing dataset was 4,765 which gave an accuracy of 83.4%. Similarly, the accuracy was calculated for the remaining algorithms for all target variables (I100, I500, N179).

It was also observed that the accuracy of predictions for all algorithms was consistently better for true negative cases compared to true positives.

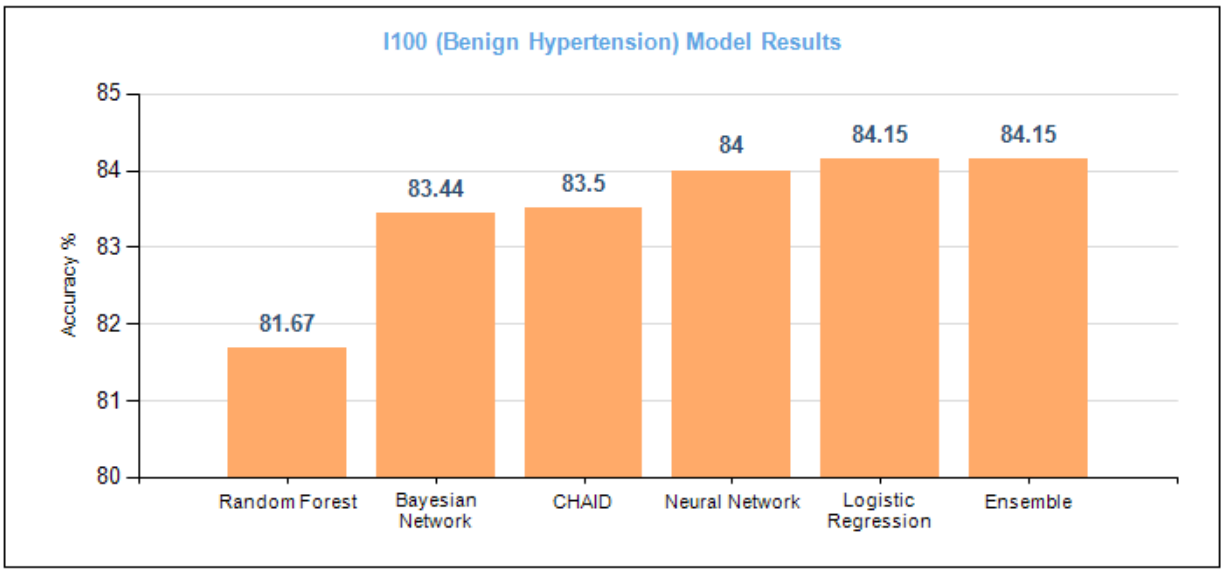


Figure 44 Predictive Modeling - I100 Results

Figure 44 shows the accuracy of the trained models for target variable I100 with a total of 4,765 patients. Ensemble and Logistic Regression had the highest accuracy for predicting patients with or without I100 (hypertension). Both these algorithms recorded identical accuracies of 84.15%. Bayesian Network, CHAID, Neural Network and Logistic regression made accurate predictions for 3,976, 3,979 and 4,003 patients, respectively, giving an accuracy as shown in Figure 44. The low accuracy (81.7%) of Random Forest can be attributed to overfitting problem which is one of the drawbacks of this algorithm. This dataset has 83% patients without hypertension and 17% (796) patients who were diagnosed with I100 (hypertension). For patients without hypertension, the six algorithms have an average accuracy of 97.2%. However, the average accuracy for those with

hypertension is only 15%. The reason for this low accuracy is the small number of patients in this group for the training dataset. Specifically, there were 2,656 patients with hypertension which is only 18.9% of the total patients (14,016). This is the reason for 70:30 split of the testing and training datasets of the total (2,656) patients diagnosed with target variable I100. A smaller number of patients in the training dataset would have resulted in an even lower accuracy. The chosen distribution also ensured that both training and testing datasets contained patients in proportion to the entire database.

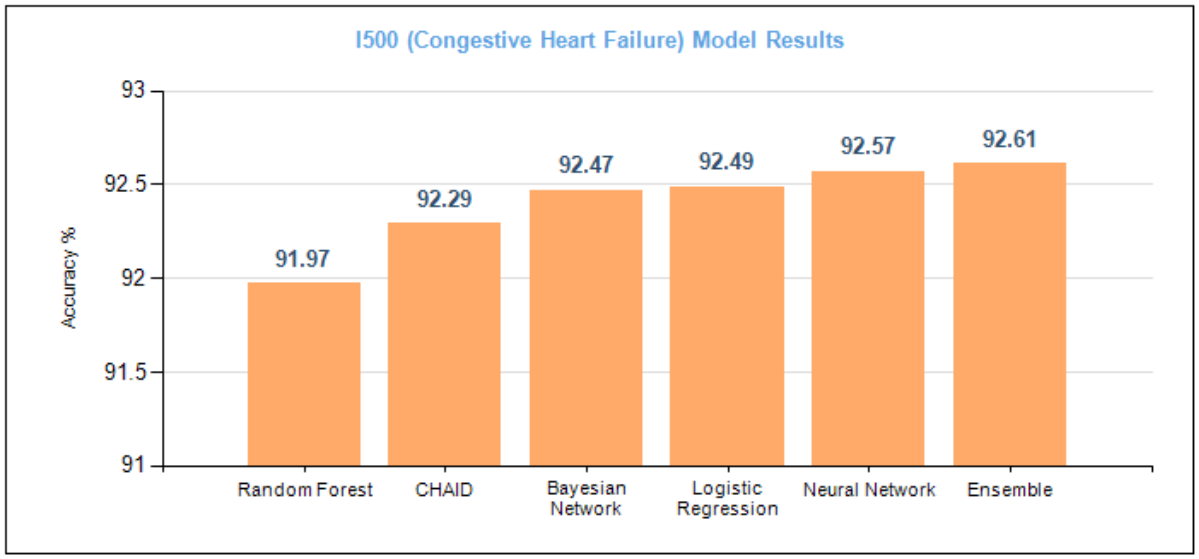


Figure 45 Predictive Modeling - I500 Results

Figure 45 shows the accuracy of the trained models for target variable I500 (Congestive Heart Failure) for a total of 4,959 patients. Ensemble recorded 92.61% accuracy followed by Neural Network with 92.57%. Logistic Regression, Bayesian Network and CHAID had

accuracies of 92.5%, 92.5%, and 92.3%, respectively. There were 415 (70%) and 970 (30%) patients diagnosed with I500 in the training and testing datasets, respectively. These datasets were used for all six algorithms. The average accuracy to predict patients with and without congestive heart failure was 29.1% and 98.7%, respectively. As explained earlier, the smaller number of patients in the training dataset for this group (patients diagnosed with I500) contributed to the low accuracy. It is to be noted that diagnosed I500 patients were only 9.8% of the total number of patients (14,016) in the entire database.

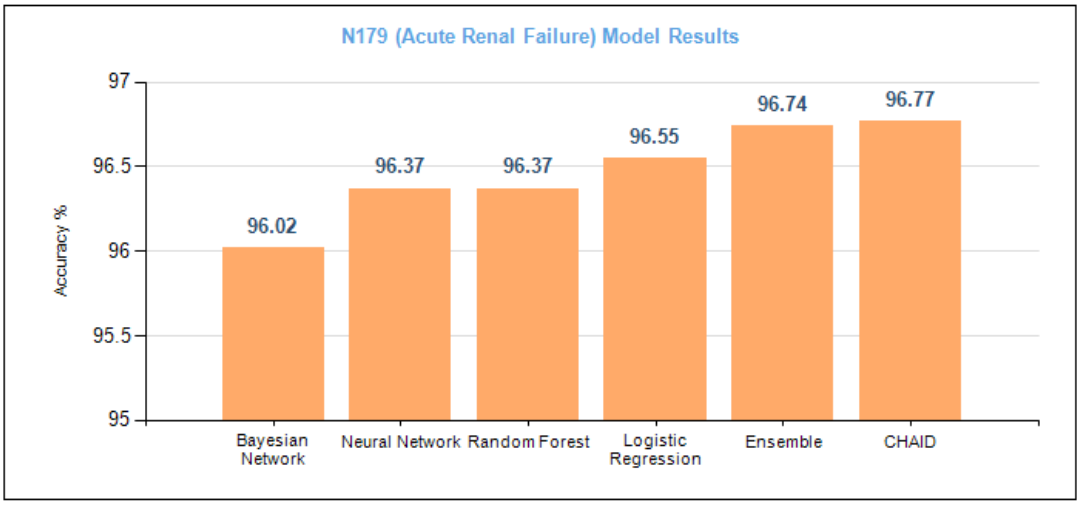


Figure 46 Predictive Modeling - N179 Results

Figure 46 shows the overall accuracy of all algorithms for target variable N179 (Acute Renal Failure). CHAID and Ensemble had accuracies of 96.77% and 96.74%, respectively followed by Logistic Regression with 96.55%. Random Forest and Neural Network recorded an identical accuracy of 96.37% and Bayesian Network had an accuracy of 96%. Within the database, there were a total of 1,303 patients who were

diagnosed with N179; these patients were split into training and testing datasets in the ratio of 70:30. The average accuracy for predicting patients with and without N179 is 63.7% (*Figure 47*) and 98.8%, respectively. It was observed that CHAID had the highest accuracy of 67.7% followed by Ensemble with 66.4%. Bayesian Network, Logistic Regression and Neural Network had accuracies of 63.8%, 63.6% and 61%, respectively. For reasons mentioned earlier, Random Forest had the lowest accuracy (59.7%) for predicting patients diagnosed with N179.

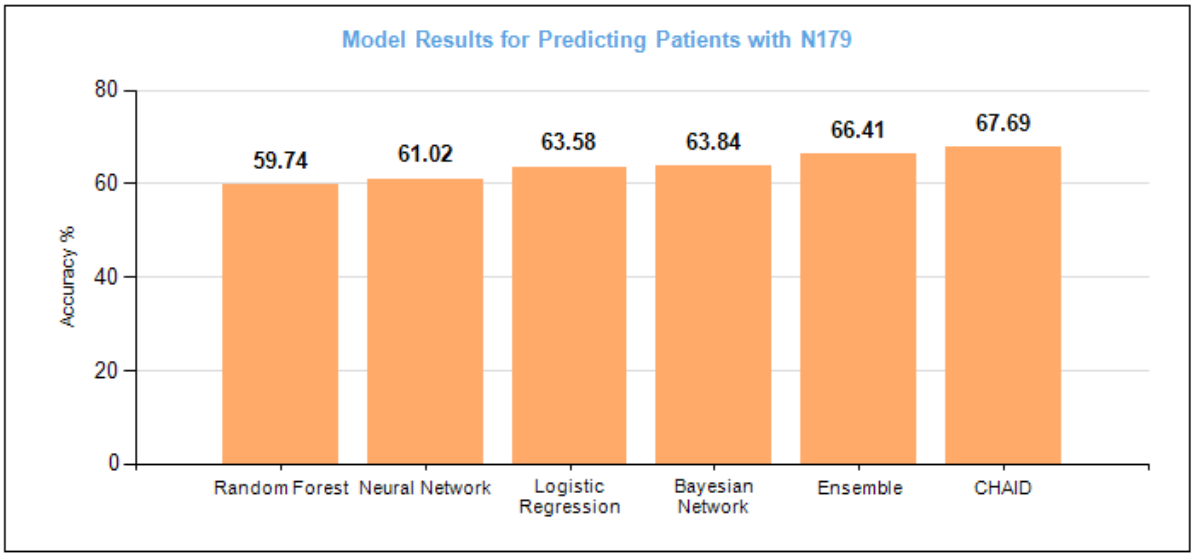


Figure 47 Predictive Modeling Accuracy for Patients with N179

It was observed that as the percentage of patients for a target variable decreased, there was an increase in accuracy of predicting true positives across all algorithms. Since, N179 had the lowest percentage of diagnosed patients, it had higher accuracy for predicting true positive cases.

4.2.5 Analysis of Variables

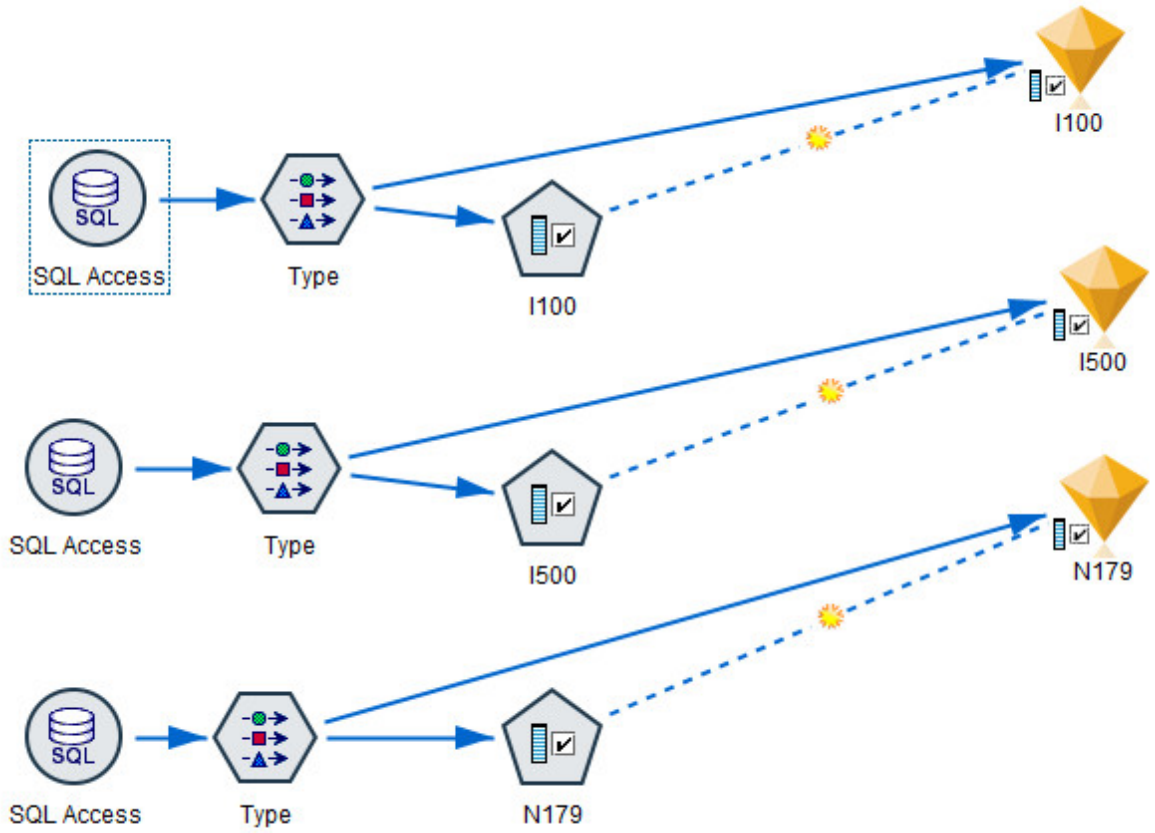


Figure 48 Predictive Modeling using Feature Selection (I100, I500, N179)

Figure 48 shows the predictive models using Feature Selection (FS) algorithm for the three target variables (I100, I500, N179) and their ranking is shown in Figure 49.

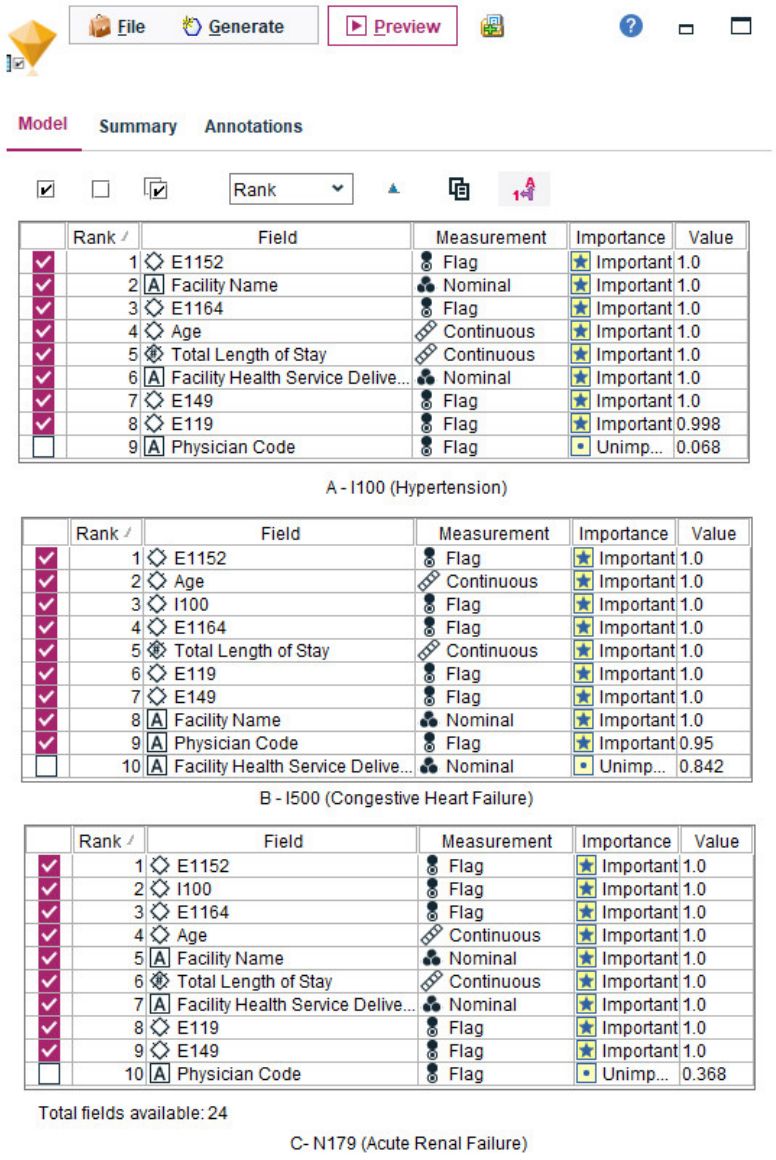


Figure 49 Feature Selection Results (I100, I500, N179)

The dataset used was identical with the exception of target variables. It can be observed that the diagnosis codes E1152, E1164, E119 and E149 were identified among the top ten variables consistently for all three target variables. This can be attributed to the large number of patients diagnosed with these codes (Table 7). It can be seen that I100 is also included as one of the top three important variables for predicting I500 as well as N179.

Codes E119 and E149 consistently rank outside the top five variables. These two codes specify diabetes patients without mention of complications which indicates that the probability of these patients to be diagnosed with other comorbidities is relatively low. 45% of patients diagnosed with E119 and 38% of patients diagnosed with E149 had no other comorbidities recorded in this dataset. In contrast, 16% of patients diagnosed with E1152 and 14% of patients diagnosed with E1164 had no other comorbidities recorded.

On further analysis, it was observed that 75% of E119 patients and 78% of E149 patients did not have either of I100, I500 and N179. Similarly, 49% of E1152 and 39% of E1164 patients were diagnosed with at least one or more of the three comorbidities (target variables) resulting in both of these codes to be ranked in the top five. An additional observation was that only two percent of the patients included all three target variables in their diagnosis. This is the reason N179 or I500 does not rank as important variables while predicting the others. The other variables such as Facility Name and Facility Health Service Delivery Area also show up as important because a majority of patients in this dataset were admitted to University Hospital of Northern British Columbia in HSDA 'Northern Interior'.

The average age of patients in the dataset was 63 years and the average Total Length of Stay was seven days. Both of these variables were ranked as important. Though Physician Code was listed as one of the top ten variables, there was no substantial relationship found by FS with any of the target variables. The other diagnosis codes were not listed as important as they all had a lower percentage of diagnosed patients.

Table 7 Top Seven Diagnosis Codes

Code	Diagnosis Description	Patients
E119	Type 2 diabetes mellitus without (mention of) complications	7,956
E1152	Type 2 diabetes mellitus with certain circulatory complications	3,763
I100	Benign hypertension	2,656
E149	Unspecified diabetes mellitus without (mention of) complication	2,105
E1164	Type 2 diabetes mellitus with poor control, so described	1,674
I500	Congestive heart failure	1,385
N179	Acute renal failure	1,303

4.2.6 Diabetes Comorbidities Assessment Tool

A physician-friendly, interactive web form has been built to predict the likelihood of a patient to be diagnosed with one of the three comorbidities (I100, I500, N179) in future. An example for predicting I100 (hypertension) using this tool is shown in

Figure 50 (input) and Figure 51 (output). The user input is given for all input variables excluding I100 which is the target variable. The Field column lists the input variables, the Storage column shows the data type, and Values column is where the user enters the input. It is to be noted that all string values need to be entered in double quotes and storage type is different from the data type of the variables which was explained earlier. Executing this web form runs the model in the background and generates output shown in Figure 51. The Ensemble algorithm is used in this case because it had the highest

accuracy for predicting I100 among all six algorithms. The web form can be connected to any of the six algorithms.

Field	Storage	Values
Patient Code	Integer	11219
E119	Integer	1
E1152	Integer	1
E149	Integer	0
E1164	Integer	1
I500	Integer	1
E1123	Integer	1
N179	Integer	1
N390	Integer	1
N0839	Integer	1
E1138	Integer	0
H251	Integer	0
E1128	Integer	1
Z22300	Integer	0
J189	Integer	0
E109	Integer	0
Z22302	Integer	1
U980	Integer	0
Z515	Integer	0
E1178	Integer	1
Facility Health Service Delivery...	String	"Northeast"
Facility Name	String	"Fort St. John General Hospital"
Age	Integer	68
Average Length of Stay	Real	9
Physician Code	String	"Has family doctor"

Figure 50 Diabetes Comorbidities Tool - User Input

Fields	Values
Patient Code	11219
E119	1
E1152	1
E149	1
E1164	1
I500 (Congestive Heart Failure)	1
E1123	1
N179 (Acute Renal Failure)	1
N390	1
N0839	1
E1128	1
Z22302	1
E1178	1
HSDA	Northeast
Facility Name	Fort St. John General Hospital
Age	68
Average Length of Stay	9
Family Doctor	Yes
Hypertension (Predicted)	**1 (Yes)**
Probablity of Hypertension (Predicted)	**64%**

Figure 51 Diabetes Comorbidities Tool - Output for I100

Figure 51 shows the output which contains the I100 diagnosis and prediction probability for a patient with specified history. For example, a predicted value of 1 indicates that the patient will have hypertension in future, and there is a probability of 64% for this to happen. This patient was diagnosed with multiple comorbidities, which included the other two target variables (I500, N179). This can be the reason for a high probability of 64%.

The prediction was in conformance with the actual data of this patient (11219) who was in fact diagnosed with hypertension. Similar assessment tools were built for other two target variables (I500, N179) using the Ensemble algorithm.

4.2.7 Summary

In the above experiments, an interesting observation was that a decrease in percentage of diagnosed patients for a target variable leads to an increase in predicted values for the corresponding patient groups. For example, I100, I500 and N179 had 18.9%, 9.9% and 9.3% of the total patients, respectively with average prediction accuracies of 83.5%, 92.4% and 96.5%, respectively. The reason is that the algorithms are able to train the models better when there is a lower number of patients. I100 and N179 had the highest (2,656) and lowest number of patients (1,303) with average corresponding accuracies of 15% and 63.7% when predicting their respective diagnosis.

It is observed that all algorithms perform relatively similar for each of the three target variables due to the following reasons:

- Auto Classifier node was used to identify the data mining algorithms with high accuracies for all three target variables.
- As mentioned in Chapter 3, only the important variables identified by FS algorithm were selected as input variables and passed to the models.
- All twenty diagnosis codes had binary data (0,1) which included the target variable that helped the five classification algorithms to make efficient predictions.

It is also to be noted that, Random Forest occasionally suffered from overfitting problem which trained models to learn the noise thereby leading to negative impact on accuracy [52].

Chapter 5

Conclusion and Future Work

Diabetes is a chronic disease whose prevalence is growing at a rapid rate throughout the world. It has also been called the biggest epidemic of the twenty-first century. The number of people with diabetes rose from 108 million in 1980 to 422 million in 2014 [53]. The global prevalence of diabetes among adults over 18 years of age rose from 4.7% in 1980 to 8.5% in 2014 [1] . In Canada, one person is diagnosed with diabetes every three minutes, and one in ten deaths are attributed to this disease. Due to this prevalence, it has received global attention and vast amounts of data has been collected. Unfortunately, this data exists in disparate repositories and has not been harnessed to its full potential. However, it is now a well-known fact that diabetic patients must monitor their health constantly because of a higher risk of developing additional comorbidities over time.

Hypertension and Acute Renal Failure have been found to be among the top four comorbidities (Figure 52) [54]. Diabetes Canada [55] also reported that in almost every clinical trial one third of the patients with Heart Failure also had diabetes. Additionally, Heart Failure occurs in diabetic patients at an earlier age at a rate which is two to four-fold higher in comparison with non-diabetic patients [55]. An early intervention and effective management is desirable to identify patients during the early stages of the disease. To this end, there have been efforts towards developing predictive models and

assessment tools for improving quality of life and reducing burden on the healthcare system. However, the existing solutions have several drawbacks as explained in Chapter 2.

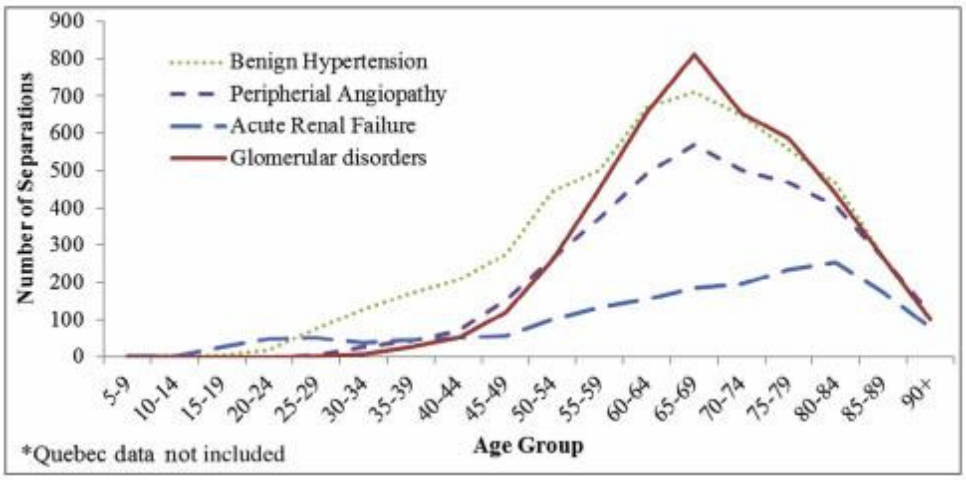


Figure 52 Comorbidities for Hospitalized Diabetic Patients in Canada [52]

One of the key shortcomings of existing research is the use of non-clinical data which is collected using surveys and self-administered questionnaires. The dataset used for this research was obtained from Northern Health which exclusively comprised of diabetic patients with either T1D, T2D or any other types of diabetes between the years 2012-2018. While this data contained only clinical records, it existed in the form of spreadsheets which made it difficult to analyze across a variety of parameters. In order to make data valuable for physicians and other stakeholders, several Key Performance Indicators (KPIs) were identified which provided insight into historical trends and patterns for using visual analytics. These metrics were presented in a visually appealing dashboard and

data was mined for predictive analysis. The developed models were then incorporated into an interactive assessment tool.

This research had two major contributions. First, Predictive models were developed to find the likelihood of one or more of three comorbidities - Benign Hypertension, Congestive Heart Failure or Acute Renal Failure using six algorithms. The model results were incorporated into a physician friendly assessment tool which is flexible to be connected to one of the six algorithms to predict diagnosis and likelihood of one of the three comorbidities. Results from the assessment tool can act as an effective guideline for healthcare professionals to identify high-risk diabetic patients thereby ensuring effective diabetes management to reduce costs on the healthcare system. Second, an interactive diabetes dashboard was developed to show an overview of the current state of diabetes for the years 2012-2018 in Northern Health. This dashboard was built with drill down capabilities to view aggregated results at finer granularities for various demographics (HSDA, LHA and Diagnoses). This research used a dataset specific to Northern Health facilities with majority of patients from Northern BC. However, both the dashboard as well as the predictive models have the capability to be extended to other regions and provinces which would reflect on the assessment tool as well.

Diabetes Dashboard

The dashboard was built using the Microsoft BI tool stack with provisions for integrating with diabetes dataset. The dashboard consists of three top-level reports. First, the main dashboard displays overall statistics for 14,021 patients who recorded 34,824 admissions

in various Northern Health facilities. Second, a diabetes comorbidities dashboard identified the prominent comorbidities for these patients. Third, a HSDA comparison dashboard provides overall statistics for the three HSDAs – NE, NI and NW. These top-level reports have drilldown capabilities to view reports at finer granularities. Several observations were made from these reports. For instance, it was interesting to note that 51% of the patients had been diagnosed with between two to five comorbidities in addition to diabetes. The three selected target variables (I100, I500, N179) were among the top ten most prominent diagnosis codes recorded in the NH dataset. There was a consistent decrease of new diabetic patients from 2012 to 2017 with a slight increase observed in 2018. In addition, it was noted that the average age of patients was found to be sixty-three.

Predictive Modeling

Patients diagnosed with diabetes can develop several other diseases over time. In this research, the focus was on identifying diabetic patients who are at a higher risk of being diagnosed with one or more of the following common comorbidities:

- I100 (Benign Hypertension)
- I500 (Congestive Heart Failure)
- N179 (Acute Renal Failure)

The reason for choosing these comorbidities was the large number of patients in the dataset who were diagnosed with at least one of these codes. For instance, there were approximately 19% of patients diagnosed with I100 (Hypertension). The other two target

variables I500 (Congestive Heart Failure) and N179 (Acute Renal Failure) also ranked among the top seven diagnosis codes with highest number of patients. Thus, these codes provide a good representation and also demonstrate how other comorbidities can be added to the study. Similarly, there are a number of data mining algorithms which are available in SPSS modeler. The following six representative algorithms were chosen to build our models:

- Bayesian Network
- Neural Network
- Random Forest
- Logistic Regression
- CHAID
- Ensemble

These six algorithms were evaluated for accuracy for the three target variables and analyzed. The important input variables for each target variable was determined by a built-in Feature Selection (FS) algorithm. It was observed that a decrease in the number of patients for target variables resulted in an increase in the accuracy of all algorithms. Another interesting observation was that Random Forest had a lower accuracy due to overfitting. Overall, an accuracy of 83.5%, 92.4% and 96.5% was observed for I100, I500 and N179, respectively.

Finally, a Diabetes Comorbidities Assessment Tool was built which took input from the user via an interactive web form and predicted the likelihood of one of the three target variables. This tool is flexible and can be connected to any one of the six algorithms to

predicts the probability of a patient to be diagnosed with one of the three comorbidities in future.

5.1 Future Work

The work presented in this thesis has demonstrated the importance of visual and predictive analytics using clinical data. However, during the process several challenges were encountered and a wish list for further work evolved. One of the characteristics of the NH diabetes dataset was that the diabetes diagnosis codes were combined with other comorbidities. For example, diagnosis code E1123 represents patients having “Type 2 diabetes mellitus with established or advanced kidney disease”. It would be more desirable to have an exclusive code for recording the type of diabetes (T1D, T2D, etc) and separate the comorbidities diagnosis of the patients. This can make it easier to segregate patients with different type of diabetes and find out specific comorbidities of patients as well. Since majority of patients in the dataset are diagnosed with T2D, it would be interesting to create dataset with only T2D patients and run the existing models for all three target variables. These results have the potential to reveal interesting correlations which are specific to T2D patients and can help healthcare professionals as well as patients to have a better understanding of their specific comorbidities.

The three selected target variables (I100, I500 and N179) had relatively fewer number of diagnosed patients in the dataset which lead to reduced accuracy in predictive modeling for those group of patients. It would be helpful to combine different diagnosis codes with

help of a Physician to increase the number of patients in these groups. For instance, the word ‘Heart’ is there in thirty-five diagnosis codes and ‘Hypertension’ was found to be in seventeen diagnosis codes. If all or at least, some of these codes can be combined, it would increase the number of diagnosed patients for the corresponding target variables. This increase of patients would reflect in the training dataset which can help enhance the accuracy of the models.

Another recommendation would be to use the six algorithms for predicting the three target variables and only choose the important variables identified by FS as shown in Figure 49. This would eliminate some of the diagnosis codes which were included earlier. This could potentially produce interesting comparative results on the performance of predictive models.

It would be interesting to capture patient migration between communities and connect it with admissions and number of patients in the corresponding facilities over the years.

Finally, adding time dimension to the metrics could allow a longitudinal study which could also predict the timelines when a comorbidity is likely to occur.

References

[1] Diabetes Canada, "What is Diabetes?," [Online]. Available: http://www.diabetes.ca/about-diabetes/types-of-diabetes. [Accessed 11 June 2020].

[2] Government of Canada, "Types of Diabetes - Canada.Ca," 20 10 2016. [Online]. Available: https://www.canada.ca/en/public-health/services/chronic-diseases/diabetes/types-diabetes.html. [Accessed 11 June 2020].

[3] Diabetes Canada, "Assess your risk of developing diabetes," 2020. [Online]. Available: https://www.diabetes.ca/en-CA/type-2-risks/risk-factors---assessments. [Accessed 11 June 2020].

[4] Government of Canada, "How to Prevent Type 2 Diabetes," 14 11 2008. [Online]. Available: https://www.canada.ca/en/public-health/services/chronic-diseases/diabetes/prevent-type-2-diabetes.html. [Accessed 11 June 2020].

[5] L. C. Rosella, M. Lebenbaum, T. Fitzpatrick, A. Zuk and G. L. Booth, "Prevalence of Prediabetes and Undiagnosed Diabetes in Canada (2007–2011) According to Fasting Plasma Glucose and HbA1c Screening Criteria," *Diabetes Care,* vol. 38, no. 7, pp. 1299-1305, July 2015.

[6] Canadian Community Health Survey, "Statistics Canada - Canadian Community Health Survey - Annual Component," [Online]. [Accessed 11 June 2020].

[7] Diabetes Canada, "WHY FEDERAL LEADERSHIP IS ESSENTIAL CONCERNING DIABETES," [Online]. Available: https://www.diabetes.ca/how-you-can-help/advocate/why-federal-leadership-is-essential. [Accessed 20 Dec 2017].

[8] Diabetes Canada, "Diabetes in Canada," February 2020. [Online]. Available: https://www.diabetes.ca/DiabetesCanadaWebsite/media/Advocacy-and-Policy/Backgrounder/2020_Backgrounder_Canada_English_FINAL.pdf. [Accessed 11 June 2020].

[9] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Computational and Structural Biotechnology Journal,* vol. 15, pp. 104-116, 2017.

[10] M. Marinov, A. S. M. Mosa and I. Yoo, "Data-mining Technologies for Diabetes: A Systematic Review," *Journal of diabetes science and technology,* vol. 5, no. 6, pp. 1549-1556, 2011.

[11] S. Sankaranarayanan and P. T. Perumal, "A Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies," in *2014 World Congress on Computing and Communication Technologies*, Trichirappalli, India, 2014.

[12] S. Chaudhuri and U. Dayal, "An overview of data warehousing and OLAP technology," *ACM SIGMOD,* vol. 26, no. 1, 1997.

[13] R. J. Koopman, K. M. Kochendorfer, J. L. Moore, D. R. Mehr, D. S. Wakefield, B. Yadamsuren, J. S. Coberly, R. L. Kruse, B. J. Wakefield and J. L. Belden, "A Diabetes Dashboard and Physician Efficiency and Accuracy in Accessing Data Needed for High-Quality Diabetes Care," *Annals of Family Medicine,* vol. 9, no. 5, pp. 385-405, 2011.

[14] S. Zahanova, A. Tsouka, M. R. Palmert and F. H. Mahmud, "The iSCREEN Electronic Diabetes Dashboard: A Tool to Improve Knowledge and Implementation of Pediatric Clinical Practice Guidelines," *Canadian Journal of Diabetes,* vol. 41, no. 6, pp. 603-612, 2017.

[15] Government of Canada, "The Canadian diabetes risk questionnaire," 29 03 2017. [Online]. Available: https://healthycanadians.gc.ca/en/canrisk?utm_source=VanityURL&utm_medium=URL&utm_campaign=publichealth.gc.ca/canrisk.. [Accessed 11 June 2020].

[16] X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," *The Kaohsiung Journal of Medical Sciences,* vol. 29, no. 2, pp. 93-99, February 2013.

[17] S. Kumari and A. Singh, "A data mining approach for the diagnosis of diabetes mellitus," in *2013 7th International Conference on Intelligent Systems and Control (ISCO)*, Coimbatore, India, 2013.

[18] J. Lindström and J. Tuomilehto, "The Diabetes Risk Score: A Practical Tool to Predict Type 2 Diabetes Risk," *Diabetes Care,* vol. 26, no. 3, pp. 725-731, 2003.

[19] V. Mohan, R. Deepa, M. Deepa, S. Somannavar and M. Datta, "A Simplified Indian Diabetes Risk Score for Screening for Undiagnosed Diabetic Subjects," *The Journal of the Associations of Physicans of India,* vol. 53, pp. 759-763, 2005.

[20] A. Althubaiti, "Information bias in health research: definition, pitfalls, and adjustment methods," *Journal of Multidisciplinary Healthcare,* no. 9, pp. 211-217, 2016.

[21] Microsoft, "What is SQL Server Management Studio (SSMS)?," 11 September 2019. [Online]. Available: https://docs.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms?view=sql-server-ver15. [Accessed 1 November 2020].

[22] IBM, "SPSS Modeler - Overview," 2020. [Online]. Available: https://www.ibm.com/ca-en/products/spss-modeler. [Accessed 22 September 2020].

[23] "Visual Studio 2019," Microsoft, 2019. [Online]. Available: https://visualstudio.microsoft.com/vs/. [Accessed 15 July 2020].

[24] A. Marcano-Cedeno and D. Andina, "Data mining for the diagnosis of type 2 diabetes," in *World Automation Congress 2012*, Puerto Vallarta, Mexico, 2012.

[25] L. Zhang, X. Shang, S. Sreedharan, X. Yan, J. Liu, S. Keel, J. Wu, W. Peng and M. He, "Predicting the Development of Type 2 Diabetes in a Large Australian Cohort Using Machine-Learning Techniques: Longitudinal Survey Study," *JMIR MEDICAL INFORMATICS,* vol. 8, no. 7, 2020.

[26] R. S. Anand, P. Stey, S. Jain, D. R. Biron, H. Bhatt, K. Monteiro, E. Feller, M. L. Ranney, I. N. Sarkar and E. S. Chen, "Predicting Mortality in Diabetic ICU Patients Using Machine Learning and Severity Indices," *AMIA Joint Summits on Translation Science Proceedings,* vol. 2018, no. 1, pp. 310-319, 2018.

[27] H. S. Kim, A. M. Shin, M. K. Kim and Y. N. Kim, "Comorbidity Study on Type 2 Diabetes Mellitus Using Data Mining," *Korean J Intern Med.,* vol. 27, no. 2, pp. 197-202, 2012.

[28] A. Dagliati, S. Marini, L. Sacchi, G. Cogni, M. Teliti, V. Tibollo, P. D. Cata, L. Chiovato and R. Bellazzi, "Machine Learning Methods to Predict Diabetes Complications," *Journal of Diabetes Science and Technology,* vol. 12, no. 2, pp. 295-302, 2018.

[29] P. Kekäläinen, H. Sarlund, K. Pyörälä and M. Laakso, "Hyperinsulinemia cluster predicts the development of type 2 diabetes independently of family history of diabetes," *Diabetes Care,* vol. 22, no. 1, pp. 86-92, 1999.

[30] K. E. Heikes, D. M. Eddy, B. Arondekar and L. Schlessinger, "Diabetes Risk Calculator: A Simple Tool for Detecting Undiagnosed Diabetes and Pre-Diabetes," *Diabetes Care,* vol. 5, pp. 1040-1045, 2008.

[31] M. Lau, H. Campbell, T. Tang, D. J S Thompson and T. Elliott, "Impact of Patient Use of an Online Patient Portal on Diabetes Outcomes," *Canadian Journal of Diabetes,* vol. 38, no. 1, pp. 17-21, 2014.

[32] A. Dagliati, L. Sacchi, V. Tibollo, G. Cogni, M. Teliti, A. Martinez-Millana, V. Traver, D. Segagni, J. Posada, M. Ottaviano, G. Fico, M. T. Arredondo, P. D. Cata, L. Chiovato and R. Be, "A dashboard-based system for supporting diabetes care," *Journal of the American Medical Informatics Association,* vol. 25, no. 5, pp. 538-547, 2018.

[33] G. Stiglic and M. Pajnkihar, "Evaluation of Major Online Diabetes Risk Calculators and Computerized Predictive Models," *PLoS One,* vol. 10, no. 11, 2015.

[34] D. Pyle, Data preparation for data mining, San Francisco, California: Morgan Kaufmann Publishers, Inc., 1999.

[35] S. García, S. Ramírez-Gallego, J. Luengo, J. . M. Benítez and F. Herrera, "Big data preprocessing: methods and prospects," *Big Data Analytics,* vol. 1, no. 1, p. 9, 2016.

[36] R. E. Bellman, Adaptive control processes: a guided tour, Princeton, New Jersey: Princeton Legacy Library, 2015.

[37] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering,* vol. 40, no. 1, pp. 16-28, 2014.

[38] N. Suhaimi and A. Ismail, "Comparing the Performance of Logistic Regression and Artificial Neural Networks Models: An Application to Type 2 Diabetes Mellitus," 2012. [Online]. Available: https://www.academia.edu/7511501/Comparing_the_Performance_of_Logistic_Regression_and_Artificial_Neural_Networks_Models_An_Application_to_Type_2_Diabetes_Mellitus. [Accessed 1 November 2020].

[39] M. M. Mijwel, "Artificial Neural Networks Advantages and Disadvantages," 2018 January. [Online]. Available: https://www.researchgate.net/profile/Maad_Mijwil/publication/323665827_Artificial_Neural_Networks_Advantages_and_Disadvantages/links/5aa2c01faca272d448b5a23d/Artificial-Neural-Networks-Advantages-and-Disadvantages.pdf. [Accessed 27 September 2020].

[40] H. Esmaily, M. Tayefi, H. Doosti, D. Ghayour-Mobarhan, H. Nezami and A. Amirabadizadeh, "A Comparison between Decision Tree and Random Forest in Determining the Risk Factors Associated with Type 2 Diabetes," *Journal of Research in Health Sciences,* vol. 18, no. 2, p. 412, 2018.

[41] S. B. Kotsiantis, I. Zaharakis and P. Pintelas, Supervised Machine Learning: A Review of Classification Techniques, Greece: IOS Press, 2007.

[42] Y. Guo, G. Bai, Y. Hu, "Using bayes network for prediction of type-2 diabetes," in *Internet Technology And Secured Transactions*, London, 2012.

[43] F. M. Díaz-Pérez, "CHAID algorithm as an appropriate analytical method for tourism market segmentation," *Journal of Destination Marketing and Management,* vol. 5, no. 3, pp. 275-282, 2016.

[44] G.Reachad,Michault, H.Bihan, C.Paulino, R.Cohena, H.Le Clésiau, "Patients' impatience is an independent determinant of poor diabetes control," *Diabetes & Metabolism,* vol. 37, no. 6, pp. 497-504, 2011.

[45] Urology Care Foundation, "Kidney Failure: Symptoms, Causes & Diagnosis - Urology Care Foundation," 2020. [Online]. Available: https://www.urologyhealth.org/urologic-conditions/kidney-(renal)-failure#:~:text=What%20is%20Kidney%20(Renal)%20Failure,kidney%20(or%20renal)%20failure.. [Accessed 08 10 2020].

[46] Microsoft, "Microsoft SQL documentation," [Online]. Available: https://docs.microsoft.com/en-us/sql/?view=sql-server-ver15. [Accessed 1 November 2020].

[47] Microsoft, "What is SQL Server Reporting Services (SSRS)?," 05 June 2019. [Online]. Available: https://docs.microsoft.com/en-us/sql/reporting-services/create-deploy-and-manage-mobile-and-paginated-reports?view=sql-server-ver15. [Accessed 27 September 2020].

[48] Tableau, "Tableau," [Online]. Available: https://www.tableau.com/. [Accessed 08 10 2020].

[49] Qlik, "Qlik," [Online]. Available: https://www.qlik.com/. [Accessed 08 10 2020].

[50] Datawrapper, "Datawrapper," [Online]. Available: https://www.datawrapper.de/. [Accessed 08 10 2020].

[51] FusionCharts, "FusionCharts," [Online]. Available: https://www.fusioncharts.com/. [Accessed 08 10 2020].

[52] T. Hastie, R. Tibshirani and J. Friedman, "Random Forests," in *The Elements of Statistical Learning*, New York, Springer, 2008, pp. 587-603.

[53] World Health Organization, "Diabetes," 8 June 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed 1 November 2020].

[54] A. Wielgosz, S. Dai, P. Walsh, J. McCrea-Logie and E. Celebican, "Comorbid Conditions in Canadians Hospitalized Because of Diabetes," *Canadian Journal of Diabetes,* vol. 42, pp. 106-111, 2018.

[55] K. A. Connelly, R. E. Gilbert and P. Liu, "Treatment of Diabetes in People With Heart Failure," 2018. [Online]. Available: https://guidelines.diabetes.ca/cpg/chapter28. [Accessed 1 November 2020].

[56] Microsoft, "ASP.NET Overview | Microsoft Docs," 08 October 2019. [Online]. Available: https://docs.microsoft.com/en-us/aspnet/overview. [Accessed 27 September 2020].

[57] "Supervised and Unsupervised Learning," 2011. [Online]. Available: https://sites.astro.caltech.edu/~george/aybi199/Donalek_classif1.pdf. [Accessed 2 Nov 2020].

[58] M. R. DEVI, "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus," *International Journal of Applied Engineering Research,* vol. 11, no. 1, pp. 727-730, 2016.

[59] G. Shmueli, "To Explain or to Predict?," *Statistical Science,* vol. 25, no. 3, pp. 289-310, 2010.