ICD-10 Code Prediction using Machine Learning

MSc Research Project
Data Analytics

Sarath Kumar Samynathan
Student ID: 20185774

School of Computing
National College of Ireland

Supervisor: Giovani Estrada

National College of Ireland
MSc Project Submission Sheet

School of Computing

Student Name:	Sarath Kumar Samynathan		
Student ID:	x20185774		
Programme:	Data Analytics	**Year:**	2021-2022
Module:	Research Project		
Supervisor:	Giovani Estrada		
Submission Due Date:	31/01/2022		
Project Title:	Improvised ICD-10 (International Classification of Diseases 10th Revision) Code Prediction using Machine Learning		
Word Count:	8533 **Page Count:** 26		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Sarath Kumar Samynathan

Date: 31/01/2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

ICD-10 Code Prediction using Machine Learning

Sarath Kumar Samynathan
20185774

Abstract

Unstructured data such as free text available in Electronic Health Records used in medical organizations are complex and hard to handle manually. Machine learning can be used to convert free text into vectors of tokens with the help of natural language processing and predict the outcome based on such conversion. Our aim of this project was to predict the ICD codes based on the synonyms of the diseases and we have predicted with the help of various machine learning algorithms and neural networks. We have used Random Forest, Support Vector, Logistic regression, Naive Bayes, KNN and MLP classifier to predict the ICD codes based on synonyms. We have also used the multilanguage embedding LASER so that the work can be used in multiple languages. The algorithm with highest accuracy is a random forest classifier, achieving 98% accuracy. Our results show that a reliable way to search for medical synonyms is made possible with traditional machine learning techniques. We therefore think that any user of such an application can successfully predict ICD codes based on synonyms.

Keywords: electronic health records, Random forest, Support vector classifier, MLP classifier, Streamlit.

1. Introduction

Most of the medical data are stored by researchers in a structured format in the data which include the information of the patients such as blood concentration, weight, height, age, birth date, oxygen level, etc. This data is structured and can be analyzed easily with the help of various techniques and is organized in relational database that can be easily extracted and analyzed. In case of unstructured data that contain irregular format, such data can be extracted by people but is considered as time being technique and is quite challenging task (Cedeño Moreno and Vargas-Lombardo, 2018).

NLP or Natural Language Processing has been widely used to extract free text data with the help of various techniques and is considered as an application that makes computer understand very easily as well as deal with text data to perform different outputs.

There are different models in Natural Language Processing that helps in identifying different patterns inside the text data and summaries the information in such a way that a machine can understand those patterns from the textual data to give the desired output.

Medical researchers had used Natural Language Processing in the past to solve various problems such as summarizing broad textual data and extracting information from them, extracting the patterns and keywords and performing other functions. In this days, Natural Language Processing have become much advanced in finding the patterns from healthcare data that are unstructured in nature and such data can be trained in different models to achieve different outputs (Cedeño Moreno and Vargas-Lombardo, 2018).

EHR or Electronic Health Records are the collection of medical records containing different information of the patients that are stored electronically in healthcare institutions or patients' databases. Which such process, the medical data can be extracted easily that contained different information on the patients and allows different research as to gain information that can be shared among other Healthcare organisation in order to increase the communication between centres and use of information in an effective way (Shabbeer, 2020).

1.1 Background

Electronic health records include different types of data that contain information such as medical history, status of immunization, medication and allergies, lab test results, images from different tests, vital signs as well as other patients' information such as weight, age, height, information of billing, etc (Shabbeer, 2020).

Such information of different diseases present in patients can be found and extracted with the help of different types of data and can be represented with the help of International Statistical Classification of disease also known as ICD 10 and Related Health Problems 10th revision code. This database contains different standards of diagnosis in order to establish different representation of diagnosis that can be applied in clinical and Healthcare Research and can be used in evaluating and improving different quality in healthcare domain and can be used in improving health insurance subsidies and facilities (Rubenstein, 2015).

The International Statistical Classification of disease also known as ICD is considered as a list of medical classification that is released by the World Health Organisation. It contains different codes that include information about the signs and symptoms of the patient's disease, abnormal findings, different circumstances and complaints, cause of injuries and other health related problems such as information of patient discharge diagnosis and the diagnosis during admission of patients. Such information can be used in improving their quality and also can be used by different organizations in an effective manner. This data set was first published in 1893 and was widely used in the application of improving Health Insurance. Such code classification can also be used in various clinical research studies in health care systems and can be evaluated to improve the overall quality of the healthcare sector. For instance, the national health insurance in Taiwan had been using this code to evaluate the premium subsidies in different groups such as diagnostic related groups and other payment systems. That is is why this type of code has become very important for reimbursement in health insurance schemes (Rubenstein, 2015).

There are 22 chapters available in ICT-10 CM code ICD-10 and it differs from ICT-9 in the form of structures and concepts present in both the code. ICD-9 composed of different symbols and is not adequate enough to represent different categories of diseases and is complicated. After the conversion of code from ICD-9 to ICD-10, specificity is increased in the diagnosis of clinical diseases and increases a multitude of different codes that can be implemented. ICD-9 contains 13000 codes whereas ICD-10 cm contains 68000 codes. There are different changes and different factors of expansion in ICD-10 compared to ICD-9 that can be implemented to find out the latest procedures and diagnosis.

ICD-10 code has been divided into different categories such as Clinical Modification as well as Procedure Coding System category. ICD-10 CM which is clinical modification can test different disease diagnosis and can be listed below. The first three initials that are present in ICD-10 code represent the category of different diagnosis and the next three initials are related to the etiology corresponding to that diagnosis.

ICD-9 Code contain around 3-5 characters which is increased to 3-7 characters in ICD-10 Clinical Modification code. This is why the complexity of finding and analyzing the code is increased in ICD-10 CM code which explains different information of clinical disease.

ICD-10 PCS code which is Procedure Coding System contain around 87000 codes and the characters present in such code can be 34 different values of 10 digits ranging from 0 to 9 and 24 letters ranging from A-H , J-N and P- Z that can be applied in different character in this code. The letters such as I and O are not included in this code in order to avoid different confusion that resembles with the number 0 and 1 (Rubenstein, 2015).

1.2 Motivation

Medical data are difficult to analyze as they are written in different languages and different hospitals and when a patient visits to gain different information from different hospitals for medical treatment, medical data would be analyzed in different process such as history of the patients, pathology reports, complaints received, notes during lab tests, etc. These medical reports cannot be easily studied by data scientist as they are written in different languages and contain large number of free text data. This is why medical organizations have professional coders having licenses in order to extract data that contain information from such medical data to classify ICD codes. This type of information requires lot of time and experience for a professional coder and these codes can be applied in Diagnosis Related Groups for different patient. Those patients who are outside the hospitals are suggested by physicians and without proper experience, this type of suggestion could be considered as a loss and can create a lot of problems to the patients (Holzhauer, 2016).

Also the medical system consumes lots of maintenance cost as the coders hired requires a lot of money to analyze the information from their data and improving quality in such data can be an issue to different Healthcare sectors as they improve the overall maintenance cost in the entire sector.

This is why Natural Language Processing and machine learning is useful and can be used by anyone who possess both Medical and data mining experience. Machine learning can extract information from this medical data and Natural Language Processing can obtain information from different languages where medical data is written in free text in different languages. Machine learning is also beneficial as it can process large amount of data quickly and can easily classify diseases using the ICD codes (Stueve and Dozal, 2020).

By performing a deep study on the previous research on similar topics will help us understand how ICD codes are classified with the help of machine learning and other techniques by different researchers and how this could be beneficial in applying in real life medical data that could be helpful in healthcare organizations.

1.3 Rationale

Machine learning becomes an important topic when it comes to classify different diseases between ICD codes and different programming languages combined with some experience on them is important to perform different machine learning and Natural Language Processing algorithms. In this project we are going to use Python because it already contains the libraries we need for both natural language processing and machine learning. The Python programming language is very popular among data scientists and models can easily be evaluated with the help of performance metrics such as accuracy, precision, recall, F1 score, etc. These metrics will help analysts to understand how good the model can perform and classify ICD codes and how effectively it can be applied in real life applications. There are other programming languages, such as R, that are good in statistical analysis but do not contain large number of libraries that can be easily accessible and can be applied in both NLP and machine learning practices.

1.4 Aims and Objectives

The aim of the project is to classify ICD codes based on the synonyms of the disease with the help of machine learning and Natural Language Processing algorithms. The objectives of this project are

- To perform machine learning algorithms to classify ICD codes based on the synonyms
- To perform Natural Language Processing algorithms to process large number of text data that can be fed into machine learning algorithms
- To perform a deep literature study to understand the effectiveness of the machine learning algorithms in Healthcare data

- To evaluate different algorithms with the help of performance metrics and study the effectiveness of each model

1.5 Research Questions

Is it possible to efficiently classify ICD codes based on synonyms? To what extent can machine learning algorithms be applied to this problem, and how can they be applied in real life applications

2. Related works

Several other research studies had performed different approaches to classify ICD codes based on different features and we'll discuss some of the approaches used by them.

Different machine learning techniques such as relevance feedback, bias classifier, K Nearest Neighbour classifier had been updated in the study (Diao et al., 2021) where they used to classify ICD coding based on the discharge summaries of the patients inside the hospital. Also they had found in the study that the ensemble methods had achieved best result and experimented with another approach that is rule based approach which can measure the patterns and represent regular expressions.

A hybrid based approach that depends on machine learning and handcrafted rules was organized in a study conducted by (Nagarajan and Satya Sravani, 2015) where they had made a comparison with Multinomial Logistic Regression algorithm and a decision tree algorithm and then he evaluated the data with the help of CMC challenge data set on classifying free text of clinical data with the help of NLP techniques, SVM techniques and evaluated the technique with the help of n-gram features that was obtained in MIMIC-II data set.

The authors conducted in a study (Weibrecht, Endel and Zechmeister, 2019) found out the different training algorithms are required in different data that contain different sizes and have different number of distinct code. Feature selection in necessary for small data set and also the study was evaluated on distinguishing ICD codes on a data that contain more than 70000 text EMR records that were obtained from the University of Kentucky Medical Centre. All such dataset were tagged with ICD-9 codes.

The feature selection has been performed in a study (Luk et al., 2020) where they used feature selection in both structured and unstructured data in different integration. They performed two integrations in the classification process and found that later integration of feature selection give better results compared to the early integration. They used ICD-9 and ICD-10 medical codes and tagged the documents related to them.

Neural network has been tested in different electronic health records data and has become popular for ICD coding. In a report of CLEF eHealth evaluation lab in 2019, neural network had shown promising results and even considered as mainstream model for ICD code and neural network is essentially helpful in performing different computer vision task and in natural language processing field as they are used to extract features from the raw data and does not require any special feature selection or feature engineering steps.

In a study conducted by (Harerimana, Kim and Jang, 2021), they had performed an attention mechanism in MIMIC-III data set where they performed LSTM network to represent level of the words and the character. In the study, they developed an attention model that can be helpful in predicting 50 most top ICD codes from the MIMIC-III data set with the help of LSTM network.

The deep learning concepts such as bidirectional GRU unit has been proposed in this study (Atutxa et al., 2019) where they used ICD 10 data sets to predict ICD 10 codes based on the free text on the autopsy reports and their death certificates of the patients in the data. Another model presented a tree of sequence LSTM model where they developed a model to capture the relationships among different codes that are hierarchical in nature.

A Study conducted by (Huang, Osorio and Sy, 2019) had founded that deep learning model easily outperforms Support Vector Machine models in predicting ICD-9 top 10 codes in MIMIC-III data set. Also another study had confirmed in the same study that ICD-9 coding on both MIMIC-II and MIMIC-III data set easily outperforms Support Vector Machine that are flat and hierarchical based. The latter worked in the study had also found out that convolutional network is successful in extracting the features from the texts and can be useful in text classification in order to learn the Global features and extract large number of contents from the documents.

A concept of modal specific machine learning model has been applied in both structured, unstructured and semi structured data and used in ensemble model approach in a study conducted by (Dervilis et al., 2018) that can be used as an integration in all modal specific models in order to create ICD codes. A deep learning model is used to handle both structured and semi structured data while the tabular data present in the study are then transformed into binary features that acted as an input and are fed into design decision tree algorithm. Another text-based classification was performed and represented as joint level word embedding problem in a study conducted by this where they had implemented an attention layered Framework that can find out the relationship between the embeddings their existing between labels and text sequences. These techniques presented in this study is later evaluated in MIMIC-III and MIMIC-II data set and had achieved less results compared to any neural network model that was further represented in this paper.

Another study (Huang, 2021) conducted a deep learning network called HA-GRU model that is also known as Hierarchical attention based bidirectional gated recurrent Unit model that is helpful in extracting the features that are relevant respond to each labels in ICD-10 codes. The results found in the study are later compared on MIMIC-II and MIMIC-III data sets.

Another study conducted by (Zhou, Cui and Wang, 2021) applied the concept of multi label classification with the help of convolutional neural network that added the concept of attention and convolutional networks mechanism. They had also added regularizers on the network in predicting ICD codes and to improve classification results from the records. They had tested on ICD 9 coding on MIMIC-III data set where they had also extended the study by apply maximum pooling in all the channels. In their study, they also achieved state-of-the-art results from this data set and the concepts applied in the study .

The use of various language model had been popular in Natural Language Processing where they are used in medical domain as well Bidirectional Encoder Representation from Transformers also known as BERT model is used in a study conducted by (Mandia, 2020) where multi head attention network had been experimented and various recall values are used to evaluate the model. It is found out that the recall values had been improved when it is compared to other neural network model such as CNN and LCM while predicting ICD 10 code at CLEF eHealth report in 2019.

Also different hierarchical models are used for text classification where in one study (Inoue and Tsukahara, 2016), hierarchical dependency were observed and experimented between different classes and different random conditional field had been applied for classification of text in health domain.

A study conducted by (Raghav, Ragavachari and Ravi, 2021) had experimented generalized error that are observed from multi class text and they used hierarchical classified with the concept of DMOZ hierarchy and other concepts such as International Patent Classification to distinct the taxonomy. Also they used the technique of pruning to decrease the error and also experimented with the help of a meta classifier. The features which are observed after returning from the meta classifier are then obtained from different bounds of generalization of the errors. In this study they had found that the hierarchical loss function is also been deeply studied in other approaches that do not consist of any deep learning algorithms.

3. Design Methodology

3.1 Data Description

Our data comes from a publicly accessible source, Kaggle. It contains a lot of information in the form of HTML files scraped from the icd10data.com website. Each HTML file is assigned to a certain ICD code category. The full ICD code detail data are included in this HTML file. There is a section named ‘Synonym' in the html file where we will provide this synonyms data as input for the model to train based on the synonyms text to predict the ICD code (The Lancet, 2015). The goal is to utilize ICD code synonyms in the html file and extract the specific information of the synonyms. For data preparation, a separate script must be written. After the data has been prepared, it must be in a tabular format so that the model may be used to train the data. The predicted value will be ICD code, and the training input text will be synonyms of the ICD code. For the model's better comprehension of the data, we'll remove the punctuation and stop words from the synonyms text. Each phrase will be tokenized before being given to the work embedding model for text categorization.

3.2 Feature extraction

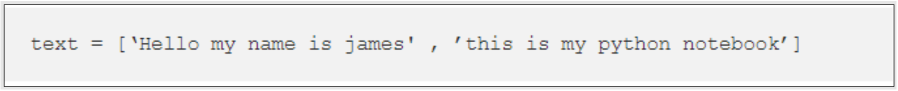
Our objective is to build a model that predicts ICD-10 codes from the free-form text that can subsequently be used to simulate the job of coders in the hospital environment. In our model, we utilize the Natural Language Processing Toolkit (NLTK) to conduct basic text preprocessing, followed by the development of a neural network model for learning the features from the input texts. Among other things, the preprocessing technique includes spell correction, case conversion, deleting stop words, tokenization, and removing unusual words. The Scikit-Learn module partitions the preprocessed data into two sets: one for training and one for validation. This layer is the first one of many that make up the neural network model. It is the term embedding layer, which refers to a set of natural language processing methods that include mapping words or phrases from a lexicon into vectors of real numbers, the first layer of the neural network model (Bengfort and Bilbro, 2019). Following that, each tokenized word is encoded into its associated word embedding using the word2vec and GloVe methods, which are well-known for their capacity to hold semantic and grammatical information in vectors (Yang and Mao, 2016).

3.3 Feature Vectorization

Machine cannot determine different characters and free text available in the data which is why it is necessary to convert into numbers by the machine. There are different vectorizer method that converts text to number in which Count vectorizer and tfidf vectorizer are the most commonly used vectorizer.

- **Count vectorizer**

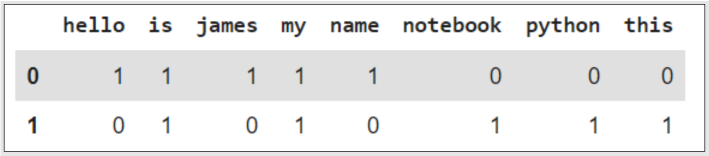
Count vectorizer is used to represent the word into numbers where unique words are represented as per the count present in the data. This method is considered as an easy method that can be directly applied into machine learning and deep learning algorithms for classification of text data (Sekhar et al., 2020). For example



```
text = [‘Hello my name is james' , ’this is my python notebook’]
```

Fig 1: Sentences present in the data

The above two sentences in the ‘text’ data can be converted as the following format

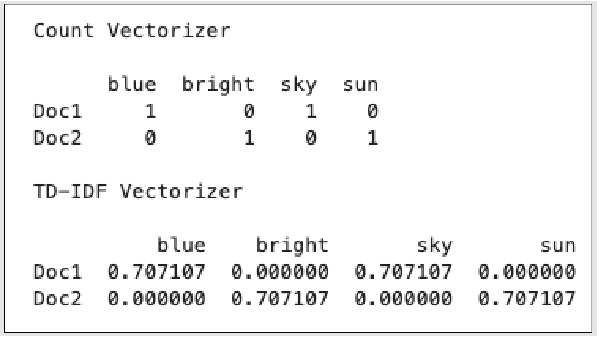


	hello	is	james	my	name	notebook	python	this
0	1	1	1	1	1	0	0	0
1	0	1	0	1	0	1	1	1

Fig 2: Count vectorizer method

- **Tfidf vectorizer**

Tfidf also known as Term Frequency Inverse Document Frequency represents the documents of words based on their weights. This metric is used to measure the elements of a word in a particular text. The value increases proportional to the number of the times a word appears in a text and is balanced with the help of the word frequency present in the data (Zhang, Shi and Wei, 2020b).



```
Count Vectorizer

      blue  bright  sky  sun
Doc1     1       0    1    0
Doc2     0       1    0    1

TD-IDF Vectorizer

          blue    bright       sky       sun
Doc1  0.707107  0.000000  0.707107  0.000000
Doc2  0.000000  0.707107  0.000000  0.707107
```

Fig 3: Tfidf vectorizer Example

3.4 Building of models

Model building includes building of different machine learning algorithms where different parameters will be studied and models will be trained on the training data. These model building step will be used in real life applications if the model found to be successful in predicting the ICD codes.

3.5 Model Evaluation

There are many performance metrics that can be used to evaluate the performance of a particular machine learning models. Although our model contains lot of classes that cannot be identified with the help of a proper performance metrics such as confusion Matrix, Precision, recall and F1 score, we will be evaluating a model with the help of accuracy score. An accuracy score is a combined score that gives us a ratio determined from the correctly classified classes by the algorithm to the miss classified classes. This score will be compared in all the algorithms in this research in order to evaluate the performance of the model given by each algorithm.

3.6 Model Selection

This step is done where different number of times is decided to train the model in order to see the range of the accuracies given by different number of times. There are different methods performed in model selection in which K-fold cross validation is used in this research where the model will be trained five times and the accuracies of each time will be observed with the help of box plot on this model.

4. Design Specification

4.1 Architectures Used

In this step, we will be explaining about the algorithms we will be using in this research. Also we will be understanding the interface which will be used for real life deployment in the web.

- **Random Forest Classifier**

Random forest is defined as an ensemble technique that can be easily applied in both regression and classification problems. The main and overall task of performing a random forest algorithm is sampling in several observations containing variables that are present in the data which go for training the model. From the data, different decision trees are then developed on the basis on conditions present inside the data. After the building of decision trees classification is performed by taking major votes of the trees and regression is performed from the average observations.

Sampling of observations is done randomly in this algorithm and features are selected that are capable to deliver significance contribution in various decision trees. In such manner, random forest creates trees which depend on each other which will further penalize the accuracies obtained from the data.

In random forest algorithm, around two third of entire observation is defined inside the training data. If X is called the total number of columns present in the data, then the algorithm performs classification by considering sqrt(X) and regression by considering X/3 (Mantas et al., 2018).

A diagrammatic representation of random forest as well as bagging algorithms is explained with different scenario. In bagging algorithms, only particular observations but all the columns present in the data were selected by the algorithms. But in random forest algorithm, only few observations and few columns present in the data were selected randomly by the algorithm. The main aim of building such individual trees is to prevent correlation existing between trees.

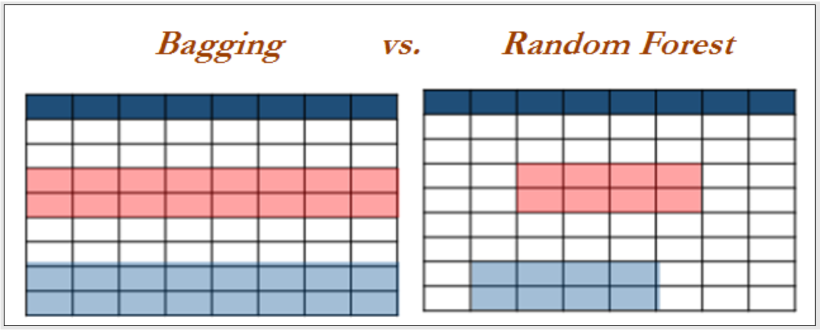


Fig 4: Bagging VS Random Forest

The following representation reviews the working of random forest trees. It can be seen that each individual tree has grown in a separate manner and a selected sample represents the depth and during the final stage, voting will be performed in order to pick the best class from the entire individual trees.

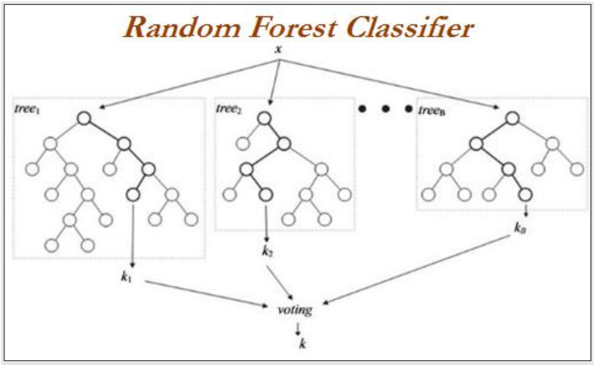


Fig 5: Random Forest Classifier

- **Naïve Bayes**

Naive Bayes is a classification algorithm that is used in classification of distinct labels and is based on the concept of conditional probability. Naive Bayes is used only in classification and is very famous algorithm for text classification techniques which is why this algorithm is used in our research. It is based on the concept of conditional probability that is used to calculate and find out the relationships between dependent events with the help of Bayes theorem (Jegadeeshwaran and Sugumaran, 2015). Let us understand how the conditional probability works.

Let say A and B are two events and we tried to calculate the probability of A where B already occurred is known as conditional Probability and is given in the following equation

$$P(A|B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Now an email classification technique ix explained below where we are predicting if the email is spam or not if it contains the word ‘lottery’ in it. The overall probability of email being spam is considered as 10% which is also known as prior probability. Now we had extra information that the word ‘lottery’ in all messages is with a probability of 4% which is known as marginal likelihood. After the implementation of the Bayes theorem in it, the posterior probability can be determined that will tell us how likely the message is spam containing the word ‘lottery’ in it (Jegadeeshwaran and Sugumaran, 2015).

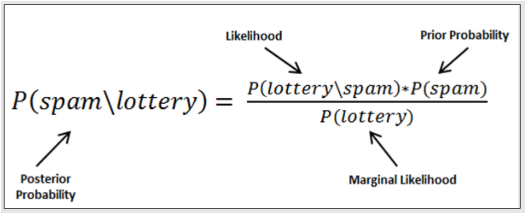
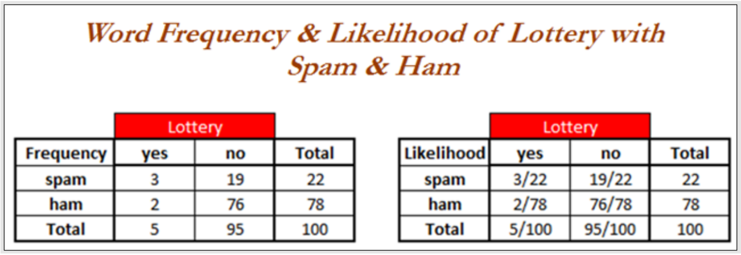


Fig 6: Example of spam classifier



Word Frequency & Likelihood of Lottery with Spam & Ham

	Lottery		
Frequency	yes	no	Total
spam	3	19	22
ham	2	76	78
Total	5	95	100

	Lottery		
Likelihood	yes	no	Total
spam	3/22	19/22	22
ham	2/78	76/78	78
Total	5/100	95/100	100

Fig 7: Word Frequency and likelihood

The above *Fig. 7* tells that the frequency of the word where ‘lottery’ word is appeared in both spam and ham messages and the above likelihood is calculated from the table explained above. The likelihood table explains that the probability of the message being spam containing the word ‘lottery’ is P(Lottery\Spam)= 3/22 = 0.13 which tells there is a 13% chance of a message being spam that contains the word ‘lottery’ in it. Also we can calculate the probability of any spam message that if it contains the word ‘lottery’ as P(Spam ∩ Lottery) with P(Lottery), which means (3/22)*(22/100) / (4/100) = 0.75.

- **Support Vector Machine**

Support Vector Machine is a black box technique that extracts the complex pattern existing in the data with the help of different classifier such as maximum margin classifier, support vector machine as well as support vector classifier. It is used in both regression and classification technique (Singh, 2016). A maximum margin classifier works on the principle of a hyperplane that is used to classify the data which is defined as a plane surface that is used to divide the data using the following equation

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

A Hyperplane is beneficial even when the data can be easily separated and fails in such data that are not separable in nature. In those cases a support vector classifier is helpful to distinguish such data.

.

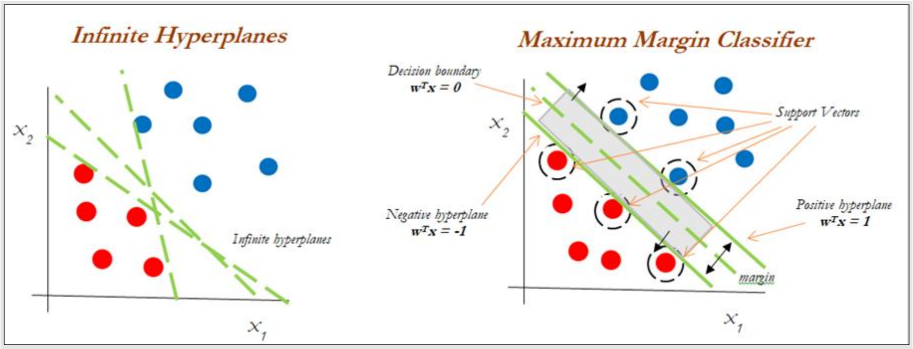


Fig 8: Working of Hyperplane

Fig. 8 explains how an infinite hyperplane can be implemented to easily distinguish the data and a maximum margin classifier can also be applied to classify them. There are some data points which touch to the extreme side of the maximum margin classifier that are known as support vectors.

Support Vector Classifier distinguishes non-separable data with the help of a hyper parameter. A very high hyper parameter value makes the model to give robust performance and the less valued parameter with makes flexible model to give better performance (Singh, 2016).

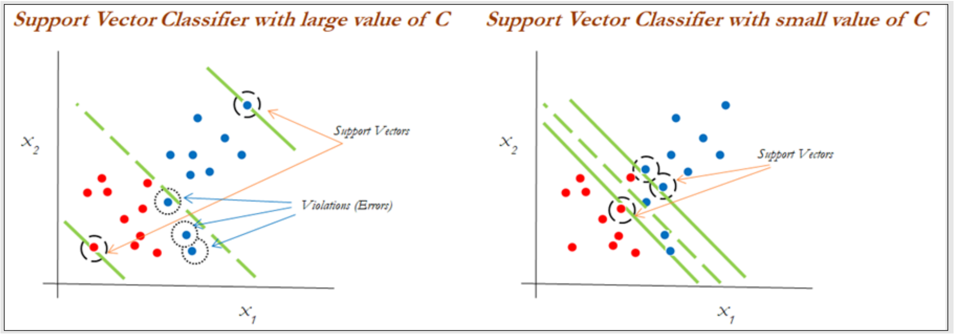

Support Vector Classifier with large value of C
x_2
Support Vectors
Violations (Errors)
x_1
Support Vector Classifier with small value of C
x_2
Support Vectors
x_1

Fig 9: Support Vector Classifier

The diagram reveals that the hyperplane with less value of hyper parameter tends to classify the data in a better way compared to the hyperplane with high value of the hyper parameter.

In some cases a Support Vector Classifier fails to separate the data that are non separable and contain more than one dimension. In such cases, support vector machine is used where a kernel trick is applied to separate the data using different degrees to convert the data into higher dimension (Singh, 2016).

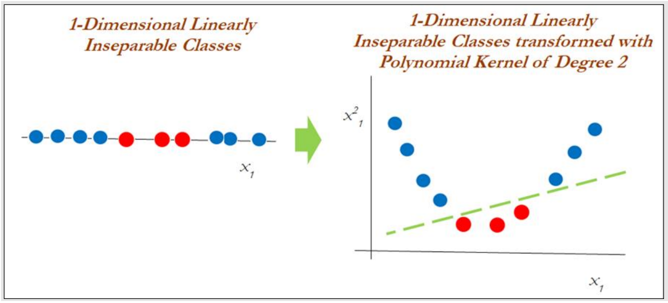

1-Dimensional Linearly Inseparable Classes
1-Dimensional Linearly Inseparable Classes transformed with Polynomial Kernel of Degree 2
x_1
x^2_1
x_1

Fig 10: One dimensional data

The above diagram reveals how the data which is non separable and present in a straight line has been separated with the help of a straight line by converting it into two dimensions.

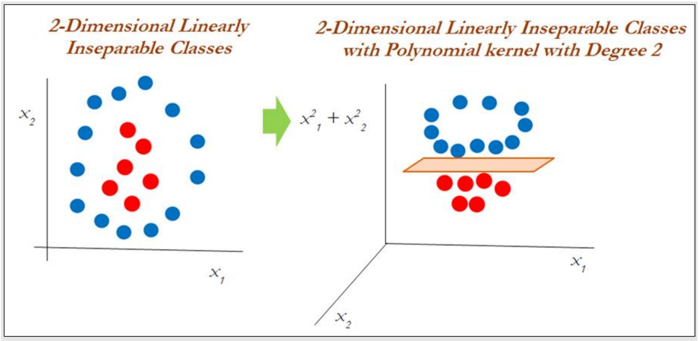

2-Dimensional Linearly Inseparable Classes
2-Dimensional Linearly Inseparable Classes with Polynomial kernel with Degree 2
x_2
x_1
$x^2_1 + x^2_2$
x_1
x_2

Fig 11: Two dimensional data

The above *Fig. 11* explains when the data is represented in two dimensions and is not separable; a two dimensional surface such as a plane surface is used to segregate the data by converting the data into 3 dimensions with the help of a kernel (Singh, 2016).

- **Logistic classifier**

Logistic regression is used in classification task to predict the likelihood of a dependent variable. The dependent variable is represented as a binary value where the binary value is decided upon a threshold value. The aim of the Logistic regression is to discover a relationship that exists between the likelihood as well as characteristic of a particular output variable. The Logistic regression based on cost function called a sigmoid function whereas in linear regression a linear function is used.

The logistic regression is represented as follows

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta 0 + \ \beta 1X$$

In the equation, p(x)/(1-p(x)) is considered as the ratio of the chances of success to the failure. In this type of regression, a linear regression equation will be translated to the logit function that will be giving an output of 0 or 1 (Kist and Silvestrini, 2015).

- **MLP Classifier**

MLP classifier also known as multilayer perceptron classifier which is a class of a neural network and sometime represent feed forward artificial neural network that contain different number of perceptron’s.

A multilayer perceptron contains at least three layers that include hidden layer, input layer as well as output layer. Each layer in an MLP classifier is considered as a neuron that acts as an activation function except for the input layers. It is used in supervised learning where a technique known as back propagation is used to train the data.

MLP is different from a linear perceptron as it contains multiple layers as well as non linear activation function and it is useful in those data that cannot be easily separated.

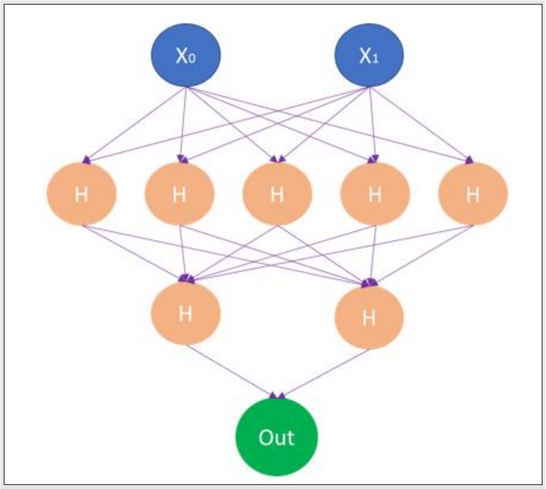


Fig 12: Neural Network classifier

From the above *Fig.12*, we can see that X_0 and X_1 are the inputs which are passed to number of hidden layers (H) and Out is the output layer which gives the desired output after passing through a non linear activation function (Pahuja and Kumar, 2021).

- **KNN classifier**

KNN or K-Nearest Neighbour is considered as a non-parametric machine learning that does not learn anything during training of the model and classify only during the test data. It can also be considered as a lazy learning algorithm as it does not generalise or create models. It only predicts the test data with the help of nearest training observation that is considered time consuming as it predicts the data based on the closest distance of the train data.

It is very time consuming which is why it is not used in complex data analysis where it will take plenty of time to predict the test data. Also it cannot capture Complex patterns unlike any other machine learning models as because it only predicts based on locating the nearest neighbors of the train data.

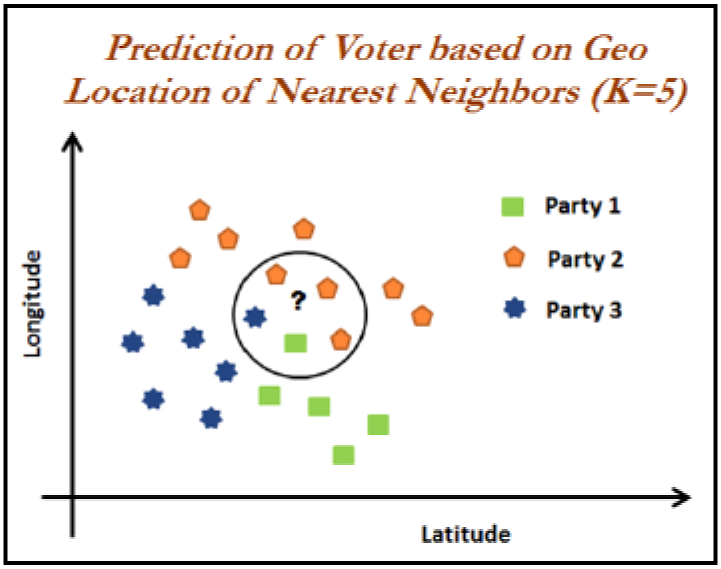
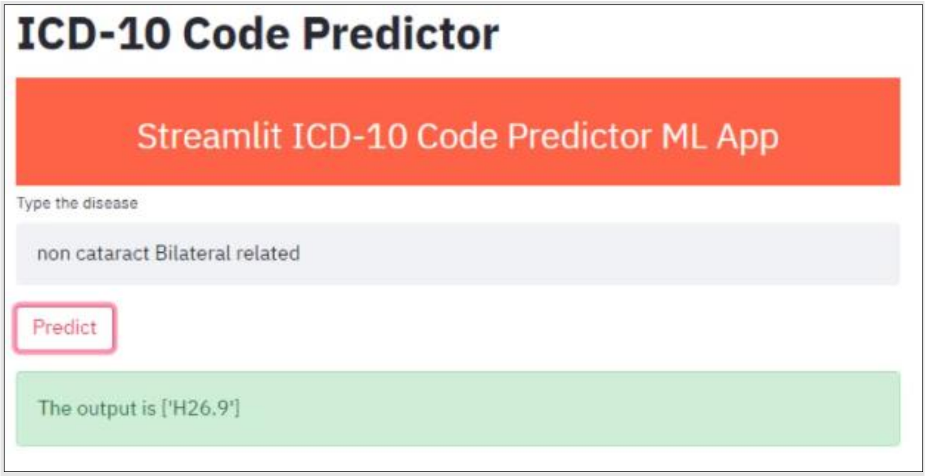

Prediction of Voter based on Geo Location of Nearest Neighbors (K=5)
Longitude
Latitude
?
🟩 Party 1
⬟ Party 2
✸ Party 3

Fig 13: KNN classifier

In the above *Fig. 13*, we can see how KNN algorithm compares the votes casted in different parties and the prediction will be the Party 2 as these points are closer to the prediction.

4.2 Model deployment

The model is deployed with the help of Streamlit library. It is open source in nature and is accessible to the public. It is built in Python languages and can be used for apps for building and running different machine learning algorithms.



ICD-10 Code Predictor

Streamlit ICD-10 Code Predictor ML App

Type the disease

non cataract Bilateral related

The output is ['H26.9']

Fig 14: Web interface of our model

5. Implementation specification

5.1 Collection of data

The data is collected from Kaggle which contain information in HTML format and web scraping is performed to scrape the data from ICD 10 data website. The HTML file that is kept on the website contains the category of ICD 10 and other features related to it. We will extract only the synonyms from the data to predict the ICD code related to it. So the aim will be storing the data into the system and importing the data with the help of Pandas into Jupyter notebook where we will take the synonyms and ICD code related to it to perform machine learning algorithms.

5.2 Data exploration

We will explore the data in this step which contain the synonyms and the ICD code related to it. We will explore the number of columns extracted from the scraped data and perform random shuffling where we will shuffle the word present in the synonyms feature and will the look into the presence of missing values in the data and perform data splitting.

5.3 Feature vectorization

In this step, we will perform Tfidf vectorizer where the words will be converted into numerical data in the form of frequency of words present in the data

This step will also undergo text cleaning to remove the stopwords present in the synonyms and perform removal of punctuations and other symbols present in the synonyms of the data.

5.4 Splitting the data

We'll divide the data into two sections in this step: one for training and one for testing. We have roughly 80,000 records in all. We will construct our test data from the train data because we do not have enough entries for each class. We'll pick 20,000 records at random from train data and swap the text in the synonyms so that the machine can't predict readily.

5.5 Model Building

In this step, we will perform model buildings in the preprocessed data where we will perform four machine learning algorithms such as random forest classifier, support vector classifier, Naive Bayes classifier and Logistic regression model. Also we will perform Multilayer Perceptron model which will be a neural network model and the results will be compared with the help of accuracy score in all the models.

Before performing multilayer perceptron model, we will perform LASER embedding in our features which is considered as a language model published by Facebook known as Language-Agnostic Sentence Representation. It is a pre trained representation of words that is considered as a probability distribution over different number of words and get the probability of a particular word appearing in a text. This language embedding is beneficial as it supports multiple languages unlike any other language models that are trained in English or any few local languages.

5.6 Evaluation

In this step, we will evaluate the model with the help of accuracy score and also perform cross-validation to train the model 5 times and determine the accuracy score from each training with the help of box plots and compare the results.

6. Evaluation

The results obtained from the machine learning algorithms performed in this research contain the accuracy scores where the model where the accuracy score is determined with the help of evaluation metrics in the table 1.

Table 1: Accuracies given by models

Model	Accuracy score
Random Forest classifier	98%
Support vector classifier	85.12%
Multinomial Naïve Bayes classifier	15.12%
Bernoulli Naïve Bayes classifier	0.88%
Logistic Regression	45.5%
KNN	86.48%
MLP classifier	77.4%

From all the accuracies observed in the table, we can see that random forest classifier gave the highest accuracy without performing any cross validation in this model. In our research, we have used Tfidf vectorizer in order to convert all the synonyms into numeric model where we have performed cleaning of the data such as removing of punctuation, symbols as well as stopwords. We did try to keep all models with the same embedding, but the MLP gave me ‘out of ram’ error. Computer couldn’t hold more than 80,000 records in memory. Then we have reduced the size of records count for SVM, KNN, RF, and NB for generating Tf-IDF vector. From all this data preprocessing, we have also performed LASER embedding where we have considered it as a language model that can predict the probability based on the words appearing in a test and supports multiple languages. With this embedding, we have performed MLP classifier with gave around 77.4% accuracy and is the highest considered to machine learning algorithms. The most common code is C38 since it deals with the Malignant neoplasm of heart diseases

7. Discussion

The main aim of this research was to predict the ICD codes based on the synonyms of the diseases with the help of Natural Language Processing in Machine Learning algorithms. In this research, we have found out that machine learning algorithms can predict ICD codes based on synonyms and the accuracy indicates that Support Vector Classifier can distinguish the ICD codes in a better manner compared to other machine learning algorithms. Also the accuracy score achieved by MLP classifier is the highest compared to machine learning algorithms.

In the research hypothesis, we have observed that the machine learning algorithms can predict ICD codes but cannot be applied in real life applications as the accuracies obtained by them is very less which is why we need to improve the model with other concepts. Also other neural network algorithms can be applied as MLP classifier gave 77.4% accuracy and different feature selection techniques can also be applied to involve more features that can be helpful in predicting the ICD codes with better accuracy.

We have used the LASER embedding technique and that is a major advantage of our model as because in the future, if any medical data come in a different language rather than English, our model can predict that as well.

Also we have built an interface with the help of Streamlit in our research which can be used by anyone to predict the ICD codes based on synonyms. This interface can be helpful to various healthcare institutions where ICD code can be predicted very easily.

8. Conclusion and Future work

Free text data which are recorded in Electronic Health Records in medical organizations are difficult to be studied as they contain lots of information and extraction of patterns from them is quite difficult with the help of manual techniques. Also experienced coder is necessary with proper license which can increase the overall maintenance expenditure in any medical organisation due to their high salary and maintenance of database and resources that store these electronic health records data.

These health records can be processed with the help of Natural Language Processing with text data can be easily studied by performing feature vectorisation as well as other word embeddings techniques while language models are used to predict different outputs based on the free text available in the text data of Healthcare institutions. This research is based on the prediction of ICD codes with the help of machine learning algorithms based on the synonyms that contain information about particular disease.

From the research, we have also performed various literature reviews where the researchers had predicted ICD codes before from different data sets with the help of different algorithms and achieved different results. In our research, we have performed a research topic in order to confirm if machine learning algorithms are efficient enough to predict ICD codes and can be applied in real life applications. Throughout the investigation, we have found out that machine learning algorithms can predict based on synonyms and can give moderate accuracy but cannot be applied in real life applications as very less accuracies are given by different machine learning algorithms tested in this research. This can be improved by selecting the top ICD codes or by performing feature selection techniques where several other features should be added to expect more information from them to predict ICD codes.

Also different deep learning algorithms can be applied to improve the result as we have seen that MLP classifier given on 77.4 percent accuracy which can also be further improved by adding more number of features or by tuning the model with other combination of best parameters that can increase the accuracy to a particular extent.

This model can be improved and applied in real life applications with different Healthcare organizations can adopt this idea of predicting ICD codes based on synonyms. Also different research and medical experts can find patterns inside the synonyms that can be helpful in predicting ICD codes and can save maintenance cost and other expenditures that were required in Manual prediction of ICD codes. These medical codes can also be helpful in understanding the disease of particular patients where lab procedures can decrease and decrease the risk of predicting a patient's disease.

Also various other approaches can be studied to understand the efficiency of predicting ICD codes and other features can also be studied so that this can be applied in different Healthcare sectors to collect those news features that are helpful in predicting ICD codes. There was errors given by the model and is more in machine learning algorithms as there are lots of ICD codes and the model learnt only a few patterns from the features. There are about 29,535 distinct classes present in the data. So, it is necessary to reduce the number of classes present in ICD codes and increase the accuracy by different algorithms.

In this research, neural network and machine learning algorithms had been performed and literature review have been carried out to study the performance of the approaches used by previous research. Out of all the researchers conducted in this project, we have understood that MLP classifier can give the highest accuracy based on the synonyms and no other deep learning algorithms is tested in this research.

So the future work of this project involves testing of other deep learning algorithms that can give a better accuracy on predicting the ICD codes based on synonyms. Also different other features can be added and feature selection techniques can be applied to select the top important features which can give better accuracy in predicting ICD codes. Also this research is not good enough to be applied in real life application as the accuracy is very less which is why we need to improve it with other algorithms.

After improvement with different other algorithms and testing, this model can be applied in real life applications to predict the ICD codes based on synonyms and this can be helpful in various Healthcare sectors where other features can also play an important role in predicting ICD codes.

The future work also includes hyper parameter tuning where different parameters should be tested in deep learning algorithms that is important in predicting ICD codes and expanding the data is required as more number of data can help the neural network models to understand the patterns in a better way and give better accuracy.

9. Acknowledgement

I would firstly want to thank my thesis guide **Giovani Estrada** at **National College of Ireland** . The door to Gio's Office was constantly open at whatever points I kept running into an inconvenience spot or had an inquiry concerning my exploration or composing. He reliably enabled this paper to be my very own work, however controlled me in the privilege the bearing at whatever point he thought I required it. Without his enthusiastic interest and info, the research couldn't have been effectively finished.

References

Cedeño Moreno, D. and Vargas-Lombardo, M. (2018). Design and Construction of a NLP Based Knowledge Extraction Methodology in the Medical Domain Applied to Clinical Information. *Healthcare Informatics Research*, 24(4), p.376.

Holzhauer, B. (2016). Meta-analysis of aggregate data on medical events. *Statistics in Medicine*, 36(5), pp.723–737.

Rubenstein, J.N. (2015). How Will the Transition to ICD-10 Affect Urology Coding? An Analysis of ICD-9 Code Use from a Large Group Practice. *Urology Practice*, 2(6), pp.312–316.

Shabbeer, S. (2020). Health Care Analysis Using Machine Learning for Mortality Risk and Readmission Risk Prediction based on EHR Data. *Journal of Advanced Research in Dynamical and Control Systems*, 12(3), pp.295–303.

Stueve, S. and Dozal, L.W. (2020). Identifying Propaganda: Comparing NLP Machine Learning Models on Propagandistic News Articles. *SSRN Electronic Journal*.

Atutxa, A., de Ilarraza, A.D., Gojenola, K., Oronoz, M. and Perez-de-Viñaspre, O. (2019). Interpretable deep learning to map diagnostic texts to ICD-10 codes. *International Journal of Medical Informatics*, 129, pp.49–59.

Dervilis, N., Simpson, T.E., Wagg, D.J. and Worden, K. (2018). Nonlinear modal analysis via non-parametric machine learning tools. *Strain*, 55(1), p.e12297.

Diao, X., Huo, Y., Zhao, S., Yuan, J., Cui, M., Wang, Y., Lian, X. and Zhao, W. (2021). Automated ICD coding for primary diagnosis via clinically interpretable machine learning. *International Journal of Medical Informatics*, 153, p.104543.

Harerimana, G., Kim, J.W. and Jang, B. (2021). A deep attention model to forecast the Length Of Stay and the in-hospital mortality right on admission from ICD codes and demographic data. *Journal of Biomedical Informatics*, 118, p.103778.

Huang, J., Osorio, C. and Sy, L.W. (2019). An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Computer Methods and Programs in Biomedicine*, 177, pp.141–153.

Huang, W. (2021). Drug-Drug Interaction Extraction Based on Bidirectional Gated Recurrent Unit networks and Capsule Networks. *Open Access Journal of Biomedical Science*, 3(3).

Luk, H.M., Allanson, E., Ming, W.-K. and Leung, W.C. (2020). Improving Diagnostic Classification of Stillbirths and Neonatal Deaths Using ICD-PM (International Classification of Diseases for Perinatal Mortality) Codes: Validation Study. *JMIR Medical Informatics*, 8(8), p.e20071.

Nagarajan, E. and Satya Sravani, V. (2015). Knowledge Abstraction from MIMIC II Using Apriori Algorithm for Clinical Decision Support System. *Indian Journal of Science and Technology*, 8(8), p.728.

Weibrecht, N., Endel, F. and Zechmeister, M. (2019). Evaluating ATC-ICD: Assessing the relationship between selected medication and diseases with machine learning. *International Journal of Population Data Science*, 4(3).

Inoue, R. and Tsukahara, M. (2016). Travel Pattern Analysis from Trajectories Based on Hierarchical Classification of Stays. *International Conference on GIScience Short Paper Proceedings*, 1.

Mandia, S. (2020). Accuracy of Diagnosis Coding Based On ICD-10. *Asian Pacific Journal of Health Sciences*, 7(1), pp.43–47.

Raghav, R., Ragavachari, S. and Ravi, H. (2021). Cosmonautics Patent Analytics using Cooperative Patent Classification Hierarchy. *Journal of Student Research*, 10(3).

Zhou, Y., Cui, S. and Wang, Y. (2021). Machine Learning Based Embedded Code Multi-Label Classification. *IEEE Access*, 9, pp.150187–150200.

Bengfort, B. and Bilbro, R. (2019). Yellowbrick: Visualizing the Scikit-Learn Model Selection Process. *Journal of Open Source Software*, 4(35), p.1075.

Jegadeeshwaran, R. and Sugumaran, V. (2015). A comparative study of Naïve Bayes classifier and Bayes Net classifier for fault diagnosis of automobile hydraulic brake system. *International Journal of Decision Support Systems*, 1(3), p.247.

Kist, M.J. and Silvestrini, R.T. (2015). Incorporating Confidence Intervals on the Decision Threshold in Logistic Regression. *Quality and Reliability Engineering International*, 32(5), pp.1769–1784.

Mantas, C.J., Castellano, J.G., Moral-García, S. and Abellán, J. (2018). A comparison of random forest based algorithms: random credal random forest versus oblique random forest. *Soft Computing*.

Pahuja, R. and Kumar, A. (2021). Sound-spectrogram based automatic bird species recognition using MLP classifier. *Applied Acoustics*, 180, p.108077.

Sekhar, S.R.M., Siddesh, G.M., Raj, M. and Manvi, S.S. (2020). Protein class prediction based on Count Vectorizer and long short term memory. *International Journal of Information Technology*, 13(1), pp.341–348.

Singh, B. (2016). Retina Recognition Using Support Vector Machine. *International Journal Of Engineering And Computer Science*.

The Lancet (2015). ICD-10: there's a code for that. *The Lancet*, 386(10002), p.1420.

Yang, X. and Mao, K. (2016). Learning multi-prototype word embedding from single-prototype word embedding with integrated knowledge. *Expert Systems with Applications*, 56, pp.291–299.

Yu, C.Y. and Shan, J. (2015a). Research on the Web Chinese Keywords Extraction Algorithm Based on the Improved TFIDF. *Applied Mechanics and Materials*, 727-728, pp.915–919.

Yu, C.Y. and Shan, J. (2015b). Research on the Web Chinese Keywords Extraction Algorithm Based on the Improved TFIDF. *Applied Mechanics and Materials*, 727-728, pp.915–919.

Zhang, X., Shi, Y. and Wei, H. (2020a). Research on TFIDF Algorithm Based on Weighting of Distribution Factors. *Journal of Physics: Conference Series*, 1621, p.012007.

Zhang, X., Shi, Y. and Wei, H. (2020b). Research on TFIDF Algorithm Based on Weighting of Distribution Factors. *Journal of Physics: Conference Series*, 1621, p.012007.