Finding Cohorts in Clinical Data

Developed by:
Kavan Mauleshkumar Patel

Guided by:
Dr Quan Sun
Dr Michael Lockyer

Submitted to:

Orion Health

&

The Department of Engineering, computer, and mathematical sciences

Year: 2021

Table of Contents

List of Figures

List of Tables

Abstract

Clinical trials play a key role in carrying out medical intelligence, but performing clinical trials gets difficult because of patient requirement phase; it is widely believed that patient requirement phase is one of the time consuming and critical, in which incorrect patient selection can lead to failure of clinical trial. In order to solve these problems Orion Health is establishing long term project named Smart Cohort. This long-term project contains a module name Cohort Finding, which aims to build a machine learning model that can predict whether patients has meet with selected criteria or not and can also visualize cohorts based on patients. In order to build these models, advance natural language processing methodologies, dimensionality reduction techniques and various machine learning models (Logistic Regression, Support Vector Machine, Random Forest, and Naive Bayes) were used. For predicting criteria Random Forest model performed best with measurable performance on eight criteria of n2c2 dataset. Along with-it t-sne based model was able to visualize cohorts based on patients. This two-machine learning model will reduce time for patient recrement phase. On the other hand, to increase the probability of selecting correct patient casual analysis was undertaken using DoWhy library and also suggested how it can be integrated in Smart Cohort.

Keywords – Clinical Trials, Model, Smart Cohort, Finding Cohorts, Criteria Prediction, Cohort Visualization, Impact Analysis, Logistic Regression, Support Vector Machine, Random Forest, Naïve Bayes, t-Sne, DoWhy.

Acknowledgements

Foremost, I would like to express my sincere gratitude to Dr Quan Sun (industrial supervisor) and Dr. Michael Lockyer (university supervisor) for the continuous support, patience, motivation, enthusiasm, and immense knowledge. Their guidance helped me all the time. I could not have imagined having a better advisor and mentor.

Besides my advisor's, I would like to thank Orion Health and AUT Engineering, Computer and Mathematical Sciences department for providing me with this great opportunity and learning environment.

1. Introduction

Orion Health provides health information technology, improving population health and precision medical solutions across the entire health ecosystem; this organization have been awarded globally. Company have 500+ employees in 20 offices across 13 countries. Technology provided by Orion Health is utilized to manage the health of over 100 million patients across the globe. Company has successfully deployed more than 55 large-scale regional and country wide solutions in more than 15 countries. In all, the Orion Health is offering software, services, and support to healthcare organizations over the last 28 years. The health sector will be greatly transformed by utilizing the technology and collecting and analysing individuals' health data (Health).

Currently, the company is undertaking a vast project name Smart Cohort. This project aims to build a User Interface (UI) based application, which enhance Clinical Trials. According to WHO, "clinical trials are a type of research that studies new tests and treatments and evaluates their effects on human health outcomes". Usually, clinical trials are carried out in five phases, which are as follows: Trial Design, Site Selection, Patient Recruitment, Outcome Tracking, and Medical Intelligence. In other words, the first trial is designed based on the criteria's (example: patient needs to have diabetes) and then the site is selected depending on the trial requirement. Once the site is nominated, suitable patients are recruited. New tests and treatments are tried on selected patients, and outcomes are tracked. Based on the track, medical intelligence is carried out.

According to survey conducted by (Antunes et al., 2019), patient enrolment is one of the most frequently reported difficulty in clinical research, additionally, three out of ten clinical trials were failed because of insufficient patient enrolment. Thus, it can be concluded that patient recruitment is one of the critical phases. This phase can be also known as cohort selection phase as patients are required based on cohorts. Patient recruitment consists of two steps, which are pre-screening and actual screening. Firstly, in pre-screening potential participants are identified utilizing various sources including labs and social media. Secondly, medical reports of these participants are manually analysed and identify the relevant list of patients who meet with the defined criteria of clinical trials. All listed patients are invited for the final screening where general question and answer (Q/A) session is held to verify the pre-screening results. If the answers are as expected, then patients get enrolled otherwise they fail to enrol. Client statistics says that for each patient it takes on an average 30 to 45 minutes to manually analyse the medical report and even after spending considerable amount of time approximately more than 50 percent of patients identified from pre-screening are failed to enrol in clinical trial. Similarly, (Haddad et al., 2018) mentioned that only 3 to 5 per cent of cancer patients had contributed to clinical trials where 20 per cent of patients were suitable for enrolment. Overall patient recruitment is one of the time-consuming and significant phase as incorrect patient selection or few patients' enrolment can lead to failure of clinical trial.

In order to target this problem Orion Health has undertaken a long-term project, Smart Cohort. This project targeting to build a fully automated application for cohort selection with a user-friendly interface. Smart Cohort application will work on 13 various selection criteria. Using this application, potential patients can be identified within a few minutes; hence, once the patients are selected, they directly undertake final screening; this application will have myriad

benefits (including resource management and cost redundancy) for health researching organizations. This research project will predominately be focusing on the minor module of Smart Cohort, which is cohort's selection. To be more specific, the project aims to develop a computational model (utilizing Machine Learning, Deep Learning, Natural Language Processing and Computational Linguistics) that can detect whether the patient has met with the selection criteria or not. Example: clinical trial requires patients who are diabetic and has been drug abusers than from the given medical report of the patient, model will predict that this patient has diabetes or not and it has been a drug abuser or not. In addition to it, the study also attempted to build an unsupervised model that can identify two groups of cohorts (Example: Diabetes and Non-Diabetes). Moreover, impact analysis will be performed on MIMIC-III dataset to identify whether Smart Cohort application can use impact analysis or not.

Myriad studies have proposed and implemented exclusive solutions to effectively select correct patients quickly based on the medical. Xiong et al. describe that to solve this problem, electronic medical reports are screened, but it is still time-consuming as myriad reports need to be reviewed manually (2021). In recent years with the help of advancement in Natural Language Processing, important medical information is extracted from electronic medical reports; and on top of this computational methods (such as (Chen et al., 2019) Rule Based, (Oleynik et al., 2019) Machine learning, (Segura-Bedmar & Raez, 2019); (Xiong et al., 2021) Deep learning, and (Vydiswaran et al., 2019) Hybrid methods) are utilized for cohort selection.

Various studies have also visualized patients based on the cohorts. (Oleynik et al., 2019) Visualized patients based on the cohorts, which are created depending on the pelvic organ variability. Likewise, (Diaz-Papkovich et al., 2019) used dimensionality reduction technique (Principal Component Analysis (PCA)) for genomic cohort visualization based on patients. However, both of this research has not considered date of the diagnosis, which is critical information. (Xiong et al., 2021) has built a pipeline for data extraction and also mention that it will help in identify date of diagnosis, which is significant information for care of patient and research.

According to Sharma and Kiciman, Machine learning models are established under correlation analyses and pattern recognition, that are not enough for casual inference because in real world project it is important to answer, "What if" and "Why" questions (2020). (Prosperi et al., 2020) Biomedical observational studies are widely affected by confounders and selected bias. Additionally, predication models are mostly misinterpreted based on casual effects. (Sharma & Kiciman, 2020b) has developed a DoWhy library which can perform impact analysis based on the assumptions and statistical models. Along with it also have capability to validate estimation of casual effects.

The scope of this project (Smart Cohort) will be only limited to Orion Health. However, developed methodology in this research paper will not only be limited to the specific hospital or organization because the design will be based on the n2c2 and MIMIC III researching dataset so methodology can be used in any health organization that is undertaking medical research.

The structure of this report is in the following format. First, it will introduce the general background of the Orion Health, Clinical Trials and the importance of Cohort Selection; additionally, this section will also describe the motivation and objective of the research. The

following section will compare and contrast the existing research in this area and its limitation; side-by-side, it will provide details about the research gap. Furthermore, the methodology section will emphasize the comprehensive pathway of the research experiment. Moreover, the analysis findings will be declared, and on the basis of that, the discussion will be carried out. Lastly, a brief summary will be mentioned, along with limitations and future work.

2. Research Questions/Objectives

Orion has undertaken a long-term project named Smart Cohort. This research project will focus on the minor module of the long-term project which is to find cohorts in various selection criteria tags. The predominant aim of this research project is to establish a model (using Machine Learning, Natural Language Processing and Computational Linguistics) which can identify whether the patient matches specific selection criteria or not. In addition, the study also targets to build an unsupervised model which can visualize two groups of cohorts (example: Diabetes and Non-Diabetes). Moreover, to identify correct patients from cohorts it is also essential to perform impact analysis.

This work will significantly contribute to Orion Health in building a Smart Cohorts application for their client; utilizing this application the client (a hospital activity performing medical research) will be able to find the appropriate patients for the innumerable types of Cohort analysis. The designed methodology will not be restricted to specific hospital research; it can benefit the health research industry in terms of saving time and massive capital. In order to achieve the goal of this research the following sub-objectives need to be researched:

- Identify suitable methodology and the best performing Machine Learning model for predicting whether patients have met with specific criteria or not.
- Establish appropriate unsupervised methodology and identify best unsupervised techniques that can used for cohort visualization.
- Can impact analysis be used in Smart Cohort?

3. Literature Review

3.1 Overview

This research focused on the three objectives where each of them has different methodology. Hence this section will review previous linked and similar research to establish accurate methodology for each of the objectives. Along-side, study will also focus on analysing and finalizing applicable techniques (such as Machine Learning and Natural Language Processing algorithms) in defined methodology.

3.2 Objective One (Criteria Prediction)

Various studies have proposed several analyses for identifying and selecting a correct patient for clinical trials based on criteria prediction. (Xiong et al., 2021) The primary methods for clinical cohort selection are Rule-based Methods, Machine Learning methods, and Hybrid methods. (Antunes et al., 2019) created a system that can automatically perform the cohort selection. In order to build this system, a combination of the heuristics and different machine learning algorithms are used based on the 13 selection criteria tags. The dataset used for this

model creation was n2c2 2018 shared task track1 - Cohort selection for clinical trials; in addition, actual data was expended by using the external database MIMIC-III critical care database. To create this system, initially, the rule-based method was implemented. Hand-crafted rules were created to find text patterns from all the medical reports, and patients are selected based on these rules. In some tags, this method was outperforming, which means it could not identify whether the patient met the selected criteria or not. To overcome this problem, two rule-based classifiers were developed where one was used to submit the results to the n2c2 Shared-Task; once the submission is completed, another rules-based classifier will perform comprehensive error analysis on the training dataset and improve specific previous rules. However, it was found that this manual modification could lead to the problem of overfitting in the training phase. Moreover, classical machine learning models were implemented. Firstly, raw text was converted into the vectorized form using the Bag-of-Word approach; all the words were converted into lower case, and stop words were removed in the tokenization step; additionally, bigrams and trigrams were found to be not significantly, so unigrams were utilized. Lastly, using API's named scikit-learn and xgboost, the following machine learning algorithms were implemented: AdaBoost, Bagging, Decision Tree, Gradient Boosting, and XG Boost. Furthermore, two deep learning algorithms were experimented, which are Artificial Neural Networks (NN) and Convolutional Neural Networks (CNN). To implement these deep learning models' raw text was converted into the word embeddings; this methodology was applied on both datasets (n2c2 Shared-Task and MIMIC-III) using word2vec architecture; hundred thousand unique words were finalized in the final vector. On the top of this vector, NN and CNN were implemented with different configurations, but for all the tagging criteria, it was providing a lower F1 score in comparison with other machine learning models and rule-based models. In the final step, these three methods were combined as a single system or software; hence, on the selected tag from all models best-fit model will be found, and cohort patients will be listed based on it. Overall, this approach has provided substantial results; in training dataset average micro was 0.9143, whereas on testing dataset it reduces to 0.8844. On the other hand, the average macro on the training dataset was 0.8596, and on the testing dataset it was 0.7271. However, the result shows the clear notation of the overfitting because the macro-average of the testing set is roughly 13 per cent smaller than the macro-average of the training set.

This research methodology does have various drawbacks. One of the major limitations is the lack of pre-processing. (Yang & Hong) stated that the abbreviation could explain domain knowledge of the document, so it could be essential to handle it in the initial stage of the data cleaning otherwise will be removed in the other phase of the cleaning process (2017). (Yafoz & Mouhoub, 2020) cited that lemmatization is one of the most common tasks in NLP-related projects because it deletes the inflectional ending and modifies the words into the base form, which is known as the lemma. Moreover, (Kanakaraj & Guddeti, 2015) applied stemming to convert all the words into the base form, which help to capture semantics and improve classification accuracy. However, in (Antunes et al., 2019) research design, lemmatization or stemming was not implemented on the medical report. Another key drawback of this study is utilizing the traditional Bag of Word (BOW), while converting medical reports into the vector space. Although using the BOW model, all the words are considered equally important in the document; but Term Frequency-Inverse Document Frequency allocates specific weight to each word in the document. These weights are allocated based on the word occurrent in the specific document and word occurring in a set of documents (Kherwa & Bansal, 2020). On FIRE 2011

dataset (Mishra & Vishwakarma, 2015) experimented various vector space models including BOW and TF-IDF; it was found that TF-IDF was performed best among other vector space models. Additionally, the downside of this approach is that it does not capture short term dependencies. (Sethy & Ramabhadran, 2008) used n-grams and a vector space model to secure all the dependency of the words. This method has provided reasonable results in finding cohort modules so it can be used in the Orion project. However, limitations (including advanced pre-processing steps, and utilizing TF-IDF) mentioned above will also be considered and implemented, which might ultimately improve the results.

Similarly, (Segura-Bedmar & Raez, 2019), have also used the n2c2 Shard Task dataset and deep learning algorithms for cohort selection; however, CNN models which were established are totally different. Four various deep learning architectures were proposed; the following are the four models: Simple CNN, Deep CNN consisting of three convolutional blocks, RNN along with gated recurrent units, which is GRUs, and Hybrid model, which comprises a CNN followed by a GRU-RNN. These approaches are extensions to the classical CNN architecture. The activation function at the output layer for all the models was sigmoid. Firstly, medical reports were converted into matrices by utilizing random initialization and word embeddings methods; in other words, records of each patient were represented in the form of a matrix of word embeddings. The first approach is simple CNN, a matrix of word embeddings is passed through the convolutional layer, which will apply a series of filters with various sizes; this filter will produce feature vectors, which will be input to the max-pooling layer. This layer will capture the model relevant features depending on the maximum values generated by different filters. Relevant features generated by max-polling layers will be the input to the FF layer. The second approach is Deep CNN consisting of three convolutional blocks. The parallel architecture was followed to Simple CNN; the only difference was that in this model, three convolutional blocks were added with different filter sizes (64, 128, and 256), and two consecutive convolutional layers were there in each block. The third methodology was based on the RNN; this approach is an extension to the Simple CNN where medical reports are processed token by token and store the semantics of the previously used tokens into the hidden layers. The final approach is Hybrid CNN-RNN, which is the combination of the CNN and RNN; Features were learned by training the CNN and then selected patient features will be the input to the RNN, which will decide whether the patient is the perfect match for the selected tag or not. Likewise, (Oleynik et al., 2019) developed a deep learning model named Long Short-Term Memory utilizing manually created features along with rule extracted trigger module and word2vec embeddings. Moreover, for input representation scheme to LSTM, self-trained and pre-trained embedding were used, along with information of subworlds. In all deep learning models which are developed in both the studies were acceptable in terms of prediction, but it was found that on average F1 score and AUC score was comparably less than other classical machine learning models (Ada Boost, Decision Tree, Gradient Boosting, XGB) implemented in (Antunes et al., 2019). Hence, while developing Cohort finding model for Orion Health deep learning models will not use.

(Xiong et al., 2021), recommended, recently developed BERT for cohort selection. The methodology used in this study is also known as the four-layer unified MRC method; in which input to the BERT model was provided in the form of <Cube, Question> and this cub and question were mapped based on Name Entity Recognition. Consequently, it captures cross-criterion relations and provides label prediction. As a result, it was concluded that MRC with

simple rules is able to select cohort patients with a computable averaged micro F1 score of 0.9163, which is higher than the average micro F1 score obtained by the classical machine learning model in (Antunes et al., 2019). In this research of Orion health project (finding cohort) BERT embedding model will be consternated, and results will be compared to the classical machine learning models.

3.3 Objective Two (Cohort Visualization)

According to Dunn Jr et al., in clinical research, visual analytics is correct practice for examine parameters across patients (2017). In this cohort finding research, it is essential to visualize patients based on criterial. (Zhang et al., 2012) mentioned that to establish healthcare discoveries patient cohort analysis analyse medical and diagnostic histories of patients. There are also various studies that has visualize patients based on cohorts to find patients level insights.

Research (Grossmann et al., 2019) has visualized the cohorts based on the pelvic organ variability. This study used four step processes where initial step is data cleaning. In this phase source data has been transformed into a form which can be easier to visualize. Along-side, time was identified from medical reports and based on that time reports was arranged. Next step was to convert this patient reports into patient level embedding. Lastly, to visualize embedded report dimensionality reduction techniques was implemented. Basically, this method converts multidimension data into two-dimension/three-dimension. Lastly, this reduced dimension was visualized based on cohorts. It can be concluded that for this project initial step needs to be patient level embedding, but to implement it is necessary to arrange patient reports based on recorded date. (Fu et al., 2020) stated that important information for patient care and research is date of diagnosis. In addition to it research has built a pipeline for date extraction that is build using NLP tools and techniques. Parsedatetime. spaCy, and regular expression were tested to identify dates from clinical notes. Parsedatetime is a package that is built to parse human-readable date-time from given text. Additionally, spaCy is NLP tools, which is used for Name Entity Recognition; provided a text it identifies date entity, but the only limitation is that to tag a date a string must follow constraints of grammar. Lastly, regex is built on bases of regular expression (regex) that extract dates from provided text spans. These all function were experimented on i2b2 challenges dataset and as a result it was found that any of these functions were able to identify more than 90 precent of correct date. (Pypi, n.d.) datefinder function can identify all sorts of date like string from provided document, but in medical notices all number are not always dates; for example, “1.4 mg” is identified as date “1-4” (first day of April). To overcome this problem, first it is essential to identify medical terminology. (Kormilitzin et al., 2021), establish a Med7 library that is used to recognize drug names, route of administration, frequency, strength, duration, form, and frequency from provided medical report. This model was fine-tunned on the 2 million free text patients report from MIMIC-III corpus. It can be concluded that for finding cohort research project first medical terminology need to be identify and then for accurate date detection various pre-defined and manual rules library will be tried to acquire higher accuracy. Next step is to distinguish, sort and embedded medical reports based on recorded date. Similar embedding techniques (TF-IDF, BERT, Orion Document Embedding) that are used in criteria predication model will be applied. High dimension embedding need to be converted into two/three dimensions for cohort visualization, so dimensionality reduction methods need to be applied. (Diaz-Papkovich et al., 2019) has

visualize the genomic cohorts utilizing dimensionality reduction techniques. First, most commonly used method was used, which is Principal Component Analysis (PCA). This method recognizes and classifies directions in the genotype space, which explain least variance. Using positions of patients and ranks of variance patients' genotypes could be summarize. However, it was found that projection generated by PCA might hide finer features of data and it's difficult to interpret so to overcome this problem t-Distributed Stochastic Neighbour Embedding (t-SNE) was implemented. t-SNE grouped patients which are roughly related to continent of origin, and this groups are establish by considering small ethnic sub-groups that visible within the large clusters. Still, this method outperforms on huge datasets, when a large number of neighbourhood optimal outlines creates convergence to the globally selected solution. This issue can be overlooked for finding cohort project as research is based on the small sample (n2c2 dataset).

Overall, for cohort visualization in this research project of Orion Health, first, a datetime extraction function will we created using, manual rules (regular expression), datafinder, spaCy, and Med7. Second, medical report will be embedded using TF-IDF, BERT, and Orion Document Embedding. Third, dimension will be reduced utilizing PCA and t-SNE.

3.4 Objective Three (Impact Analysis)

According to Sharma and Kiciman, in a decision-driven computer system, it is crucial to predict and comprehend causal effects of investigation correctly. A machine learning model that is built on correlation analysis and pattern reorganization is insufficient for decision making; a casual inference test needs to be involved (2020). Similarly, (Sharma & Kiciman, 2020a) stated that machine learning models are build based on the correlation analyses and pattern recognition, which are insufficient for casual inference analysis; moreover specifically, correlation analyses are not sufficient for answering "Why", and "What if" questions. In addition to it, (Domingos, 2012) mentioned that widely machine learning is applied on observational data, in which predictive variable are not under control of the learner because it opposed to the experimental data. Using observational data only few algorithms can extract casual information, but their applicability's are fairly restricted. However, correlation will provide a sign of a potential casual connection, which can be further investigated, to verify actual causation. For example, it was found the cholesterol and high glucose are highly correlated that does not mean that high glucose is caused by cholesterol, it can only be concluded that this both activities occur simultaneously, but to find actually causation of high glucose casual inference analysis need to carry out. (Prosperi et al., 2020) When it comes to biomedical research, observational research is mostly affected by confounders and selection bias. Besides, prediction models are mostly mistakenly used to draw casual inference effects, but in reality, neither their features nor their predictions essentially have a casual explanation. (Sharma & Kiciman) emphasize that there are libraries that can apply state of the art casual inference methodologies, and this can speed up the assumption of casual inference, but it was found that one of the major challenges is the practice of modelling assumptions and consequences of these assumptions for casual recognition and estimations. Furthermore, verifying and testing is another predominant challenge in casual inference analysis, not like supervise machine learning algorithms that can be verify utilizing testing dataset, casual inference analysis often has no ground truth answer on hand; hence examining core assumptions and utilizing sensitivity experiments is critical to gain self-assurance of results.

However, this all challenges can be targeted by using DoWhy casual inference library (2020b). This is end-to-end library for casual inference analysis, and it is built on the recent research in robustness checks and modelling assumptions. Therefore, in this research DoWhy library will be used to perform impact analysis on MIMIC-III dataset. (Sharma & Kiciman, 2020b) DoWhy is established on a simple unifying language for casual inference analysis, and it will be implemented using four steps, which are Model the casual question, Identify the casual estimand, Estimate the casual effect, and Refute the obtained estimate. To be more specific, firstly, casual graphical model will be created based on the prior knowledge about some of the variables and rest of the variables will automatically be consider as potential confounders. Secondly, in modelling casual question phase, all the possible ways will identify for detecting casual effect depending on the graphical model and for this identification do-calculus and graph-based criteria will be used. Following are the supporting criteria: Back-door criterion, Front-door criterion, Instrumental Variables, and Mediation. Thirdly, casual effects will be estimated using various methods that are based on both back-door criterion and instrumental features. Alongside it also support a permutation test and non-parametric confidence intervals for testing adopted estimation methodologies including Propensity-based Stratification, Propensity Score Matching, Inverse Propensity Weighting, Linear Regression, Generalized Linear Models, Binary Instrument/Wald Estimator, Two-stage least squares, Regression discontinuity, and Two-stage linear regression. Lastly, obtained estimations will be refute, to validate estimated effects from casual estimator, various methods can be used including Add Random Common Cause, Placebo Treatment, Dummy Outcome, Simulated Outcome, Add Unobserved Common Cause, Data Subset Validation, and Bootstrap Validation. This research project will follow the same four step process using DoWhy library.

Overall, based on the above-aforementioned research, it can be concluded that cohort finding problem can be formulated in several ways, and it can be targeted using a wide range of methodology including Rule-Based Models, Machine Learning Models, Natural Language Processing and Deep Learning Models. Yet, the Deep Learning model is not effective compared to the Machine Learning and BERT model. Also, cleaning medical reports is a core module of the whole project as it can be resealable affect accuracy. Besides, all the investigations followed the Knowledge Discovery in Databases method along with NLP techniques. Lastly, causal inference analysis will be conducted using DoWhy. These studies also offered an in-depth understanding of the NLP, Machine Learning Models and BERT, along with a clear pathway to find cohorts from the clinical reports.

4. Methodology

4.1 Overview

Depending on the three various objectives methodology is formulated for each of them. For criterial prediction and cohort visualization Knowledge Discovery in Databases (KDD) methodology will be followed. KDD is most broadly used approach for Data Mining and Machine Learning related problems. According to Maimon and Rokach, KDD is the "organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets" (2005). The structure of the experiment in all the reviewed studies was following KDD methodology along with NLP techniques. Moreover, impact analysis will be performed using four step processes of casual inference, which are Modelling, Identifying, Estimating,

and Refuting. (Antunes et al., 2019), stated that DoWhy library is build based on a simple unifying language for casual inference.

4.2 Actual Data Description:

For finding cohort research two researching dataset will be used, which are n2c2 Shared Task 2018 National NLP Clinical Challenges and MIMIC III. Overall, both of the dataset consists of patient medical reports which is written in human language (English).

n2c2 data consist of small training dataset that contain 202 annotated xml files medical report of patients, including 887 medical narratives. Additionally, this dataset also has testing set which contain 86 medical reports (xml file) of various patients and narratives of patients in each report are further classified based on recorded date; in total there are 377 narratives. Individual xml file has a sequence of 2 to 5 narratives and was interpreted based on patient level along with 13 criteria (as mentioned in Table 1) status, which are "met" or "not met". It was found the only seven criteria were balanced in the dataset (with more or less similar classes)

Criteria	Balanced/Unbalanced	Criteria Description
Abdominal	Balanced	History of intra-abdominal surgery
Advanced-cad	Balanced	Advanced cardiovascular disease
Alcohol-abuse	Imbalanced	Current weekly alcohol consumption is over recommended limits.
Asp-for-mi	Semibalanced	Use of aspirin to prevent myocardial infarction
Creatinine	Balanced	Serum creatinine above the normal limit.
Dietsupp-2mos	Balanced	Use of dietary supplements in the last two months
Drug-abuse	Imbalanced	Drug abuse
English	Imbalanced	The patient can speak English
Hba1c	Balanced	Glycated hemoglobin levels between 6.5% and 9.5%.
Keto-1yr	Imbalanced	Ketoacidosis in the last year
Major-diabetes	Balanced	Major complication due to diabetes.
Makes-decisions	Imbalanced	The patient can make decisions by himself
Mi-6mos	Imbalanced	Myocardial infarction in the last six months.

Table 1 Overview of the target classification criteria in the 2018 n2c2 shared task track 1

MIMIC-III is vast database which is obtained from the Medical Information Mart for Intensive Care (MIMIC-III) and it consists of comprehensive, time-stamped information for more than 46,000 ICU patients at Beth Israel Deaconess Medical Centre (BIDMC). To be more specific information include name of hospital, sex, data-of-birth (DOB), date of admission, clinical observations, microbiology test results, and diagnosis codes (formatted in International Classification of Diseases 9th Edition, Clinical Modification [ICD-9-CM]). Finding cohort require this dataset to be used in criteria prediction mode for imbalanced dataset; but all the information's are not essential; based on objective (criteria prediction) only medical reports (written in human readable language, English) and ICD-9-CM will be used.

4.3 Objective One (Criteria Prediction)

In order to predict whether patient has been met with the selected criteria or not KDD methodology has been formulated based on problem. Figure 1 illustrates the outline of the

formulated KDD methodology in five step processes. First, actual dataset will be processed through the Data Cleaning phase, which will clean-up the medical reports using advance NLP techniques. Second, clean medical reports will be converted into the embedding. Third, using feature selection appropriate information will be extracted based on the criteria, and this extraction will be achieved from embedding. Fourth, utilizing relevant features embedding various machine learning model will be implemented for each of the criteria. Fifth, all applied models will be evaluated, and best fit model will be selected.

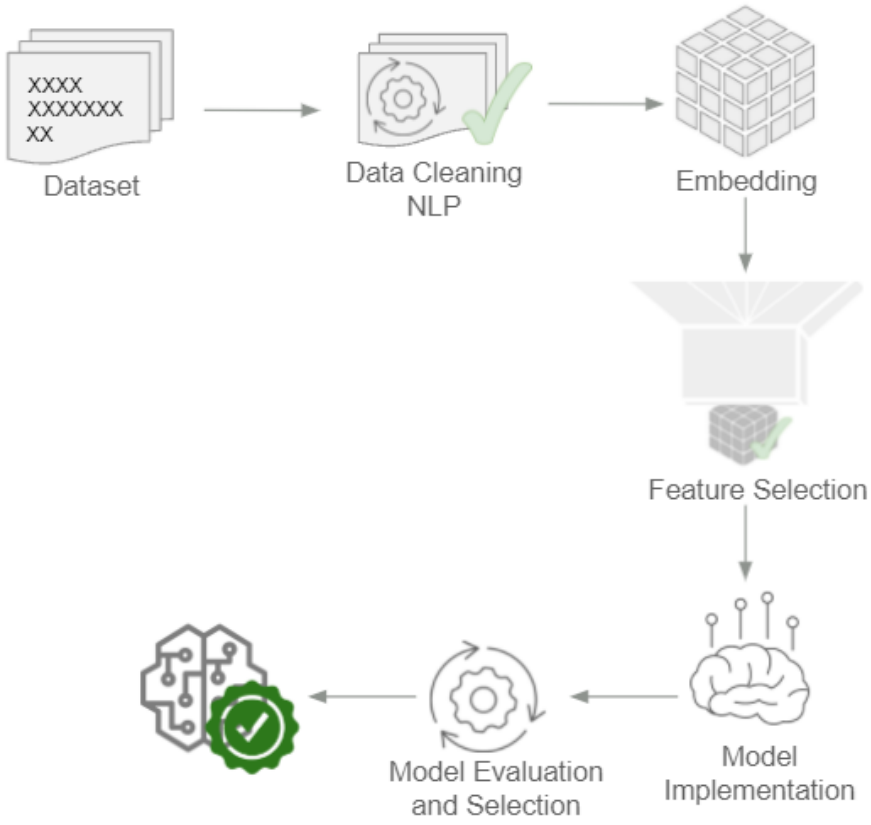


Figure 1 Overall Methodology for Criteria Prediction

4.3.1 Dataset

This objective first develops a machine learning model based on the n2c2 training dataset; additionally, also train a model based on combination of the sample set of MIMIC-III and n2c2 training dataset. Initial task is formulated both datasets.

Figure 2 represents process of n2c2 datasets construction. In this process status of all criteria's will be retrieved from the medical reports and based on criteria dataset is constructed.

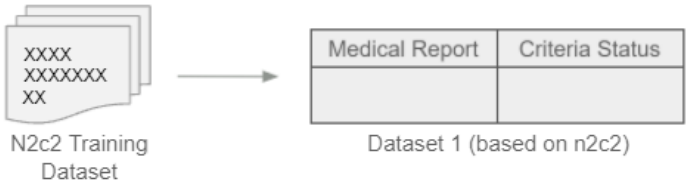


Figure 2 N2c2 Dataset

Figure 3 emphasize the process of creating second dataset which is combination of n2c2 and MIMIC-III. As final dataset medical reports are provided for each patients along with the criteria status.

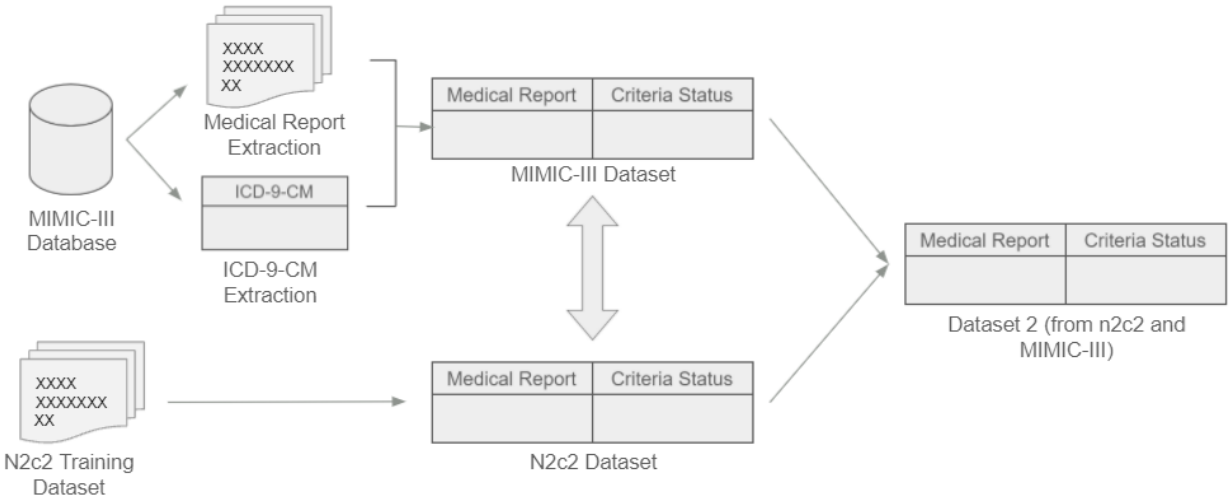


Figure 3 MIMIC-III and N2c2 Dataset Merging

4.3.2 Data cleaning (NLP)

According to (Marinov & Efremov, 2019), data cleaning is important phase in field of machine learning, particularly in NLP. (Lattar et al., 2020) mentioned that quality of data has become a critical issue and this issue becomes more critical in medical domain where need for valuable decision making is essential. In this situation the requirement for data cleaning to enhance quality of data become crucial. Figure 4 represents, method used for cleaning medical reports.

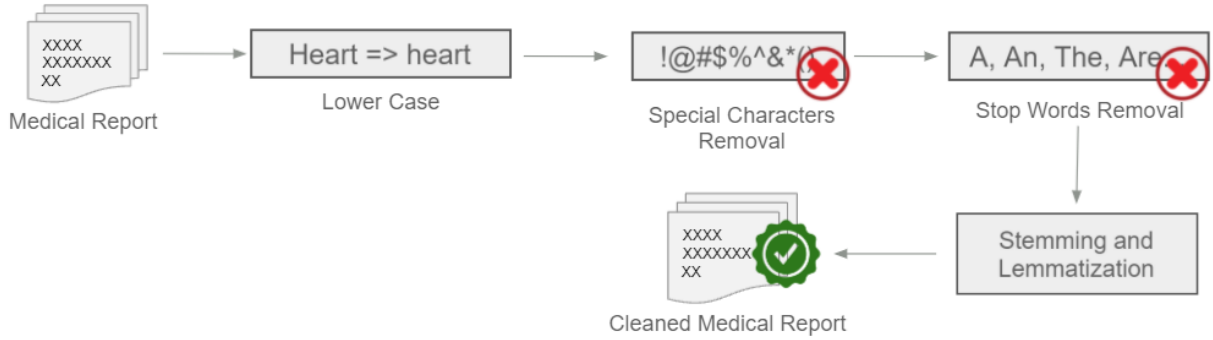


Figure 4 Data Cleaning Process for Criteria Prediction

In data cleaning process, first all the text will convert into lower caps. Then all the special characters stop word will be removed. Once unnecessary words are being removed stemming and lemmatization will be performed. As a outcome of data cleaning phase cleaned medical reports are extracted.

4.3.3 Embedding

Machine Learning algorithms are directly unable to understand human written medical report or any other text; hence it need to convert into embedding. (Wikipedia, n.d.) Embedding

usually represents words of text into the real valued vector that capture the actual meaning of the word in a way that words which are closer in the vector space are more similar in the meaning. In this project medical report will be embedded using TF-IDF and BERT Embedding (Developed by Orion Health). According to Balakrishnan and Lloyd-Yemoh (2014), term frequency-inverse document frequency (TF-IDF) is used to convert document into vector space, which is also known as embeddings. TF calculate how often particular medical term occur in the medical report and TF can be calculated as mention in equation 1.

$$tf_{t,d} = \frac{N_w}{N} \quad (1)$$

Where

N_w *= Number of time medical term occurs in the medical report*

N *= Total number of medical terms in medical report*

Term Frequency (tft,d) of an individual term *t* is the number of occurrence of that term into the specific document *d* where document in this research project is medical reports Once the occurrence of the word or terms is calculate in individual medical report. ((Alharthi et al., 2017); (Zhao et al., 2018)) Inverse document frequency (IDF) will we used to distinguish all the terms specified in the document (medical report) and this will be done by assigning weights to individual words based on the occurrence of words in set of documents (medical report). To calculate IDF following formula 2 has been used:

$$idf_t = log\left(\frac{N}{df_t}\right) \quad (2)$$

Where

N = Number of medical reports

df_t *= Number of Documents where term t occur*

Once the TF and IDF is calculated for all terms, it needs to be multiply to calculate TF-IDF, as show in equation 3 ((Alharthi et al., 2017); (Zhao et al., 2018)):

$$tf - idf_{t,d} = tf_{t,d} * idf_t \quad (3)$$

Orion Health has developed their own Bidirectional Encoder Representations for Transformers (BERT) embedding model which will be used to perform embedding. (Mu et al., 2021) Working of BERT can be described in two phases, first is training BERT that will understand the language (medical report language) and then in second step fine tuning of BERT will be implemented based on the problem that needs to be solve; in this research it based on the report embedding.

Specifically in phase one, by tanning two unsupervised task simultaneities (Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) model will understand what language and context is. MLM BERT will consider a sentence with random words that are filled with masks, and output of this make tokens will be produced. Parallelly, NSP BERT will consider two different sentences and identify whether sentence two is following the provided first sentence or not. Moreover, every word of the sentence is a particular token and then using pre-

trained embedding model these tokens are converted into embeddings ($E_1\ to\ E_n$). Output layer will be the binary variable of NSP that will be one if sentence B follows sentence A else it is zero. In this layer word vectors will be present that are denoted by $T_1\ to\ T_n$ and this will contain output of the MLM problem, hence number of word vectors that are outputted are equal to the input word vector

Second phase is fine tuning of BERT, in this phase only output layer of BERT is replaced with the new set of output layers, which will be related to the embedding of the medical report. Updated output layer will provide an embedding of each medical report based on the sentences. Once output layer will be replaced supervised training will be implemented using actual dataset.

Once both embedding will be generated using both methodology (TF-IDF and BERT) then final results will be compared.

4.2.4 Feature Selection

All the information provided in medical report are not always relevant so using created embedding significant features will be identified in this phase. According to the Tang and Zhang, in data mining feature selection is valuable method for reducing the dimensionality of data (2020). In medical diagnosis, it is significant to identify features related to the disease because these identified features will help to eliminate unnecessary attributes from the actual dataset (medical report) and improve the performance of the model.

The Boruta feature selection algorithm is wrapped around random forest algorithm. This algorithm considers fluctuations in the average precision loss of generated trees in the forest and then to measure the improvement it utilizes the average drop accuracy. Primarily depending on the actual feature set, correlation between predicted value and feature will be removed by generating mixed shadow feature set prior to selection; this is more beneficial for processing biomedical data consist of strong correlation features. The Boruta feature selection algorithm involves the following steps:

- Hybrid feature *N = [M,P]* will be created by randomly scrambling the original sample feature (*M*) which will generate shadow feature *P* from the sample.
- Elevation of the correlation between features and dependent variables will be accomplished and mixed feature set will be disorder.
- Based on the mixed feature set random forest model will be build and importance of all features will be calculated. Average reduce accuracy Z value will be consider as a measure of the feature significance. Greater the Z value more important feature is and greatest Z value in the shadow feature will be denotated as Max_Z.
- Feature will be considered important only if the Z value of the specific feature will be higher than Max_Z and bilateral equality test result are varying; else feature will be considered unimportant and will be removed.
- This process will be continued until all the features are confirmed or rejected. Algorithm will stop if it reaches to the maximum iterations.

4.3.5 Model Implementation

Once important features will be identified for each criterion, the next step is to build various machine learning models. Predicting whether patient has been meeting with selected criteria or not; this is a binary classification problem so in this study four classification algorithm will be experimented, which are Logistic Regression, Support Vector Machine, Nave Bayes, and Random Forest.

4.3.5.1 Logistic Regression

(Kleinbaum et al., 2002) Logistic regression is used to measure the relationship among a categorical dependent variable and independent variables by plotting probability scores of dependent variables. This statistical model is based on the logistic function that model a dependent binary variable; still complex extensions are exist. Term logit refers to "log odds", probability that is modelled. As show in equation 4 this term is defined as the ratio of the probability that an even occurred to the probability that event does not occur:

$$odds = \frac{p(event)}{1 - P(event)} \tag{4}$$

In logistic regression, sigmoid function is used to model a probability using a curve where predictor domain X can be anything and the range of $p(X)$ is between 0 and 1. Equation 5 represent the mathematical formulation of sigmoid function.

$$Sigmoid\ Function: p(X) = \frac{1}{1 - e^{-\beta X}} \tag{5}$$

In machine learning goal is to estimate parameter to make predictions. The parameters in the equation in the two-class classification in logistic regression is $\hat{\beta}$ vector. This regression utilizes maximum likelihood for parameter estimation. First all the training data will be considered and splatted into two groups based on their labels (cohorts). For each sample labelled "1" the goal is to estimate the vector $\hat{\beta}$ such that $\widehat{p(X)}$ is close to 1 as possible. On the other hand, for the sample group with labelled "0" goal is to estimate the vector $\hat{\beta}$ hat such that $1 - \widehat{p(x)}$ is close to 1 as possible. Mathematical formulation can be written as shown in equation 5,6, and 7.

$$Sz = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{5}$$

Where

$z = Dependent\ Variable$
$X_i = Independent\ Variable$
$\beta_i, \alpha = Unknown\ Parameters$

$$z = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{6}$$

$$f(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-(-\alpha + \sum \beta_i X_i)}} \tag{7}$$

There are various assumptions of logistic regression, which are as follows:

- A linear relationship is not assumed between independent and dependent variables.
- Error term (residuals) are not needed to be normally distributed.

- Homoscedasticity is not necessary.
- Dependent variable is not calculated based on interval or ratio.

4.3.5.2 Support Vector Machine

Support Vector Machine (SVM) is a linear statistical model used for classification and regression problems. Overall idea of Support vector machine is that it creates a hyperplane or a line that separates the classes. Widodo and Yang (2007) suggested the detail working of the SVM as following:

Firstly, sample dataset (medical reports) is assumed to have two classes ("Met" and "Not Met") and each class associated with labels, which can be "Met" (y_i =1) or "Not Met" (y_i = 0). If data is linear, it is possible to establish hyperplane $f(x) = 0$ such that it separates the provided data (medical reports). Equation 8 represents its mathematical formulation.

$$f(x) = W^T X + b = \sum_{j=1}^{M} w_j x_j + b = 0, \quad (8)$$

Where w is M-dimensional vector and b is a scalar.

Position of the separating hyperplane will be defined by vector (w) and scalar (b). Moreover, hyperplane will be separated by decision function which is made using sign $f(x)$. Constraints needs to be satisfied in distinct separating hyperplane. Mathematical formulation can be written as shown in equation 9 or 10.

$$\begin{aligned} f(x_i) = 1 \quad if \quad y_i = 1, \\ f(x_i) = -1 \quad if \quad y_i = -1 \end{aligned} \quad (9)$$

or

$$y_i f(x_i) = y_i(W^T X_i + b) \geq 1 \quad for \ i = 1,2,3, \dots, M \quad (10)$$

Create a maximum distance between the plane and the nearest data using separating hyperplane. For example, maximum margin is known as optimal separating hyperplane. Similar example of optimal hyperplane of two sets of data is presented in figure 5. There are two classes characterized with black square (negative) and while circle (positive).

The SVM attempts to spot a linear boundary between classes and adject it in a way that dotted line margin is maximized. Moreover, SVM tries to adject the boundary to confirm that the distance between the closest data point in each class and boundary is maximum. Then the boundary is located in the centre of between two margin points and this nearest data points utilized to distinguish margin, which are known as support vectors; these are embodied by grey circles and squares in figure 5. Once the support vector is selected, rest of the features are not necessary because support vectors can have all the information that are needed to define classifier.

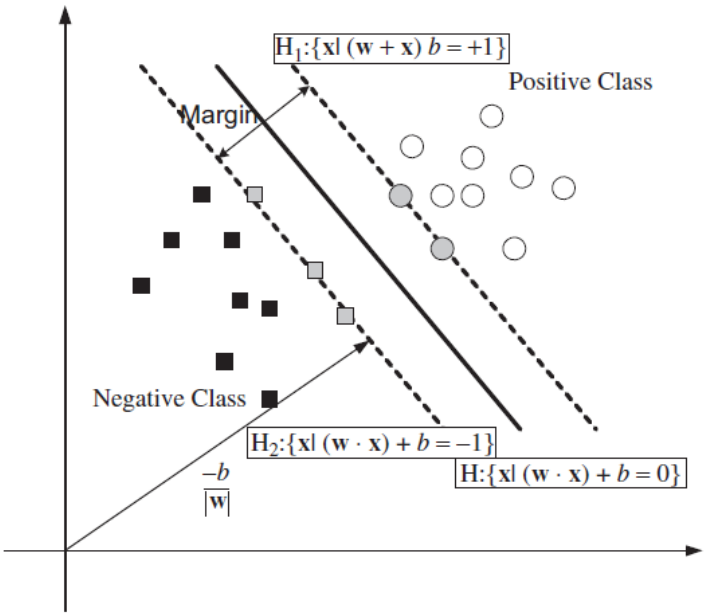


Figure 5 Support Vector Machine Example (Widodo & Yang, 2007)

4.3.5.3 Nave Bayes

According to Granik and Mesyura (2017), nave bayes classification algorithm is build based on the Bayes' Theorem, along with the assumption of independence between predictors. In other words, naïve Bayes classifier presumes that the availability of specific feature in a class is not related to the presence of any other attribute or feature.

As shown in equation 11, From $P(c)$, $P(x)$, and $P(x \mid c)$, Posterior probability $P(c \mid x)$ can be calculated using bayes theorem.

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)} \qquad (11)$$

Where

$P(c \mid x) = \ posterior\ probability\ of\ class(c, target) given\ predictor\ (x, attributes)$
$P(c) = \ prior\ probability\ of\ class.$
$P(x \mid c) = \ likelihood\ which\ is\ probability\ of\ predictor\ given\ class.$
$P(x) = \ prior\ probability\ of\ predictor.$

Working of this algorithm can be describe in 3 step processes as following:

- Generate frequency table from actual dataset
- Establish a likelihood table by calculating probabilities
- Calculate posterior probability for each class using Naive Bayesian equation (highest posterior probability class will be predicted class)

$$Pr(F \mid W) = \frac{Pr(W \mid F) * Pr(F)}{(\Pr(W \mid F) * \Pr(F) + \Pr(W \mid T) * \Pr(T))} \qquad (12)$$

Where

$Pr(F \mid W) =$ $conditional\ probability, that\ a\ patient\ is\ met\ given\ that\ word\ W\ apperes\ in\ medical\ report$

$Pr(W \mid F) =$ $conditional\ probability\ of\ finding\ word\ W\ in\ not\ met\ medical\ report.$

$Pr(F) =$ $Overall\ probability\ that\ the\ given\ medical\ report\ not\ met\ the\ criteria.$

$Pr(W \mid T) =$ $conditional\ probability\ of\ finding\ word\ W\ in\ met\ medical\ report.$

$Pr(T) =$ $Overall\ probability\ that\ the\ given\ medical\ report\ met\ the\ criteria.$

Equation 12 demonstrates mathematical formulation of nave bayes and this formula is derived from Bayes theorem.

4.3.5.4 Random Forest

Random Forest is a supervised algorithm, which uses an ensemble learning method. Generally, Ensemble learning is a method that combines various algorithms' predictions to make a single prediction; and this prediction is more accurate than the prediction made by a single model. (Liaw & Wiener, 2002) pointed out that at the time of model training, Random Forest builds numerous decision trees, and at the time of final prediction, it means the predictions of all trees; additionally, the following are the working steps of Random Forest Regression:

- From the actual dataset, draw N_{tree} Bootstrap samples.
- Grow an unpruned Regression tree for individual bootstrap samples.
- Make final prediction by aggregating the predictions from all N_{tree} Trees.

4.3.6 Model Evaluation

In order to identify which model, perform best among all, model evaluation and selection will be accomplished. All the four implemented models will be evaluated on the bases of three evaluation parameters, which are 3-Fold Cross Validation, Area Under the Curve, and F1 score.

First evaluation parameter is 3-Fold Cross Validation. Cross-validation can be defined as a resampling procedure which is used to evaluate machine learning models. In this procedure k is only the parameter that represents number of groups in which data sample will be split. Here we have considered k value as three hence it is known as 3-Fold Cross Validation. This evaluation method will assess how model is performing on the unseen data. Mostly widely this method is used in evaluation because it normally results into less biased. Figure 6 illustrate working diagram of 3-Fold Cross Validation and following is general workflow of it:

- Randomly shuffle dataset (set of medical report)
- Split data into three parts
- For each part follow following steps:
    - Consider selected part as testing data set
    - Consider remaining parts as training dataset
    - Build a machine learning model based on the training set and evaluate it based on testing set
    - Hold the evaluation score and remove a model
- Using model evaluation scores summarize the skill of the model.

Second evaluation parameter is Area Under Curve. AUC is used to represents quantify separability, which is calculated from Receiver operating characteristic (ROC) probability

curve. In other words, it tells how much model is capable to differentiate between two classes. ROC summarize all confusion matrixes that is produced by each threshold. Greater the AUC is better the model is. Following steps will be followed to calculate AUC:

- For all Machine Learning model plot ROC graph based on the true positive rate and false positive rate
- Calculate AUC from all plotted ROC graphs.
- Find name of the model which is having highest AUC value

Figure 6 3-Fold Cross Validation Working.

Third evaluation parameter is F1-Score. F-1 score is simply harmonic mean of precision and recall, as represented in equation 15. Recall is also known as true positive rate or sensitivity. Recall tells how many positive labels are predicted out of total number of positive labels exists. Equation 13 states that it is the ratio between true positive and true positive plus false negative. Precision tells how overall positive label is predicted. It is the ratio between true positive and true positive plus false positive. Mathematical formulation of Recall, Pression, and F1-Score are as following:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (13)$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (14)$$

$$F1\ Score = Harmonic\ Mean(Precision + Recall) \quad (15)$$

4.4 Objective Two (Cohort Visualization)

To visualize patients based on the cohort's following method described in figure 7 will be utilized. This methodology is designed in six phases, which are Data Cleaning, Date Time Extraction, Report Level Embedding, Patient Level Embedding, Dimensionality Reduction and

Cohort Visualization based on cohort. n2c2 dataset will be utilized for this objective and Data cleaning phase will be similar to the objective one (Criteria Prediction). Once the medical reports are cleaned, all the medical report need to separate based on the medical narratives. Next step to perform embedding on medical narratives using same TF-IDF and BERT model (generated by Orion Health), which was used in objective one (Criteria Prediction). Moreover, Medical narratives embedding will be converted into patient level embedding and from it relevant features will be identified based on criteria using Boruta feature selection (alike to criteria prediction). Lastly, to visualize patients' cohorts on two-dimension graph, dimension will be reduced using PCA and t-SNE.

Figure 7 Overall Methodology for Cohort Visualization

4.4.1 Date Time Extraction

Initial step of data cleaning will be parallel to objective one, which is criteria prediction. On the cleaned medical report, it is important to recognize dates because to perform patient level embedding, medical narratives embedding is essential and medical narratives can only be identified based on recorded date.

To extract dates from medical report user defined function was created using multiple libraries which are Med7, Spacy, and Date finder.

Working of this user define function is as shown in figure 8. Firstly, using Med7 library medical terminologies including drug names, route of administration, frequency, strength, duration, form, and frequency will be identified from cleaned medical report. Furthermore, based on the identified medical words check list will be created. Check list is nothing just a collection of all Med7 identified words.

Secondly, from cleaned medical report date will be extracted using Spacy library, but Spacy is unable to reorganize some dates and many medical terms has incorrectly identify as date. For example, "1.4 mg" is identified as date "1-4" (first day of April). To overcome this problem all

the Spacy identify dates will be verified in Check list, where if spacy identified data is in the check list, then it is not date else it is selected as date.

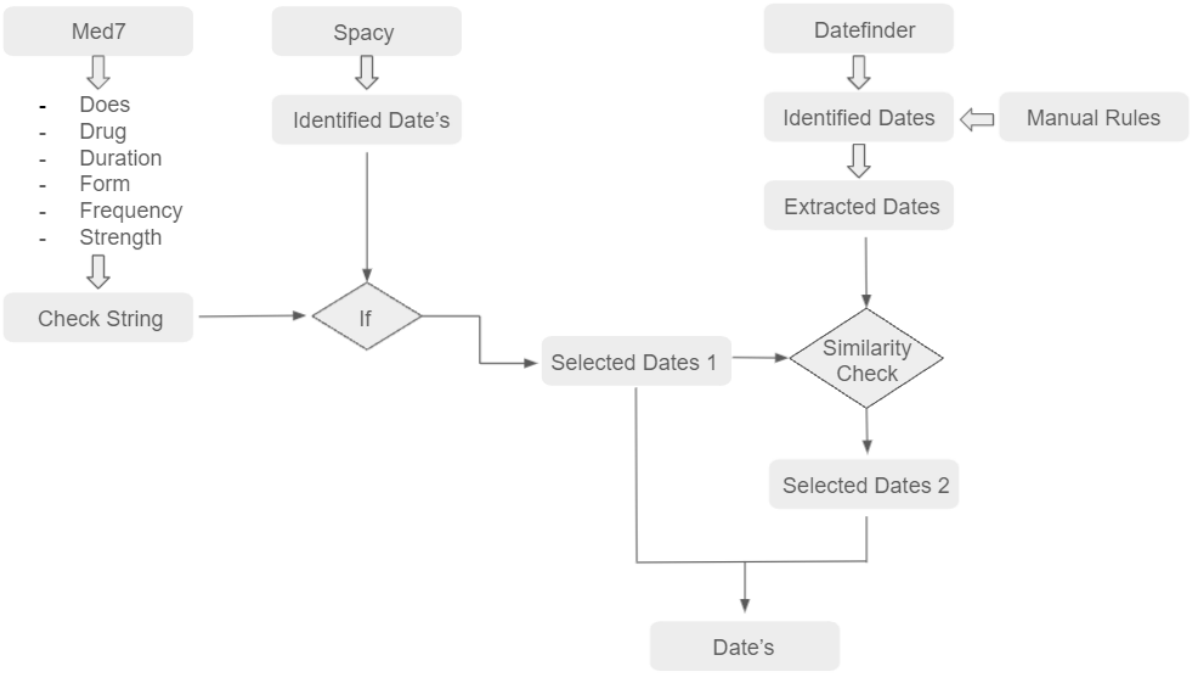


Figure 8 Workflow of Date Extraction

Thirdly, to identify remaining dates, all the dates were extracted from cleaned medical report using datefinder function.

Lastly, to check that whether dates identified dates are new or it's already been identified by Spacy similarity check will be run and if the similarity is less than threshold of 90 precent then it will be considered as new date that is not identify by spacy so it will be appended in spacy date identification list. Threshold of 90 percent is selected based on trial and error.

4.4.2 Medical Narratives Level Embedding, Patient Level Embedding, and Feature Selection

Based on the identified record dates medical narratives will separated from cleaned medical report. Moreover, on medical narratives report embedding will be applied using TF-IDF and BERT embedding. Next step is just converting this medical narrative embedding into patient level embedding, and to calculate it medical narrative embedding will averaged based on the patient id. Lastly, from patient level embedding significant features will be recognize for each criteria using same Boruta feature selection algorithm. TF-IDF, BERT, and Boruta Feature Selection same as objective one (criteria prediction).

4.4.3 Dimensionality Reduction

Dimensionality is major phase where target is to reduce multiple dimensions into two dimensions by capturing most of the variability using selected feature. In order to perform this reduction two widely used algorithms were consider, principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE).

4.4.3.1 PCA

(Lu et al., 2012) PCA is popular feature extraction technique that target to reduce the dimension by preserving as much as variance. Working of the PCA and its mathematical formulations are as following:

Goal is to use linear transformation *U* on input matrix features *X* (equation 16 and 18), which is in high dimension space *D* and target of this transformation is to generate matrix features *Z* (equation 17) into reduced feature space *M*.

$$X = [x_1\ x_2\ x_3\ \dots\ x_n]^T \quad X \in \mathbb{R}^{N*D} \tag{16}$$

$$Z = [z_1\ z_2\ z_3\ \dots\ z_n]^T \quad Z \in \mathbb{R}^{N*M} \tag{17}$$

$$X = UX \tag{18}$$

However, along with reducing dimension, information loss also needs to be minimized as mentioned in equation 19. Information is represented by covariance matrix S as shown in equation 20.

$$\max(u)\ U^T S_X U \quad U^T\ U = I \tag{19}$$

$$S_Z = \frac{1}{N}\ Z^T Z \quad S_Z \in \mathbb{R}^{M*M} \tag{20}$$

Equation 21 represent maximization problem has no upper bound on U, hence it can take any infinite value, but adding on condition that each vector in this matrix has a unique magnitude. This end up with optimization problem with quality constraints and it will be solved with Lagrange multiplier. After optimizing equation 21 it with Lagrange multiplier equation 22 is generated. In matrix equality this equation says that entry on one side equal to the other.

$$S_X U = \lambda U \tag{21}$$

$$S_X u_i = \lambda_i u_i \tag{22}$$

Equation 22 is similar to the eigenvector equation that will be solved by eigen decomposition of the covariance matrix (S_x), however, S_x is $d * d$ matric hence if diagonalizable it will lead to d pair of eigen vectors and eigen values. The concept of eigen decomposition involve splitting or decomposing a square matrix into a product of three matrices. Consider a square matrix S (equation 23), which is PDP invers as shown in equation 24 where P is matrix of eigen vector and D is diagonal matrix whose diagonal consist of eigen values as shown in equation 25.

$$S = PDP^{-1} \quad S, P, D \in \mathbb{R}^{N*N} \tag{23}$$

$$P = [u_1\ u_2\ u_3\ \dots u_n]\ u_i \epsilon\ \mathbb{R}^N \tag{24}$$

$$D = \begin{bmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_N \end{bmatrix} \tag{25}$$

This will Pairs of eigen values and vector. Say we want to find eigen values and eigen vector pair of *S* then they must satisfy the equation 26 where λ and u is one such pair; that represents the variance proportional to λ.

$$Su = \lambda u \tag{26}$$

Since the maximum number of eigen value and vector pairs for an $n * n$ matrix is n the sum of eigen values will corresponds to the entire variance of the transformation, which is shown in equation 27.

$$Total\ Variance = \sum_{i=1}^{N} \lambda_i \tag{27}$$

If we choice to project the data along the subset of this n vectors such as top D eigen vectors then the variance retain will be the sum of those eigen values (as show in equation 28).

$$Retained\ Variance = \sum_{i=1}^{D} \lambda_i \tag{28}$$

Hence the amount of information can be retained can be expressed as the percentage of the original by using formula 29.

$$\%_{info} = \frac{\sum_{i=1}^{D} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \tag{29}$$

It can be concluded that by picking the subset of eigen vector and values pair we are able to retain most information by only using fractional of the original dimension.

Eigen values are proportional to the variance retain so D eigen vector and eigen value pairs will be sorted based on eigen value and top M pair will be selected (as represented in equation 30, 31, 32 and 33).

$$(\lambda_1, u_1), (\lambda_2, u_2), (\lambda_3, u_3) \dots\ (\lambda_D, u_D) \tag{30}$$

$$\lambda_1 > \lambda_2 > \lambda_3 > \dots > \lambda_D \tag{31}$$

$$(\lambda_1, u_1), (\lambda_2, u_2), (\lambda_3, u_3) \dots\ (\lambda_M, u_M) \tag{32}$$

$$U = [u_1 \ u_2 \ u_3 \dots u_M]^T \quad (33)$$

Final equation 34 will be generated where D dimension vector, x will be transformed into M dimensional vector Z, and still, it can retain variance of data.

$$Z = U^T x \quad (34)$$

Where

$x \in \mathbb{R}^D$

$z \in \mathbb{R}^M$

$D \gg M$

Information is represented by covariance matrix S. As showing in equation 34, using a linear transformation U on input matrix features X in high dimension space D. Target is to transfer it into some matrix Z into reduced feature space M.

4.4.3.2 t-SNE

(Van Der Maaten, 2014) t-Sne is a non-linear data visualization method, which have capability to generate two-dimension visualization from high-dimension data. Considering the high dimension dataset $D = \{X_1, X_2, X_3, \dots, X_N\}$ and a function $d(X_i, X_j)$, which is used to calculate distance between pair of objects. Objective is to discover s-distribution embedding in which individual objective is characterized by a point by $E = \{Y_1, Y_2, Y_3, \dots, Y_N\}$ with $Y_i \in R$. Hence, by symmetrizing two conditional probabilities t-sne will outlines joint probabilities $P_{i,j}$ that calculate the pairwise similarity among objects (X_i and X_j).

$$p_{j|i} = \frac{\exp\left(-d\left(X_i, X_j\right)^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-d(x_i, x_k)^2 / 2\sigma_i^2\right)}, \qquad p_{i|i} = 0 \quad (35)$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}, \quad (36)$$

In equation 35 and 36, gaussian kernel bandwidth α_i is calculated, by making sure that Perplexity of the conditional distribution P_i is equal to predefine perplexity μ. Due to this condition optimal value of α_i differ per objects and its optimal value can be calculated utilizing robust root-finding method or simple binary search. When it comes to s-dimensional embedding E, similarity among two points (y_i and y_i) are calculated by applying normalized heavy tailed kernel. As shown in equation 37, similarity of embedding $q_{i,j}$ among y_i and y_j will be calculated with single degree of freedom of normalized student-t kernel.

$$q_{ij} = \frac{(1 + ||y_j - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}}, \qquad q_{ij} = 0 \quad (37)$$

Dissimilar input objects (X_i and X_j) will be shaped by low dimensional objects y_i and y_j using heavy tails of the normalized student-t kernel. This will directly affect in the low dimension embedding where space will be created that can accurately model the small-scale pairwise distances. In order to locate the embedding points (y_i) Kullback-Leibler divergence among the

joint distribution P and Q will be minimized. It mathematical formulation can be written as shown equation 38.

$$C(\varepsilon) = KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (38)$$

Equation 39 denotes that objective function focused on modelling extreme values of similar objects ($p_{i,j}$) by extreme values of neighbouring points in embedding space ($q_{i,j}$), because of irregularity of the Kullback-Leibler divergence.

$$\frac{\partial C}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij}) q_{ij} Z(y_i - y_j) \quad (39)$$

4.4.4 Cohort Visualization based on patient:

Once dimensionality is reduced to two dimensions using PCA and t-SNE, then both of the reduced dimension set will be plotted, and cohort will be interpreted based on patient. This visualization will be created for all criteria and lastly, depending on the cohorts best dimensionally reduction technique will be concluded.

4.5 Objective Three (Impact Analysis)

To identify whether impact analysis can be used in Smart Cohort or not, first casual inference problem will be formulated from MIMIC-III dataset and then it will be solved by the DoWhy. Figure 9, explained the methodology used to undertake impact analysis using DoWhy library.

Figure 9 Working flow of DoWhy Library

There are total four phases, which are Model Casual Question, Identify the Causal Estimated, Estimate Casual Effect, Refute the Obtained Estimate:

First step is generating a model casual question and then depending on the problem and prior knowledge casual graphical model needs to be created. This graph does not require to be complete one, it can be partial graph generated based on the prior knowledge about the features. Uninvolved feature will automatically consider as potential confounders.

Second step is to identify casual estimate. Depending on the casual graph all possible ways will be identified for generating required causal effect. To identify this causal effect, do-calculation and graph-based criteria will be used. Following are the identification criteria supported in Dowhy:

- Back-door
- Front-door
- Instrumental Variables
- Direct and indirect effect identification

Third step is to estimate casual effect. Objective of this phase is to estimate casual effect using statistical models. Method developed based on instrumental variables, and back-door criterion will be used. Lastly, significant of the obtained estimation will be tested by permutation test and non-parametric confidence intervals. Following are the estimation method supported in Dowhy:

- Linear Regression
- Generalized Linear Models
- Binary Instrument
- Two Stage Linear Regression
- Propensity Score Matching

Final step is refuting the obtained estimation. Once the casual effect estimation is produced by statistical model it important to verify and validate it. There are wide range of method supported in Dowhy library for validating effects of estimate from a casual estimator. Following are the refuting methods supported in Dowhy:

- Add Random Common Cause
- Placebo Treatment
- Dummy Outcome
- Simulated Outcome
- Add Unobserved Common Cause
- Data subset validation
- Bootstrap Validation

Once the casual inference problem has been solved using Dowhy Library then similar solution will be suggested in terms of Smart Cohort Project. Else, Dowhy library could be not useful in Smart Cohort Project.

5. Analysis and Findings

This section will describe the key findings of three objectives, which are criteria prediction, cohort visualization based on patients, and impact analysis.

5.1 Objective One (Criteria Prediction)

This objective aims to build a machine learning model that can predict whether patient has met with selected criteria or not. As show in table 2, it was found that model build using random forest perform well on eight criteria. This performance is evaluated using baseline research (result 1); to be more specific, if 3-Fold Cross Validation, and F1 Score is higher than (Antunes et al., 2019) study then model perform well on those criteria.

Criteria	3FCV(F1)	3FCV - Baseline	F1-Test	F1 – Baseline	AUC
Major-Diabetes	0.88	0.76	0.86	0.85	0.89
ASP-For-MI	0.82	0.52	0.88	0.59	0.5
Dietsupp-2MoS	0.77	0.67	0.74	0.70	0.72
Advance-CAD	0.86	0.73	0.86	0.84	0.89
HBA1C	0.58	0.61	0.76	0.60	0.79
Makes-Decisions	0.98	0.55	0.99	0.73	0.61
Abdominal	0.32	0.62	0.52	0.82	0.5
Alcohol-Abuse	0.27	0.48	0.07	0.58	0.49
Creatinine	0.8	0.8	0.65	0.72	0.78
Drug-Abuse	0.58	0.70	0.5	0.65	0.79
English	0.99	0.91	0.95	0.78	0.71
KETO-1Yr	-	-	-	-	-
MI-6MOS	0.48	0.56	0.5	0.48	0.66

Table 2 Comparison of Criteria Prediction Results with Baseline Research

5.1.1 Major Diabetes (Patient having Diabetes)

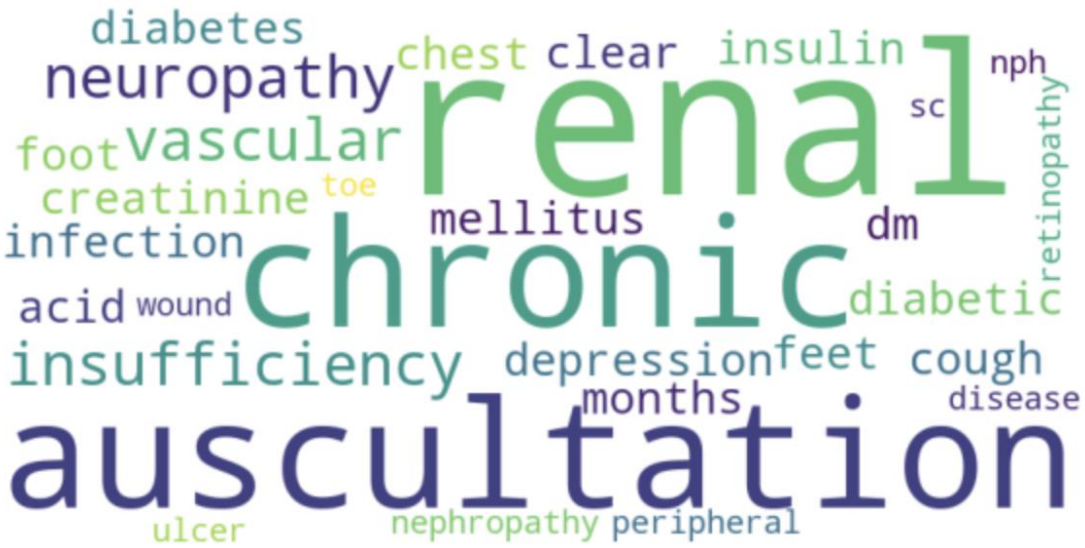
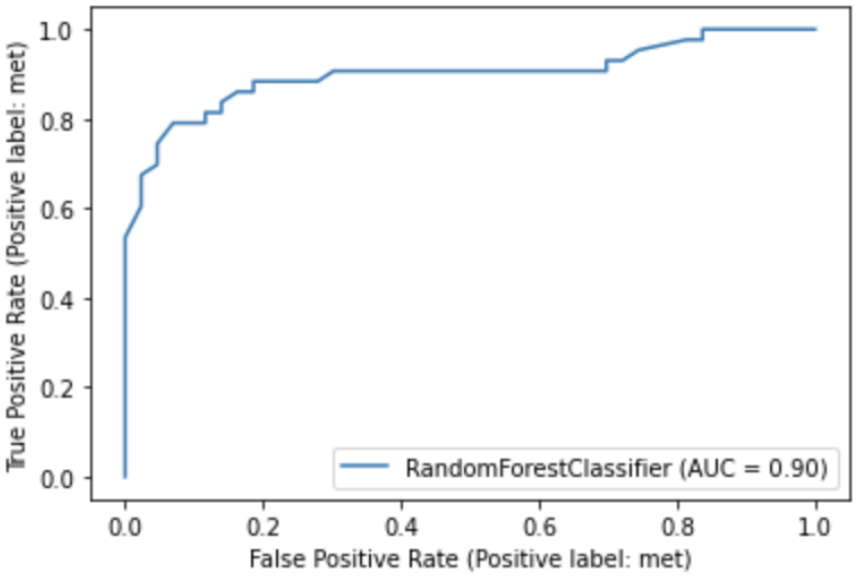


Figure 10 Word cloud representing features of Major Diabetes model

Figure 10 demonstrates the important words related to the major diabetes criteria. It was found that words such as “Diabetic”,” Insulin”, “Chronic”, ”Insufficiency” and ”Auscultation” are frequently been used in medical report if the patient is having diabetes; hence random forest model was build using these significant words, which are show in word cloud (Figure 10).




```
F1 score Traning Dataset using 3 Fold CV: 0.9
F1 score Testing Dataset: 0.85
```

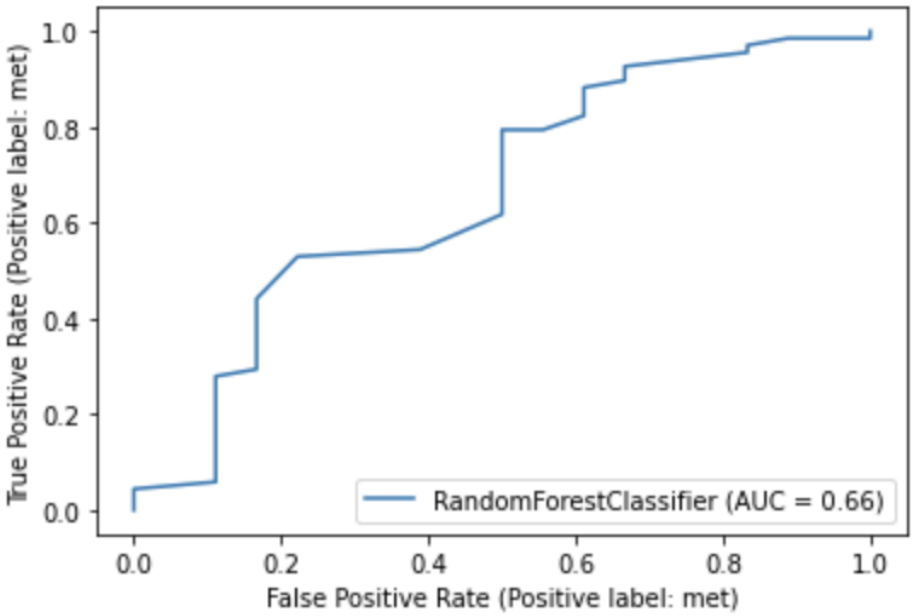
Figure 11 Evaluation results of Major Diabetes

Figure 11 exemplifies the performance of major diabetes criteria prediction model, which was build using random forest. According to the statistics, of ROC curve, calculated AUC is 0.90, which means this model is able to distinguish diabetes and non-diabetes patients, however, some time it does fail, which can be overlooked. Moreover, 3-Fold Cross Validation score was 0.90 conclude that model perform well on unseen data and 0.85 F1 score determine that model is also good at predicting diabetes patients.

5.1.2 ASP-For-MI (Use of aspirin to prevent myocardial infarction)

Figure 12 Word cloud representing features of ASP-For-MI model

Figure 12 indicates the vital words frequently used in medical report if the patient is meeting ASP-For-MI criteria. Words including “Chest”, “Pain”, “Seizure”, “Allergies”, and “Mellitus” are noticed in ASP-For-MI report. Using all these words random forest machine learning model was created to predict ASP-For-MI criteria.




```
F1 score Traning Dataset using 3 Fold CV: 0.99
F1 score Testing Dataset: 0.89
```

Figure 13 Evaluation results of ASP-For-MI

Figure 13 illustrate how ASP-For-MI criteria model perform while predicting whether patient has meet with ASP-For-MI criteria or not. 0.66 AUC suggest that model is able to distinguish between ASP-For-MI and non-ASP-For-MI patients, additionally, this model also perform excellent on unseen data with 0.99 3-Fold validation score. Lastly, it was able to predict most of the ASP-For-MI patients with 0.89 F1-Score.

5.1.3 Dietsupp-2MoS (Use of dietary supplements in the last two months)

Figure 14 shows frequently mentioned words in Dietsupp-2MoS medical report. It can be concluded that word such as “Chest”, “Pain”, “Calcium”, “Lives”, “Vitamin”, and “extremities” are closely related to the Diestupp-2MoS medical reports. Random forest model was fitted based on the words shown in word cloud (Figure 14).

Figure 15 validate the performance of Dietsupp-2Mos criteria model, which will predict that whether the patients has been met with the Dietsupp-2MoS criteria or not. By analysing evaluation parameters of this model, it was found that with 0.72 AUC it is able to clearly sperate Dietsupp-2Mos and non-Dietsupp-2MoS patients. Additionally, 0.74 3-Fold cross validation score and F1 Score concluded that model is working admirable on unseen data and also able to predict majority of Dietsupp-2Mos patients.

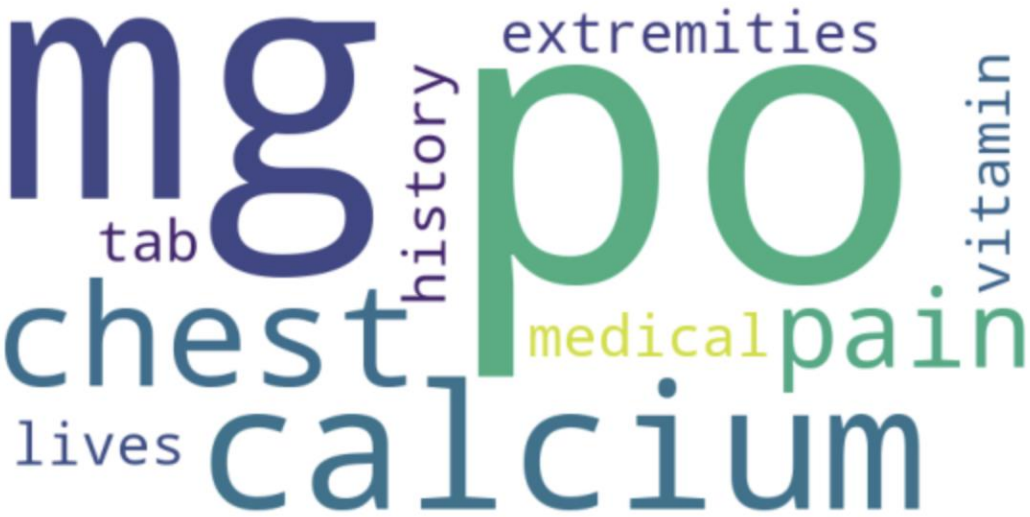
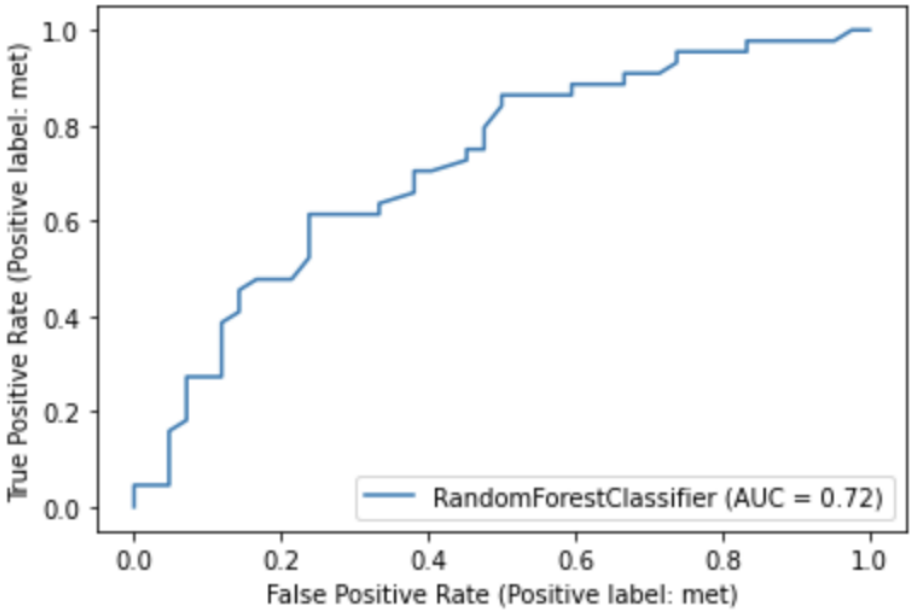


Figure 14 Word cloud representing features of Dietsupp-2MoS model




```
F1 score Traning Dataset using 3 Fold CV: 0.74
F1 score Testing Dataset: 0.74
```

Figure 15 Evaluation results of Dietsupp-2MoS

5.1.4 Advance CDA (Advanced cardiovascular disease)

Figure 16 reveals significant words that are being used in Advanced-CDA report. By examining word cloud closely related words to Advanced-CDA are identified which includes “Artery”, “Catheterization”, “Coronary”, “Cardiac bulk”, and “ischemia”. Based on these recognized words machine learning model will be created for predicting Advanced-CDA criteria.

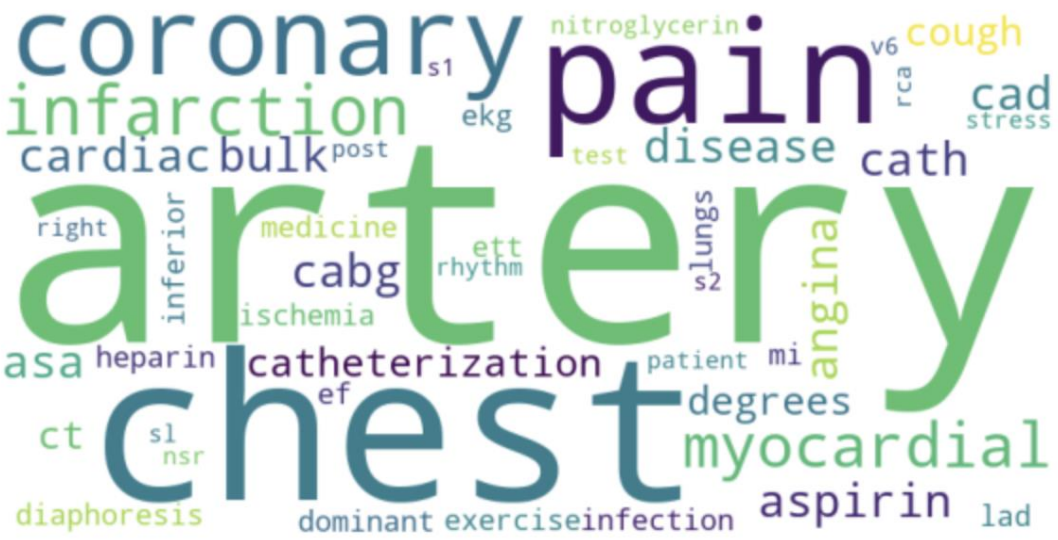
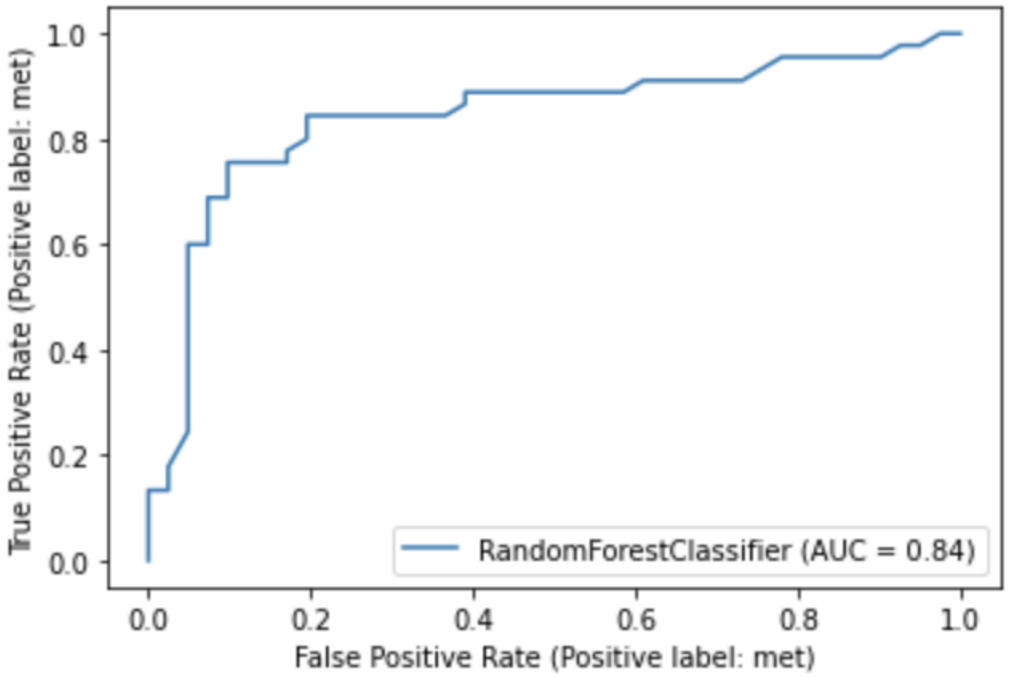


Figure 16 Word cloud representing features of Advanced-CAD model

Figure 17 measures, how well Advanced-CDA model perform while predicting patient being met with Advanced-CDA or not. It is clearly seen that AUC, 3-Fold Cross Validation, and F1-Score is higher than 0.80, which mean model is performing outstanding on class distinguishing, unseen data and predicting true value, which is met label.




```
F1 score Traning Dataset using 3 Fold CV: 0.9
F1 score Testing Dataset: 0.84
```

Figure 17 Evaluation results of Advanced-CAD

5.1.5 Makes Decisions (The patient can make decisions by himself)

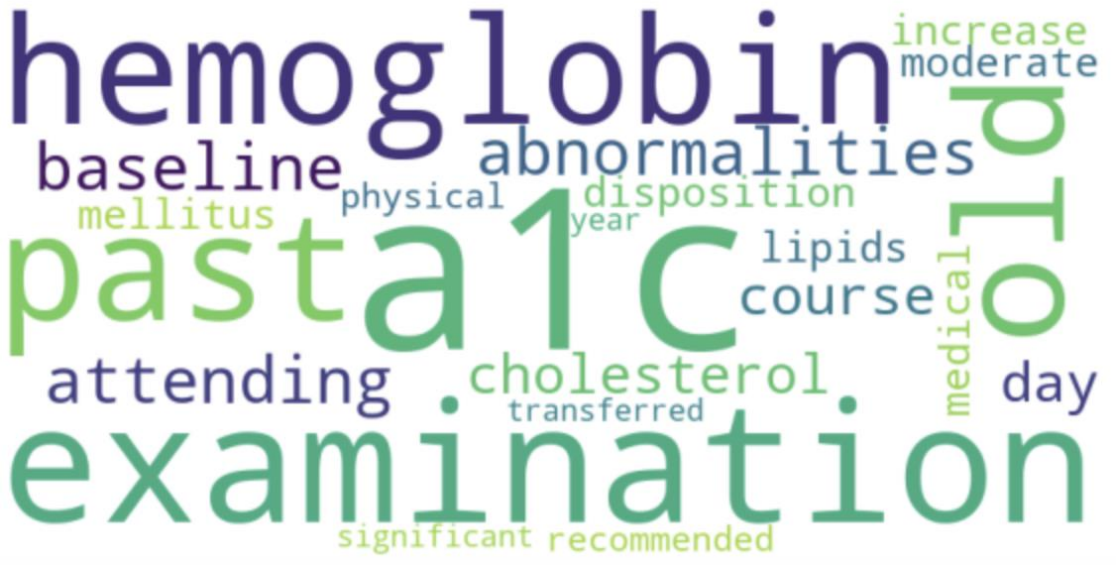
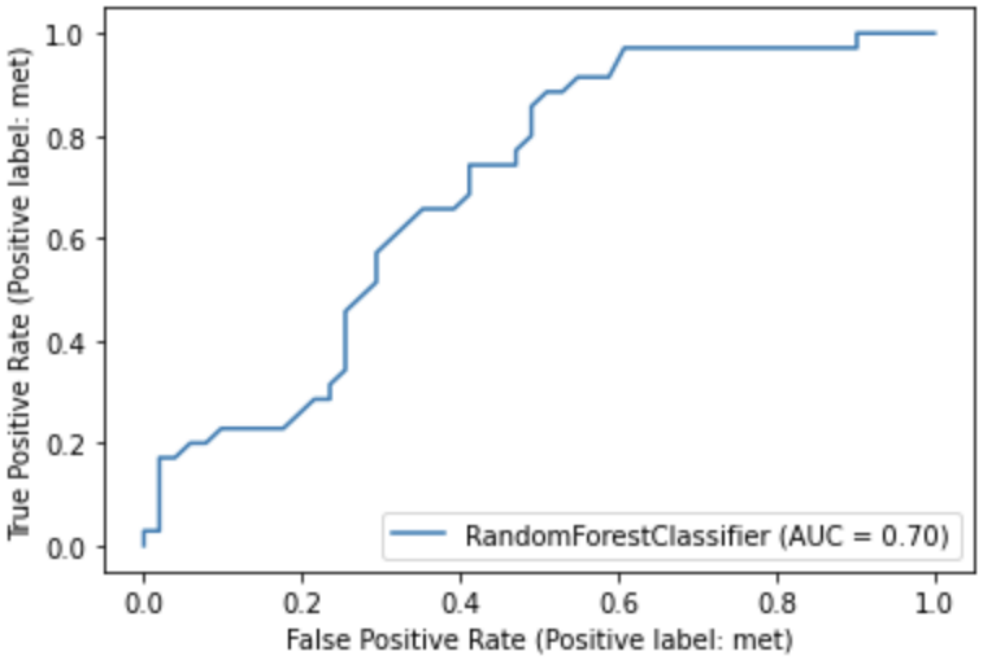


Figure 18 Word cloud representing features of Makes decision model

Figure 18 demonstrates significant words that will help to identify whether patient is able to make their own decision or not. “Examination”, “Attending”, “Past”, “Transferred”, “Old”, and “moderate” are few of the highly related words found in patients medical reports which is classified based on making decision tag. These significant words are utilized to create Making Decision criteria model.




```
F1 score Traning Dataset using 3 Fold CV: 0.83
F1 score Testing Dataset: 0.68
```

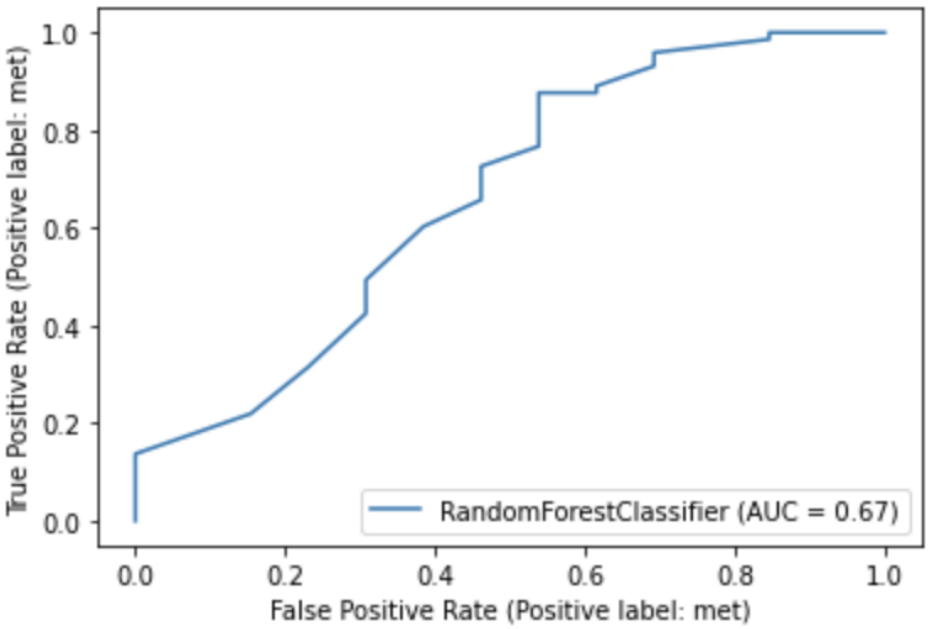
Figure 19 Evaluation results of Makes decision

Figure 19 represent performance of making decision criteria prediction model. Linear graphs represent the ROC curve from which AUC is calculates as 0.70. From this AUC, it can be concluded that model is able to differentiate patients who can make decision and who cant. Moreover, 3-fold cross validation and F1-Score are 0.83 and 0.68, respectively. By interpreting these two parameters it was found that model perform well on unseen data, but some time it is unable to identify patients who can make decision.

5.1.6 English (The patient can speak English)

Figure 20 Word cloud representing features of English model

Figure 20 exhibits word cloud of the words that are closely related to the patient who can speak English. By examining the word cloud, it was found that many English word such as “Note”,”Auscultation”,”Visit”,”Day”, and ”Date” can be used to found that patient can speak English or not. English criteria prediction model will be created based on this word cloud.



```
F1 score Traning Dataset using 3 Fold CV: 1.0
F1 score Testing Dataset: 0.93
```

Figure 21 Evaluation results of English

Figure 21 indicates the performance of English speaker criteria prediction model that was build based on the random forest. From the ROC curve, AUC was calculated as 0.67, which indicates that classification group are not well clearly distinguished, however, it does perform magnificent on classifying English speaking patient, and unseen data.

5.2 Objective Two (Cohort Visualization based on patients)

Objective two is to visualize cohort based on the patient, for this visualization two dimensionality reduction methodologies (Principal Component Analysis (PCA), and t-Distributed Stochastic Neighbor Embedding (t-SnE)) were used. By performing experiment on two criteria (Major Diabetes and Advanced-CDA) it was found that t-SnE was able to clearly sperate cluster based on the patients. Given below is the visualization for both criteria.

5.2.1 Major Diabetes (Patient having Diabetes)

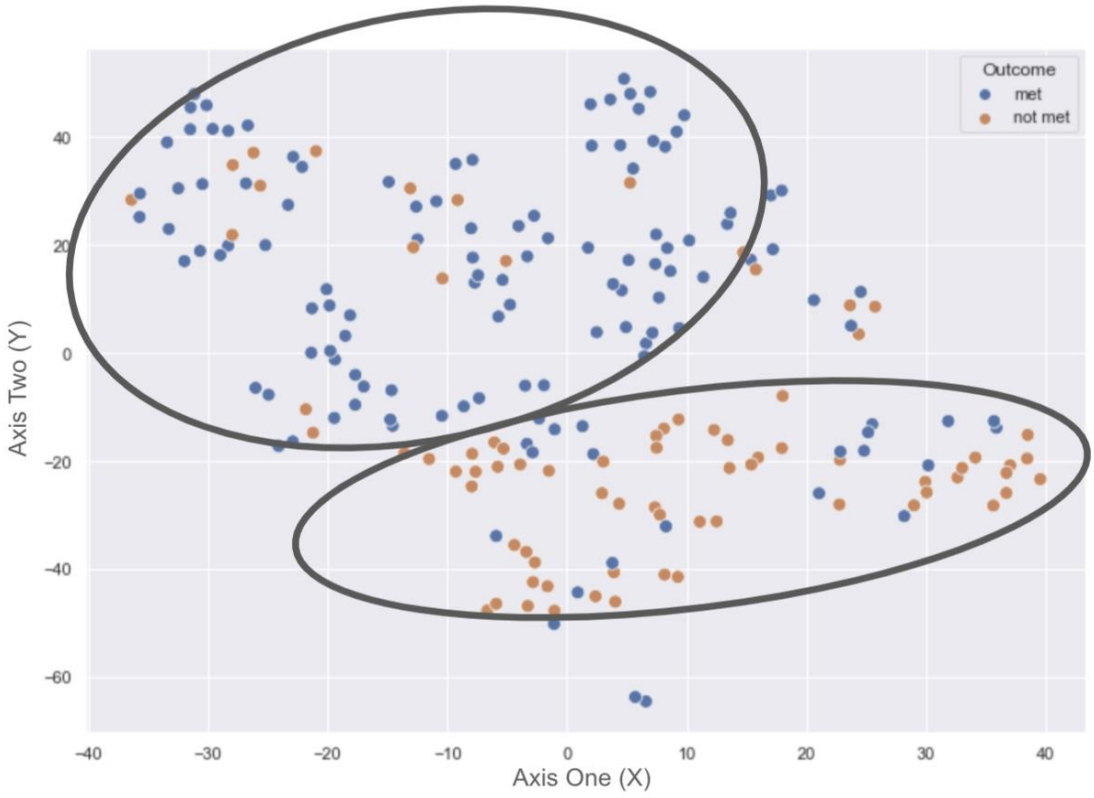


Figure 22 Cohort Visualization of Major Diabetes

Above scatter plot (Figure 22) present Major Diabetes cohort visualization depending on the patients. After reducing high dimension into two dimensions, it was plotted, and points were coloured based on the Outcome. Each point represented as a unique patient; hence list of

patients can be extracted based on the Cohort cluster. Majority of patient are clustered correctly.

5.2.2 Advanced-CDA (Advanced cardiovascular disease)

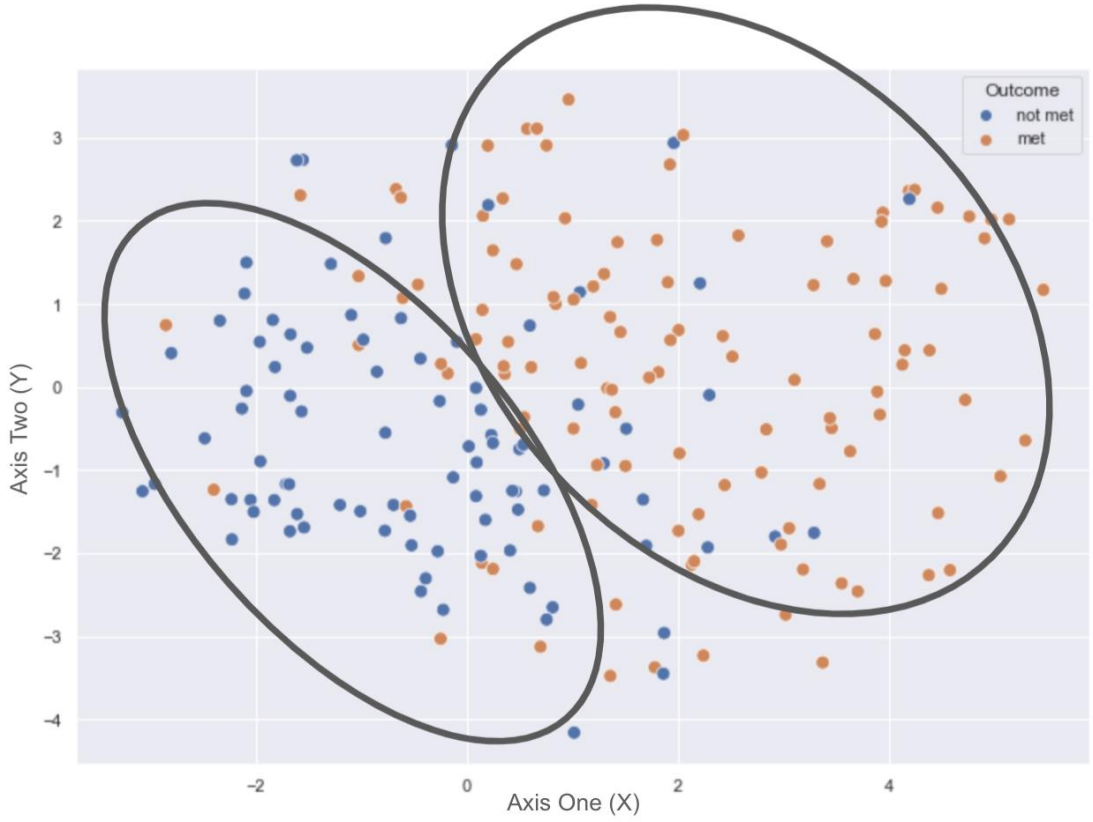


Figure 23 Cohort Visualization of Advanced-CAD

Figure 23 demonstrates Advanced-CDA cohort visualization depending on the patients. First, high dimension was reduced into two dimensions then it was plotted, and points colour are based on Outcome. Majority of patient are clustered correctly and individual point characterized as a unique patient, which means list of patients can be extracted based on the Advanced-CDA Cohort cluster.

5.3 Objective Three (Impact Analysis)

This objective focused on experimenting impact analysis using DoWhy casual inference library. Key finding from this experiment is to identify that whether casual inference can be used in Smart Cohort project or not. If it can then it will also identify that how it will be useful. Following two small experiments will be conducted using DoWhy library: Selecting five Diabetes Patients based on Glucose and Cholesterol abnormal event (experiment one), and selecting five diabetes patients based on top five abnormal lab events (experiment two).

5.3.1 Experiment One – Selecting Five Diabetes Patients based on Glucose and Cholesterol Abnormal event

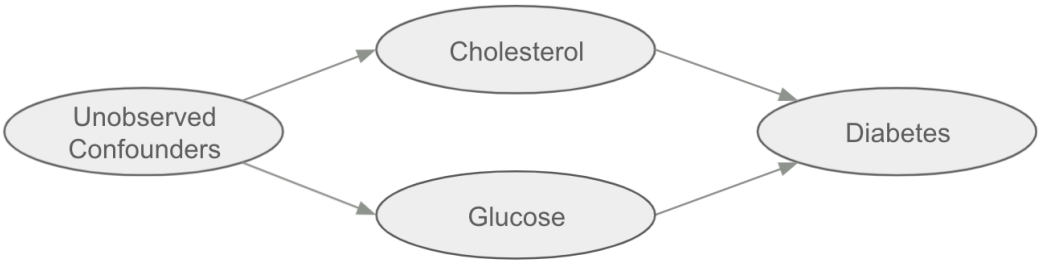


Figure 24 Casual inference example one graph

Figure 24 exemplify, casual inference problem where it says that presence of high glucose and cholesterol can lead to diabetes. Additionally, cholesterol and glucose can be affected by some unobserved confounders. As shown in figure 25, estimated causal effect found using linear regression is 55% and this was verified using random common cause method.

```
Estimated effect by linear regression: 0.5460552523374468
Estimated verification effect by random common cause: 0.5431703132078716
```

Figure 25 Casual inference estimation of example one

It can conclude that if the patient contains abnormal event related to the cholesterol and glucose then probability of patient being diabetes will increase by 55%.

This casual inference can be use in Smart Cohort. For example, using Smart Cohort ten diabetes patient id's are identified, but actually clinical trial only require five of them so to reduce the probability of selecting incorrect patient casual inference can be used and probability of each patient can be plotted as show in figure 26 This graph will help in selecting patients' id's who has higher probability of been diabetes.

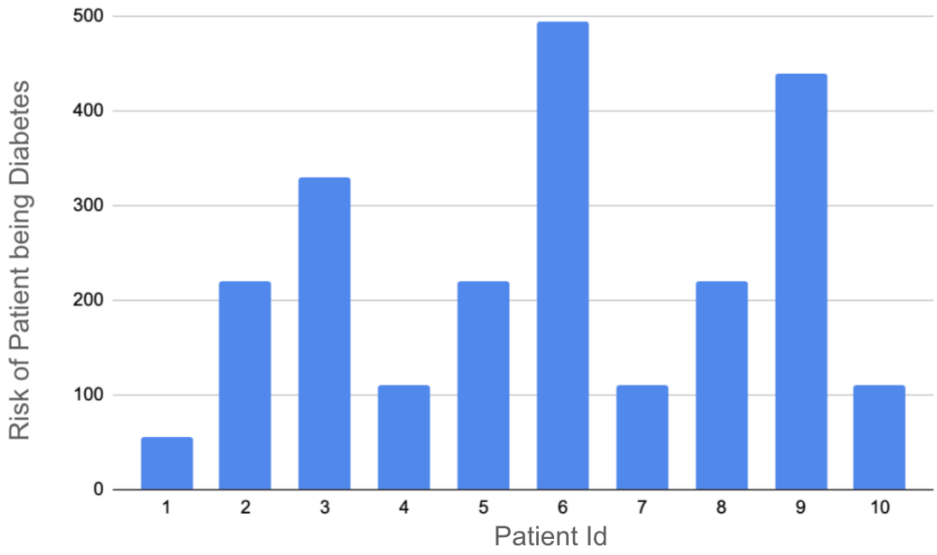


Figure 26 Casual inference example one bar graph for selection patients

5.3.2 Experiment Two – Selecting Five Diabetes Patients based on Top five Abnormal Lab Events

For verification similar experiment was performed, but this time casual problem establish is based on the top five abnormal lab event where we assume that this lab event can identify diabetes patients.

First top five abnormal lab events were identified in diabetes patients. This identification was estimated based on the occurrence of the abnormal lab event and it was found that haematocrit, haemoglobin, glucose, red blood cell, and urea nitrogen are most occurring abnormal lab event.

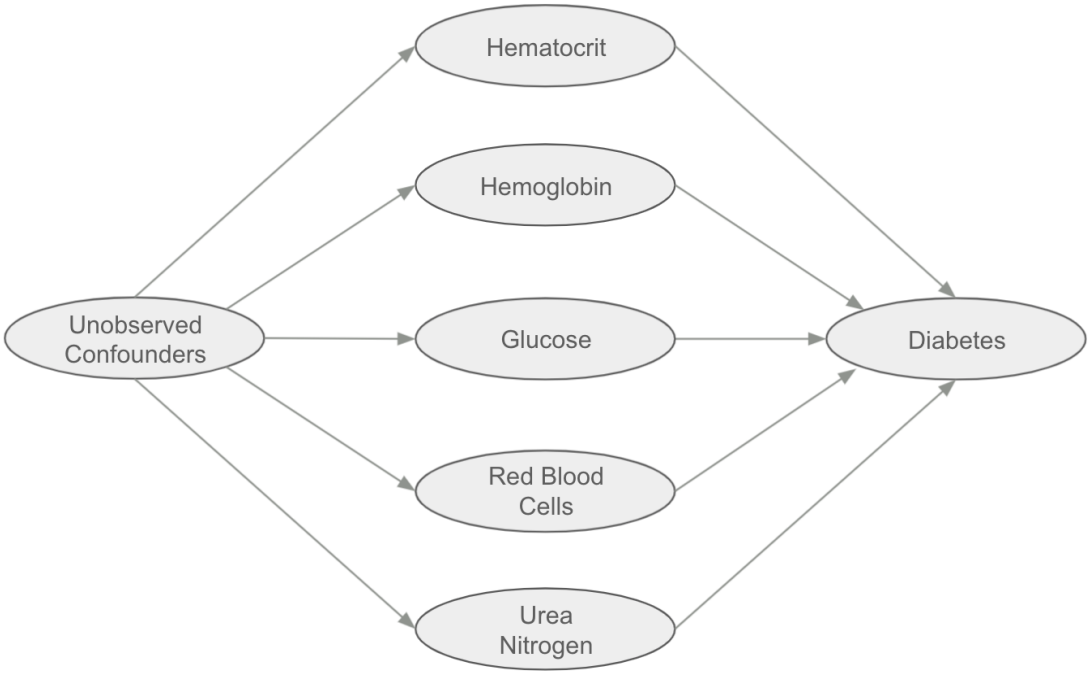


Figure 27 Casual inference example two graph

On the bases of these five abnormal lab events casual inference problem is created (as shown in figure 27). It was assumed that all these abnormal lab events are also affected by unobserved confounder.

```
Estimated effect by linear regression: 0.5067857027252
Estimated verification effect by random common cause: 0.494872095770
```

Figure 28 Casual inference estimation of example two

Figure 28 describe estimated casual effect and verified casual effect. Using linear regression estimated casual effect is approximately 0.50 and verified casual effect using random common cause is more or less similar to actual estimation. Hence it can be conclude that a increase in that haematocrit, haemoglobin, glucose, red blood cell, and urea nitrogen abnormal event can lead to the increase in probability of patient being diabetes by roughly 55%.

This casual inference can be use in Smart Cohort. For example, using Smart Cohort ten diabetes patient id’s are identified, but actually clinical trial only require five of them so to reduce the probability of selecting incorrect patient casual inference can be used and probability of each

patient can be plotted as show in figure 29. This graph will help in selecting patients id's who has higher probability of been diabetes.

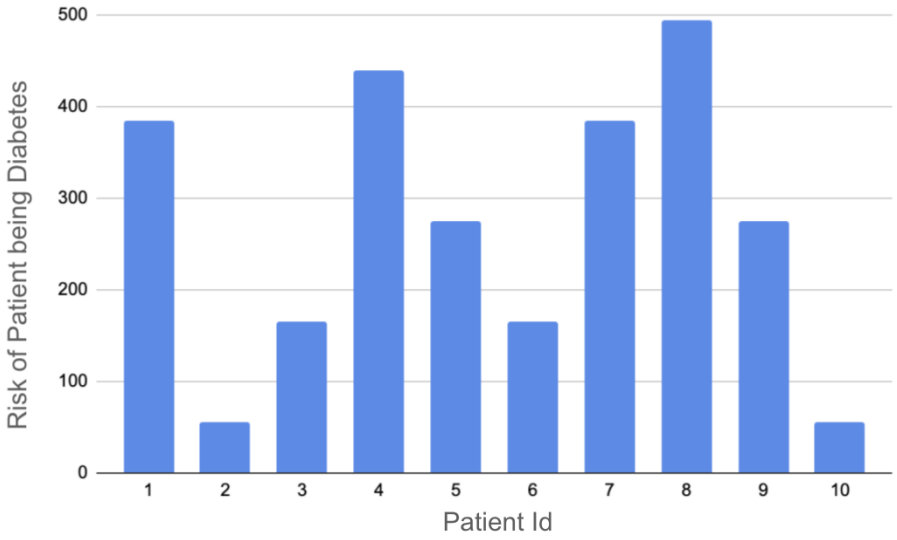


Figure 29 Casual inference example two bar graph for selection patients

6. Discussion

6.1 Overview:

This study was conducted to build a machine learning model for criteria prediction and cohort visualization. Additionally, also focused on identifying use of casual inference in Smart Cohort. 202 medical reports along with 887 medical narratives were consider from n2c2 dataset. This discussion will investigate the performance of newly developed methodology of this study in comparison with baseline research ((Antunes et al., 2019)). It will also examine best model for criteria prediction and cohort visualization. Lastly, it will explore how casual inference can be used in Smart Cohort.

This research finds that proposed method performs excellent in comparison with baseline study ((Antunes et al., 2019)); additionally, best machine learning algorithm for criteria prediction is random forest with highest 3-Fold Cross Validation score, F1-Score, and AUC. Moreover, t-SnE is concluded as better then PCA in terms of visualizing cohort based on patient. Lastly, casual inference can be used in Smart Cohort for identifying patient with highest probability from given list of patients who has met with selected criteria.

6.2 Does suggested methodology for criteria prediction perform well in comparison to base line research ((Antunes et al., 2019)):

According to the findings and results, in comparison to the baseline research suggested methodology performs well on most of the criteria's. In other words, suggested approach for criteria prediction in this paper provide higher 3-Fold Cross Validation Score and F1-Score in comparison with baseline research ((Antunes et al., 2019)). Moreover, it also provides better AUC score. To be more specific, Major-Diabetes, ASP-For-MI, Dietsupp-2Mos, Advance-CDA, HBA1C, Makes-Decisions, English, and MI-6MOS criteria perform terrific. On the other hand, Abdominal, Alcohol-Abuse, Creatinine, Drug-Abuse, and KETO-1Yr

outperformed while evaluating performance based on (Antunes et al., 2019) study, but it was noticed that (Antunes et al., 2019) experiment is using rule-based algorithms for these criteria and their machine learning algorithm is also not performing up to the expectation. Table 3, compares best performing criteria's results with the (Antunes et al., 2019) baseline research results.

Criteria's	3FCV – Baseline	3FCV	F1-Score - Baseline	F1-Score
Advanced-CDA	0.76	0.86	0.82	0.86
ASP-For-MI	0.58	0.82	0.58	0.88
Dietsupp-2MOS	0.73	0.77	0.69	0.74
HBA1C	0.57	0.58	0.62	0.76
English	0.88	0.99	0.79	0.95
Major-Diabetes	0.74	0.88	0.87	0.86
Makes-Decisions	0.75	0.98	0.83	0.99
MI-6MOS	0.59	0.48	0.47	0.5

Table 3 Best performing criteria's in comparison with baseline research

From above table (Table 3) it can be clearly concluded that 3-Fold Cross Validation score and F1 score for all eight criteria are much higher than the (Antunes et al., 2019) baseline research score (3FCV and F1-Score). Moreover, from these eight criteria except Major-Diabetes criteria, for all criteria including English and Advanced-CDA, best performing algorithm depending on 3-Fold Cross Validation is different and best performing algorithm depending on F1-Score is different but using random forest algorithm with methodology proposed in this paper will provide both higher 3-Fold Cross Validation score and F1-Score. Hence, it can be concluded that using baseline research method (Antunes et al., 2019), in most of the case machine learning algorithm perform well on unseen data may not perform excellent on testing dataset, and visa-versa, however, using methodology generated in this paper will provide high standard performance on both unseen data and testing data.

6.3 Best fit machine learning algorithm for criteria prediction:

According to (Antunes et al., 2019), with higher 3-Fold Cross Validation and F1-Score, rule-based algorithm performs best in comparison with other algorithm such as Decision Tree, Bagging, Ada Boost, and Neural Networks. Additionally, restring it to classically machine learning models only shows that depending on criteria best fit model do vary. Similarly, D2 also acquired measurable result using rule-based algorithm. Likewise, (Oleynik et al., 2019) perform experiment using rule-based classifier, shallow methods (such as Support Vector Machine and Linear Regression), and Long Short-Term Memory as a conclusion higher F1-Score was measured by rule-based algorithm. However, in this study based on 3-fold cross validation and F1 score it was found that random forest was best fit algorithm. This contradicts suggest that there might be various rule based and machine learning algorithms that fit well for criteria predictions, still it was concluded that deep learning-based algorithms are outperforming.

6.4 Best dimensionality reduction methodology for cohort visualization:

According to figure 22 and 23, it can clearly conclude that for criteria Diabetes and Advanced-CDA, t-SnE is clearly able to distinguish patients based on the cohorts, whereas PCA fails in

generating clear cohorts. Similarly, (Diaz-Papkovich et al., 2019) results concluded that in comparison with PCA, t-SnE can clearly visualize overlooked subpopulations in genomic cohorts; parallelly, (Zhan et al., 2021) approached a t-sne based method for cohorts' visualization. However, (Diaz-Papkovich et al., 2019), concluded that Uniform manifold approximation and projection (UMAP) shows clear cohorts then t-SnE; also, (De Freitas et al., 2021), visualize clear cohorts using UMAP. Overall, it can be concluded that t-sne perform well on cohort visualization in comparison to PCA, but UMAP can provide even better than tsne.

6.5 Can casual inference be used in Smart Cohort; if yes, then how?

According to example provided in 5.3.1 and 5.3.2, it can be concluded that casual inference can be used in Smart Cohort project. Experiments perform on MIMIC-III dataset using DoWhy library says that machine learning model that are created can only classify patient being met with specific criteria or not, but which patients have highest probability of meeting the criteria? In order to target this question casual inference can be used and probability can be calculated. Lastly based on likelihood, probabilistic bar graph can be generated as shown in Figure 26 and 29, which can clearly identify patients who has higher chance of meeting criteria.

7. Conclusion, Limitations and Future Work

This research aims to build a machine learning model for criteria prediction and cohort visualization machine. Additionally, study also target to identify that whether casual inference can be used in Smart Cohort or not. To undertake this research two researching dataset (n2c2 and MIMIC-III) has been used.

Fresh methodology is suggested for criteria prediction model. Using this methodology and various machine learning algorithm (Random Forest, Logistic Regression, and Support Vector Machine) numerous models were trained. Furthermore, with highest 3-Fold Cross Validation, F1 Score and AUC it was concluded that random forest fit well and work excellent on eight criteria's including Diabetes and Advanced-CDA.

Cohorts were visualized based on patients using two dimensionality reduction methods, which are PCA and t-SnE; as a result, t-SnE was able to clearly visualize cohorts and only few patients are in incorrectly grouped in cohort.

Impact analysis experiments were performed using casual inference library named DoWhy. Result of the trials suggested that casual inference can be used in calculating probability of patient been met with selected criteria and using this likelihood, probabilistic bar graph can be generated that helps in identifying the patients who has higher chance of meeting criteria.

In terms of smart cohort this research has provided more accurate methodology for criteria prediction and cohort visualization; in addition to it, study also provided clear pathway for using casual inference in Smart Cohort. Suggested methodologies and Casual Inference approach can help in selecting the correct patient based on criteria and also reduce patient recruitment time. In all Smart Cohort for clinical trial can reduce the time require for patient recruitment phase and increase the probability of selecting correct patients.

This research also has numerous limitations. First, machine learning model which are created are based on n2c2 and MIMIC-III dataset; hence if it needs to be used in real time then models need to be retrain using the suggested methodology. Moreover, criteria prediction model is only limited to eight criteria and ground facts (For example, Covid-19) are also not considered that could have specific effects.

Recommendations for further studies can be carried out by considering other clinical criteria. Moreover, Uniform Manifold Approximation and Projection algorithm could improve the quality of the cohort visualization. Furthermore, if it is not easy to cover whole population then Markov Chain Monte Carlo (MCMC) method can be utilized for criteria prediction and cohort visualization because it could improve the stability of models in real time scenario. Lastly, casual inference can be used for checking assumption of criteria prediction model and cohort visualization model; this will eventually make sure that assumption is not negatively affect the real time decisions.

References

Alharthi, H., Inkpen, D., & Szpakowicz, S. (2017). Unsupervised topic modelling in a book recommender system for new users. eCOM@ SIGIR,

Antunes, R., Silva, J. F., Pereira, A., & Matos, S. (2019). Rule-based and Machine Learning Hybrid System for Patient Cohort Selection. HEALTHINF,

Balakrishnan, V., & Lloyd-Yemoh, E. (2014). Stemming and lemmatization: a comparison of retrieval performances.

Chen, L., Gu, Y., Ji, X., Lou, C., Sun, Z., Li, H., Gao, Y., & Huang, Y. (2019). Clinical trial cohort selection based on multi-level rule-based natural language processing system. Journal of the American Medical Informatics Association, 26(11), 1218-1226.

De Freitas, J. K., Johnson, K. W., Golden, E., Nadkarni, G. N., Dudley, J. T., Bottinger, E. P., Glicksberg, B. S., & Miotto, R. (2021). Phe2vec: Automated Disease Phenotyping based on Unsupervised Embeddings from Electronic Health Records. medRxiv, 2020.2011. 2014.20231894.

Diaz-Papkovich, A., Anderson-Trocmé, L., Ben-Eghan, C., & Gravel, S. (2019). UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. PLoS genetics, 15(11), e1008432.

Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87.

Dunn Jr, W., Burgun, A., Krebs, M.-O., & Rance, B. (2017). Exploring and visualizing multidimensional data in translational research platforms. Briefings in bioinformatics, 18(6), 1044-1056.

Fu, J. T., Sholle, E., Krichevsky, S., Scandura, J., & Campion, T. R. (2020). Extracting and classifying diagnosis dates from clinical notes: A case study. Journal of Biomedical Informatics, 110, 103569.

Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. 2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON),

Grossmann, N., Casares-Magaz, O., Muren, L. P., Moiseenko, V., Einck, J. P., Gröller, M. E., & Raidou, R. G. (2019). Pelvis Runner: Visualizing Pelvic Organ Variability in a Cohort of Radiotherapy Patients. VCBM,

Haddad, T. C., Helgeson, J., Pomerleau, K., Makey, M., Lombardo, P., Coverdill, S., Urman, A., Rammage, M., Goetz, M. P., & LaRusso, N. (2018). Impact of a cognitive computing clinical trial matching system in an ambulatory oncology practice

Health, O. Our Story. Retrieved 17/10/2021, 2021, from https://orionhealth.com/nz/about-us/our-story/our-story/

Kanakaraj, M., & Guddeti, R. M. R. (2015). Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques. Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015),

Kherwa, P., & Bansal, P. (2020). Topic modeling: a comprehensive review. EAI Endorsed transactions on scalable information systems, 7(24).

Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). Logistic regression. Springer.

Kormilitzin, A., Vaci, N., Liu, Q., & Nevado-Holgado, A. (2021). Med7: a transferable clinical natural language processing model for electronic health records. Artificial Intelligence in Medicine, 118, 102086.

Lattar, H., Salem, A. B., & Ghezala, H. H. B. (2020). Does data cleaning improve heart disease prediction? Procedia Computer Science, 176, 1131-1140.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

Lu, J., Tan, Y.-P., & Wang, G. (2012). Discriminative multimanifold analysis for face recognition from a single training sample per person. IEEE transactions on pattern analysis and machine intelligence, 35(1), 39-51.

Maimon, O., & Rokach, L. (2005). Data mining and knowledge discovery handbook.

Marinov, M., & Efremov, A. (2019). Representing character sequences as sets: A simple and intuitive string encoding algorithm for NLP data cleaning. 2019 IEEE International Conference on Advanced Scientific Computing (ICASC),

Mishra, A., & Vishwakarma, S. (2015). Analysis of tf-idf model and its variant for document retrieval. 2015 international conference on computational intelligence and communication networks (cicn),

Mu, Y., Tizhoosh, H. R., Tayebi, R. M., Ross, C., Sur, M., Leber, B., & Campbell, C. J. (2021). A BERT model generates diagnostically relevant semantic embeddings from pathology synopses with active learning. Communications Medicine, 1(1), 1-13.

Oleynik, M., Kugic, A., Kasáč, Z., & Kreuzthaler, M. (2019). Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. Journal of the American Medical Informatics Association, 26(11), 1247-1254.

Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., & Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nature Machine Intelligence, 2(7), 369-375.

Pypi. datefinder 0.7.1. Retrieved 17/10/2021, 2021, from https://pypi.org/project/datefinder/

Segura-Bedmar, I., & Raez, P. (2019). Cohort selection for clinical trials using deep learning models. Journal of the American Medical Informatics Association, 26(11), 1181-1188.

Sethy, A., & Ramabhadran, B. (2008). Bag-of-word normalized n-gram models. Ninth Annual Conference of the International Speech Communication Association,

Sharma, A., & Kiciman, E. (2020a). Causal Inference and Counterfactual Reasoning. In Proceedings of the 7th ACM IKDD CoDS and 25th COMAD (pp. 369-370).

Sharma, A., & Kiciman, E. (2020b). DoWhy: An end-to-end library for causal inference. arXiv preprint arXiv:2011.04216.

Tang, R., & Zhang, X. (2020). CART Decision Tree Combined with Boruta Feature Selection for Medical Data Classification. 2020 5th IEEE International Conference on Big Data Analytics (ICBDA),

Van Der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. The Journal of Machine Learning Research, 15(1), 3221-3245.

Vydiswaran, V. V., Strayhorn, A., Zhao, X., Robinson, P., Agarwal, M., Bagazinski, E., Essiet, M., Iott, B. E., Joo, H., & Ko, P. (2019). Hybrid bag of approaches to characterize selection criteria for cohort identification. Journal of the American Medical Informatics Association, 26(11), 1172-1180.

WHO. International Clinical Trials Registry Platform (ICTRP). Retrieved 17/10/2021, 2021, from https://www.who.int/clinical-trials-registry-platform

Widodo, A., & Yang, B.-S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. Mechanical systems and signal processing, 21(6), 2560-2574.

Wikipedia. Embedding. Retrieved 17/10/2021, 2021, from https://en.wikipedia.org/wiki/Embedding

Xiong, Y., Peng, W., Chen, Q., Huang, Z., & Tang, B. (2021). A Unified Machine Reading Comprehension Framework for Cohort Selection. IEEE Journal of Biomedical and Health Informatics.

Yafoz, A., & Mouhoub, M. (2020). Analyzing Machine Learning Algorithms for Sentiments in Arabic Text. 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC),

Yang, D., & Hong, J. (2017). Performing literature review using text mining, Part II: Expanding domain knowledge with abbreviation identification. 2017 IEEE International Conference on Big Data (Big Data),

Zhan, X., Humbert-Droz, M., Mukherjee, P., & Gevaert, O. (2021). Structuring clinical text with AI: old vs. new natural language processing techniques evaluated on eight common cardiovascular diseases. medRxiv.

Zhang, Z., Gotz, D., & Perer, A. (2012). Interactive visual patient cohort analysis. Proc. of IEEE VisWeek Workshop on Visual Analytics in Health Care,

Zhao, G., Liu, Y., Zhang, W., & Wang, Y. (2018). TFIDF based feature words extraction and Topic Modeling for Short Text. Proceedings of the 2018 2Nd International Conference on Management Engineering, Software Engineering and Service Sciences,

A Appendix: Disclaimer

Auckland University of Technology
Master of Analytics
Research Project

Disclaimer:

Clients should note the general basis upon which the Auckland University of Technology undertakes its student projects on behalf of external sponsors:

While all due care and diligence will be expected to be taken by the students, (acting in data analytics, statistics, research or other professional capacities), and the Auckland University of Technology, and student efforts will be supervised by experienced AUT lecturers, it must be recognised that these projects are undertaken in the course of student instruction. There is therefore no guarantee that students will succeed in their efforts.

This inherently means that the client assumes a degree of risk. This is part of an arrangement, which is intended to be of mutual benefit. On completion of the project it is hoped that the client will receive a professionally documented project report, while the students are exposed to live external environments and problems, in a realistic project and customer context.

In consequence of the above, the students, acting in their assigned professional capacities and the Auckland University of Technology, disclaim responsibility and offer no warranty in respect of outcomes of the project both in relation to their use and results from their use.