REVA
UNIVERSITY
Bengaluru, India

A Project Report on

An Interactive Web Solution for Electronic Health Records Segmentation and Prediction

Submitted in Partial Fulfilment for Award of Degree of

Master of Business Administration
In Business Analytics

Submitted By

Sudeep Mathew
R19MBA09

Under the Guidance of

Mithun DJ
Senior Manager – Data Science
RACE, REVA University

REVA Academy for Corporate Excellence - RACE
REVA University
Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Bengaluru - 560 064
race.reva.edu.in

August, 2022

REVA
UNIVERSITY
Bengaluru, India

Candidate’s Declaration

I, **Sudeep Mathew** hereby declare that I have completed the project work towards **Master of Business Administration in Business Analytics** at, **REVA University** on the topic entitled **An Interactive Web Solution for Electronic Health Records Segmentation and Prediction** under the supervision of **Mithun Dolthody Jayaprakash Senior Manager of Data Science at Race Academy**. This report embodies the original work done by me in partial fulfilment of the requirements for the award of degree for the academic year **2022**.

Place: Bengaluru	Name of the Student: Sudeep Mathew
Date: 27-August-2022	Signature of Student

REVA UNIVERSITY

Bengaluru, India

Certificate

This is to Certify that the project work entitled **An Interactive Web Solution for Electronic Health Records Segmentation and Prediction** carried out by **Sudeep Mathew** with **SRN R19MBA09**, is a bonafide student of REVA University, is submitting the second-year project report in fulfilment for the award of **Master of Business Administration** in Business Analytics during the academic year 2022. The Project report has been tested for plagiarism, and has passed the plagiarism test with the similarity score less than 15%. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said degree.

Signature of the Guide	Signature of the Director
Name of the Guide	Name of the Director
Mr. Mithun Dolthody Jayaprakash	**Dr. Shinu Abhi**

External Viva

Names of the Examiners

1. Vaibhav Sahu, Strategic Cloud Engineer, Google
2. Abhishek Sinha, Data Science Manager, Capgemini

Place: Bengaluru

Date: 27-August-2022

REVA UNIVERSITY
Bengaluru, India

Acknowledgement

I am highly indebted to **Dr. Shinu Abhi**, Director, and Corporate Training for their guidance and constant supervision as well as for providing necessary information regarding the project and for their support in completing the project.

I would like to thank my project guide **Mr. Mithun Dolthody Jayaprakash** for the valuable guidance provided to understand the concept and execute this project. It is gratitude towards all other mentors for their valuable guidance and suggestion in learning various data science aspects and for their support. I am thankful for my classmates for their aspiring guidance, invaluable constructive criticism, and friendly advice during the project work.

I would like to acknowledge the support provided by Hon'ble Chancellor, **Dr. P Shyama Raju**, Vice Chancellor **Dr. M. Dhananjaya**, and Registrar **Dr. N. Ramesh**. It is sincere thanks to all members of the program office of RACE who are supportive of all requirements from the program office.

It is my sincere gratitude towards my parents, and my family for their kind co-operation and encouragement which helped me in the completion of this project.

Place: Bengaluru
Date: 27-August-2022

REVA UNIVERSITY
Bengaluru, India

Similarity Index Report

This is to certify that this project report titled **An Interactive Web Solution for Electronic Health Records Segmentation and Prediction** was scanned for similarity detection. Process and outcome are given below.

Software Used: Turnitin
Date of Report Generation: 25-August-2022
Similarity Index in %: 9 %
Total word count: 6607
Name of the Guide: Mithun Dolthody Jayaprakash

Place: Bengaluru
Date: 27-August-2022
Verified by: M N Dincy Dechamma

Name of the Student: Sudeep Mathew
Signature of Student

Signature
Dr. Shinu Abhi,
Director, Corporate Training

List of Abbreviations

Sl. No	Abbreviation	Long Form
1	EDC	Electronic Data Capture
2	EHR	Electronic Health Records
3	SAE	Serious Adverse Event
4	NLP	Natural Language Processing
5	CRISP – DM	Cross-Industry Standard Process for Data Mining
6	CRO	Contract Research Organisations
7	EDA	Exploratory Data Analyses
8	CHF	Chest Heart Failure
9	PCA	Principal Component Analyses
10	LAB	Laboratory
11	EMR	Electronic Medical Records
12	AUC	Area Under Curve
13	TFIDF	Term Frequency Inverse Document Frequency

List of Figures

List of Tables

Abstract

A vast variety of patient data have been collected and monitored through Electronic Health Records (EHR) using various tools in the clinical research industry and it is a concern for a pharmaceutical company is to ensure the safety of the patients who are participating in the clinical trials. It is evident that need of centralized analytics solutions for EHR datasets that deliver insights and predictability.

The objective of the project is data acquisition and data understanding and then creating a web interface for data exploration for the basics of statistics and segmentation for patient's risk profiling. Also, the project intended to create a predictive model recommending serious adverse events.

A Clinical data monitoring system is developed and deployed over a streamlit python framework as a web solution. The app consists of data exploratory features in which source data is uploaded as csv files into the application for insights and statistics. A patient's clustering model developed using k-means using machine learning for chest heart failure segmentation and it is observed that six clusters were optimal while training the model and it is incorporated into the application for predicting the segments of the patients based on the risk levels. Few machine learning model were trained on patient's historic diagnosis text data for predicting the occurrences of serious adverse events and the logistic regression model indicated 89 % of AUC score in test data and is deployed into the application for the prediction.

Keywords: Natural Language Processing, EHR, Segmentation, Serious Adverse Event Prediction

Table of Contents

Chapter 1: Introduction

Patient's safety was the primary concern for the pharmaceutical companies during the phases of drug development. In clinical research, it is important to make sure the participant is safe during the drug research. The importance of analyzing patient data is the primary concern for clinical research organizations. The analysis of patient's data is to monitor the progression of the diseases and occurrences of adverse events at the time of research. "Clinical research organization" also known as CRO helps pharmaceutical companies during the clinical research process to improve efficiency and speed. CRO supports pharmaceutical companies during the R&D phase by providing a way for some of the necessary stages in the clinical trial process. "CRO" market reached 36.7 bn in 2017 and has expected the market to reach 51 bn by 2024. The efficacy of the drugs in patients has been taken into account for final approval by the regulatory boards. Patient data is analyzed and submitted to the drug authority for the final approval for drugs to be marketed and before the approval.

Electronic health records (EHRs) contain patient diagnostic records, physician records, and records of hospital departments. For heart diseases, we can receive huge unstructured data from EHR time series. By analysing and mining, we can identify the links between diagnostic events and ultimately predict the probability of the occurrence of a serious adverse event. Adoption of EHR dataset and the increase of digitized information about patient data revolutionize the emergence of clinical research in oncology researches (Berger, 2016). One of the applications of EHR data is improvising learning system on clinical research and which helps in various application of patient selection, dosing, drug target etc. is discussed by (Eichler, 2019). Standardizing electronic health records in Indian health record system is implemented by (Pai, 2021). The comprehensive techniques for modelling EHR data is provided by (Yadav, 2018)

The web app is developed by aiming to help the medical or clinical team to monitor the safety of the patients during the clinical trials by allowing the users to explore the data through visualization and statistics, and clustering and the probability of the occurrence of SAE.
Clustering or segmentation techniques are helpful to find out underlying hidden association between each data points. This helps the business to take decision on each cluster. Detection anomaly or outlier in EHR data was implemented by (Estiri, 2019). The primary objective is

to apply unsupervised clustering techniques on EHR data the result indicated that clustering techniques produced with high sensitivity and specificity.

The occurrence of serious adverse events or SAE is one of the primary concerns that pharmaceutical companies face during clinical trial. This application intended to solve this problem by providing the probability of the occurrence of SAE by analyzing patient’s diagnosis data. A work has implemented a prediction model to detect the occurrence of adverse event such as cardiac arrest by utilizing patient data (Churpek MM, 2014).

The Application intended for the clinical or medical monitor team which help to improve overall the patient’s safety concerns and address key issues during the clinical trial.

Chapter 2: Literature Review

In one of the works by Ziyi was referring to the challenges and the perspectives of machine learning multimodal in electronics health records. And this work suggests that including structured data is not enough to achieve good result and instead his studies seek to use machine learning and deep learning models on structured and unstructured EHR dataset (Ziyi Liu, 2016).

The machine learning models in electronic health records could outperform conventional survival model for predicting mortality in coronary disease. This works includes multiple machine learning model such as cox model, random forest in 80000 patients EHR dataset and the output indicated that it outperform conventional models (Andrew J. Steele, 2018).

Adeler had proposed a risk prediction model for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. This work was to combine time series model and cox proportional model to get a range of risk prediction models and the evaluation of model is based on discriminatory statistics (Adler Perotte, 2015).

Predicting the risk of heart failure with EHR sequential data modeling is developed and this project used patient’s diagnosis data with LSTM sequential model used and the evaluation based on the utility and efficacy of the proposed solution (BO JIN, 2018).

Lutz Has performed an unsupervised machine learning model to detect patient’s subgroups in electronic health records. This project used agglomerative hierarchical clustering and kmeans clustering on patient’s lab and coded datasets. The results indicated that natural grouping is present in the dataset and hierarchical clustering provides higher quality clusters than kmeans clustering (LÜTZ, 2019).

One of the works implemented by Gabriele is for private hospital workflow optimization through k-means clustering. This work is to optimize work allocation-based staff members, patients, hospital, and location to cluster the staff members by means of the frequency of the facing time (Gabriele Spini, 2019).

K-means clustering is used for healthcare knowledge discovery is one of the works and the objective is used to discover hidden patterns by applying k-means clustering and self-organizing map (MAP) (Ahmed Alsayat, 2016). In one of the works is an unsupervised machine learning model used for the discovery of latent disease clusters using electronic health record (Wang, 2020).

Prediction of health outcomes for pediatric patients is one of work implemented and the approach of the project was to implement Bayesian model and clustering for predicting the risk of type 2 diabetics for the children between the age 10 and 14 (Hubbard, 2021). An anther work implemented using unsupervised LDA approach to cluster patients sub groups into multiple clusters using patient's health records (Mantas, 2020).

Natural language processing is used in the field of unstructured text data and this work explore the possibilities of applying NLP models in EHR datasets. Author discussed on various application of NLP such like classification models, question answering, phenotyping, knowledge graphs, medical dialogue etc. (Irene Li, 2021).

In comparative analysis of text classification approaches in electronic health records indicated that in text classification in traditional approaches could exceed the performance of the contextual embedding models such as BERT (Aurelie Mascio, 2020). A work mentioning the application of deep learning models in electronic health record by developing various deep neural network models. The proposed solution is a deep learning model for predicting the health risk of the patient (Jose Roberto Ayala Solaresa, 2019).

On other hand a work implemented by Mascio suggesting different text classification approaches on patient's text data (Mascio, 2020). This work focused on various traditional machine learning models and compared with contextual embedding BERT and identified that traditional models performed well on the text data. Bittar tried to implement suicide risk assessment using text data and the work implemented to predict the tendency to commit suicide by extracting text features from the clinical notes and the data trained SVM model (Bittar, 2020).

Various machine learning techniques on electronic health records were discussed above on segmentation as well as text classification. This project trying to attempt to create a web interface for the users to bring data and to segment patients who has chest heart failure in order to classify the patients into multiple groups based on the risk level. This project also seeks to solve the predictability of death of the patient based on the historic text data.

The previous works on electronic health records discussed above was regarding various machine learning model and application and evaluation techniques to determine the quality of the models. This review of previous works provides a framework and approaches to tackle the business problem which is discussing in coming section.

Chapter 3: Problem Statement

In clinical research, the patient's safety has become a primary concern for the pharmaceutical companies. **Early identification, prevention of disease** and **ensuring patients care** has been crucial steps for the companies. Companies finds **difficult to analyse and interpret patient's electronics health records**. Huge patient's datasets are a source of meaningful insights of a patient's health. And medical or clinical team do not have a way to **explore the data and segments patients**. The prevention of the occurrence of a serious adverse event like the **probability of the occurrence of death** must be prevented. By continuous monitoring of patients EHR records and predictive analytics reduce the risk of patient's life.

Chapter 4: Objectives of the Study

The aim of this project is focused on

- Data acquisition and data understanding of the EHR dataset and design and implement exploratory data analytics tool.
- Develop chest heart failure segmentation using k-means algorithm.
- Design patients serious adverse event prediction models using patient's diagnosis data.
- Integrate the Machine learning models into web solution and deploy the App into streamlit framework.

Chapter 5: Project Methodology

This project aims to address the safety concerns of the patients during the clinical research. The Application features that it allows the users to input the csv files which contains patient's electronic health records and then the users are able to explore the various data points and get various statistics, correlation and multivariate and univariate analysis and the next feature includes the users can brings chest failure patients data points for getting patients segmentation which helps the clinical team to classify patients based on their risk levels.

Finally, the last part of the app consists of the text classification by utilizing patient's diagnosis data, the project aims to build a serious adverse event or the probability of the occurrence of the death.

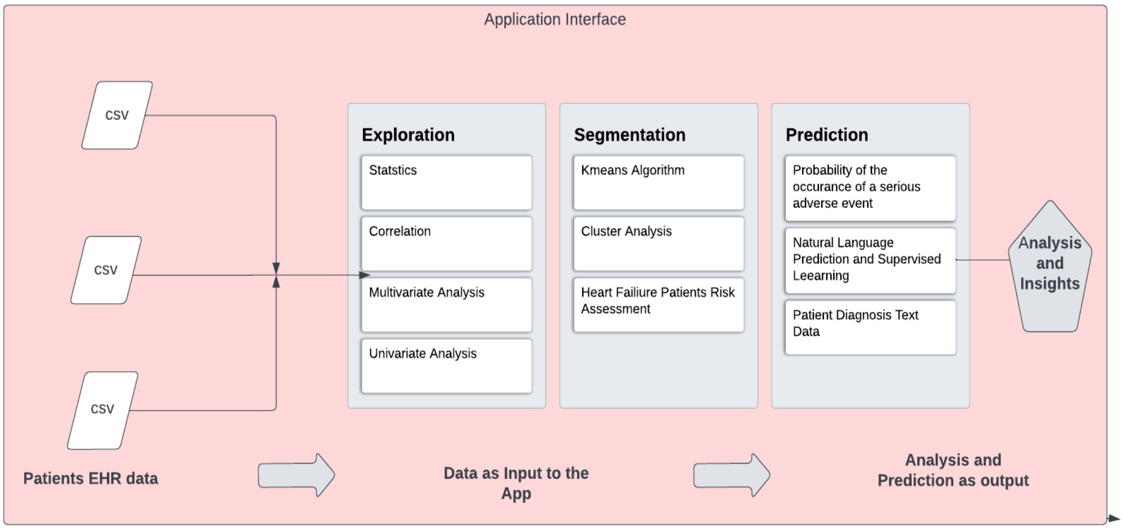


Figure No. 5.1 Project Methodology

Figure No. 5.1 illustrates the project methodology and below mentioned are the primary components in the application.

- Datasets:-Application provides an interface to connect to patients EHR data. Users can upload data in a CSV file format into the application interface.

- Application: consists of three major components which are EDA, segmentation and finally Prediction. Each component is unique in nature and is implemented for different business purpose and can be considered features that App provides.

- EDA: This feature helps the users to upload patients various datasets in a CSV file format into the app for exploring patient’s data and to gain insights and various statics that helps the business to take decisions.

- Segmentation: This feature helps to understand segments the patients into homogeneous and within each group they share similar profiles and helps the business to perform different action on each cluster groups

- Prediction: The prediction feature design and implemented on patient’s diagnosis data (text data) collected over the patients visit to the hospital. The model uses all the diagnosis text data as an input for the NLP model and output feature is whether patient has expired. NLP model created to classify the text data.

Chapter 6: Business Understanding

A CRO is accountable for planning, setup, and day-to-day execution and management of its contracted clinical research study often referred to as “clinical trial”. CRO supervising tasks such as “Handling” and “managing” the technical side data collection and medical testing comprise a significant portion of its tasks. One of the prime important tasks of CRO is to adhere that clinical trial is compliance with regulatory agency guidelines is crucial, and adhering to Good Clinical Practice (GCP) standards is part of the CRO’s role as it acts as the trial’s central hub, connecting the sponsor with other stakeholders such as “regulatory agencies”, “ethics committees”, “vendors, hospitals”, etc. The important aim of the business is to effectively handle patient’s electronics health records efficiently and to ensure the safety of the patients. There are three major issues that business tries to address through clinical data analytics tools in which the first option is to have an exploratory data analytics tool which helps the business to analyse the data effectively.

By Analysing huge chunks of dataset may not be good idea and difficult to interpret the underlying associations and meaningful insights in the data. And hence clustering or segmentation allows the users to take right decision based on each cluster that the patients reached into based on their health condition. Clustering or segmentation is a machine learning technique that the business can make use to find distinct cluster and categorize patient on each cluster.

The importance of early identifying the risk of serious adverse events such as the probability of the occurrence of the death will helps the business to early decision which helps the business to take actions proactively. The second option is to have a segmentation feature where patient’s clusters can be identified and the third option is to develop a serious adverse event prediction using patient’s diagnosis datasets.

- Medical: A medical team is a group of medical doctors who extensively uses this App for their medical data review to ensure the safety of the patients.

- Sponsor: Sponsor can also utilize the app for exploring the data, and for early identifying the occurrences of serious adverse events

Chapter 7: Data Understanding

In business understanding, business questions are framed and, in this phase, we will explore the data and how the data helps to answer each question. In this project test data is used which mimics actual production data.

Data is collected from MIMIC iii data mart and it consists of 46000 patient's data and it consists of patient's demographic, hospital admissions, vitals, labs and microbiology, diagnosis and drug administration datasets.

Table No. 7.1 Dataset and Description

No	DATASETS	DESCRIPTION
1	PATIENTS	Contains the demographics data for each patient's
2	ADMISSIONS	Consists of unique records of patient's admission to the hospital
3	D_ICD_DIAGNOSIS	Standard 1cd9_code and label for different diagnoses which is a standard dataset
4	DIAGNOSES_ICD	Diagnosis contains icd9_code for each patients visits
5	PRESCRIPTIONS	Data is related to the drugs administrated for each patient's during the admission in the hospital

Table No. 7.1 listed out all the datasets the project utilizes to build patients analytics solution

Table No. 7.2 Patient Dataset and Field Description

No	**PATIENTS**	
	Fields	***Description***
1	SUBJECT_ID	Unique id for all patient's
2	GENDER	Gender for each patient's
3	DOB	Date of birth of the patient's
4	DOD	Date of death of the patient's
5	DOD_HOSP	Date of death if the death at the hospital
7	EXPIRE_FLAG	Determine if the patients died or alive

The Table No. 7.2 list down all the demographics dataset and the description of each column

Table No. 7.3 Admission Dataset and Fields Description

No	**ADMISSIONS**	
	Fields	***Description***
1	SUBJECT_ID	Unique id for all patients
2	HADM_ID	Unique id for every hospital admission for each patient’s
3	ADMITTIME	Date and time of admissions
4	DISCHTIME	Date and time of discharge
5	DEATHTIME	Date of death if the death at the hospital
6	ADMISSION_TYPE	Admission type whether it is elective or emergency
7	ADMISSION_LOCATION	Location of the Admission
8	DISCHARGE_LOCATION	Location of the discharge
9	INSURANCE	Type of Insurance for the patient’s
10	LAUNGUAGE	Language the spoke
11	RELIGION	Religion of the patient’s
12	MARTIAL_STATUS	Determines whether patient married or not
13	ETHNICITY	Ethnic of the patient
14	ENDRGETIME	Date and time of registration end
16	DIAGNOSIS	Diagnosis of the patient’s disease
17	HOSPITAL_EXPIRE_FLAG	Whether the patient dies in hospital or not
20	HAS_CHARTEVENTS_DATA	Does chart event data available for patients

Table No. 7.3 shows the admission dataset and contains all the information of patients during the course of hospital admission

Table No. 7.4 Diagnosis Dataset and Fields Description

No	**DIAGNOSIS_ICD**	
	Fields	***Description***
1	ICD9_CODE	Standard code for the diagnosis
2	SHORT_TITLE	Short title for each diagnosis
3	LONG_TITLE	Long Title for each diagnosis

Table 7.4 shows all the columns present in the diagnosis ICD dataset which is a standard dataset contains coded values and description of each diagnosis

Table No. 7.5 Diagnosis and Codes Dataset and Field Description

No	**DIAGNOSIS**	
	Fields	***Description***
1	SUBJECT_ID	Subject_id for each patient’s
2	HADM_ID	Hospital Admission id for patients who are admitted in the hospital
3	SEQUENCE_NO	Sequence no for each record
4	ICD9_CODE	Dictionary code corresponds to patients diagnoses

Table 7.5 contains patient’s diagnosis codes for each visit to the hospital

Table No. 7.6 Prescription Dataset and Field Description

No	**PRESCRIPTION**	
	Fields	***Description***
1	SUBJECT_ID	Unique id for all patients
2	HADM_ID	Unique id for every hospital admission for each patient’s
3	ICUSTAY_ID	Unique id for patients ICU admission
4	STARTDATE	Date and time of ICU admission date and time
5	ENDDATE	ICU end date time
6	DRUG_TYPE	Type of drug administrated to patients
7	DRUG	Drug name given to patient
8	ROUTE	Route of the drug administration

Table No. 7.6 lists the prescription dataset contains patient’s prescription of the drug during the visit of the hospital

Chapter 8: Data Preparation

For data preparation, there are two distinctive steps the first one is data wrangling which is to join various datasets to create a single data frame for the modeling purpose and the second task in data preparation is the data preprocessing which consists of data cleaning, normalization etc.

8.1 Data Wrangling

Data wrangling is one of the crucial tasks in this project. There are multiple source datasets that needs to be joined or combined in order to make a large data frame for the data modeling. Python and SQL are extensively used in data wrangling. In order to build CHF patients segmentation dataset for the segmentation problem.

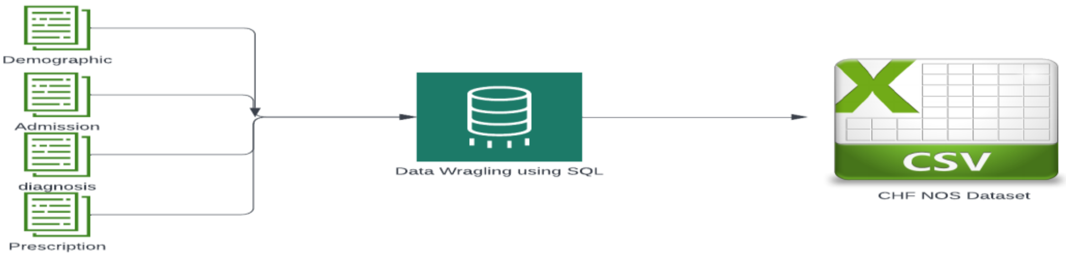


Figure No. 8.1.1 Data Wrangling for Segmentation

Figure No. 8.1.1 shows the data wrangling flow for chest heart failure dataset. The input data sets are joined by using SQL queries and later exported as data frame for the segmentation problem.

8.2 Data Pre-Processing

Data pre-processing is an essential step prior to the data analytics project that we need to ensure the data is in valid format for the visualizations and models to work. In this project data pre-processing includes data cleaning and noise removal, normalization etc. have been implemented for segmentation as well as prediction model.

8.2.1 Data Pre-processing in Segmentation

The output of the data wrangling created a single data frame for the data analysis and data modeling. The primary diagnosis of the patients with this CHF NOS formed to a data frame in the previous data wrangling steps.

For segmentation, data scaling and missing records removal were the two methods performed prior to data modeling phase. Data scaling is a method to normalize the range of independent variables or features of data. In data preprocessing, it is also known as data normalization. This project uses *standardscaler* package to scale the features of the data. And finally, all the missing values are removed from data frame by missing value removal method in python. In Figure No.

8.2.1 display data pre-processing flow for segmentation model.

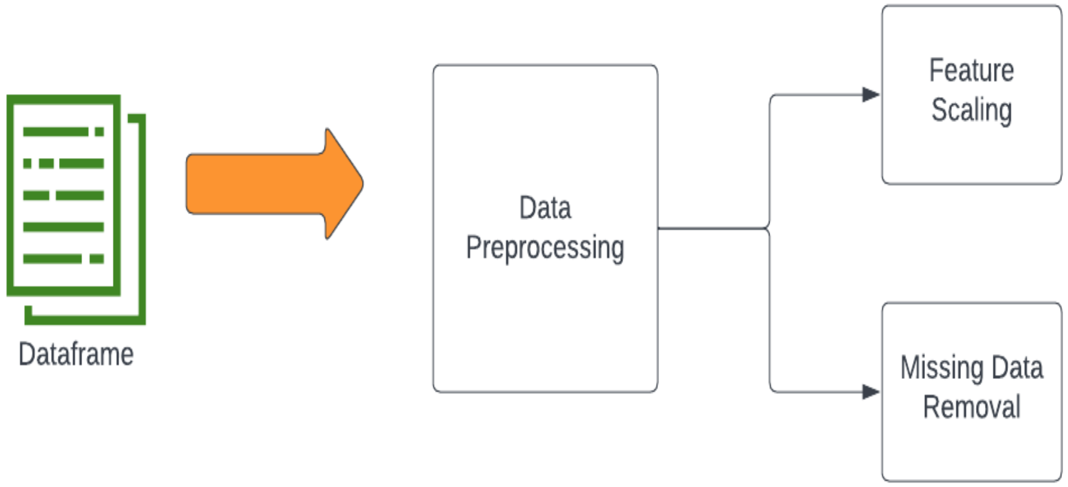


Figure No. 8.2.1 Data Pre-Processing Flow for Segmentation

8.2.2 Data pre-processing in SAE Prediction

The data frame after wrangling step contains one feature contains patient’s diagnosis text data and the label for the serious adverse event is a binary field called “expire flag” which tells whether the patient had observed any serious adverse event or not. The input text data have gone through various text pre-processing techniques prior to data modeling.

Text pre-processing steps includes text normalization, removing special characters and numbers, stop ward removal. These steps are applied on each patient's diagnosis text data for better results in modeling. Text normalization is the process converting all the text data into lower cases. Removal of unwanted special characters and integers is an important steps in text pre-processing hence those are should not need for the text prediction. Finally stop words are non-important words in the text and are most occurred words and which may not provide overall semantic meaning of text. In Figure No. 8.2.2 displays the data pre-processing for classification model.

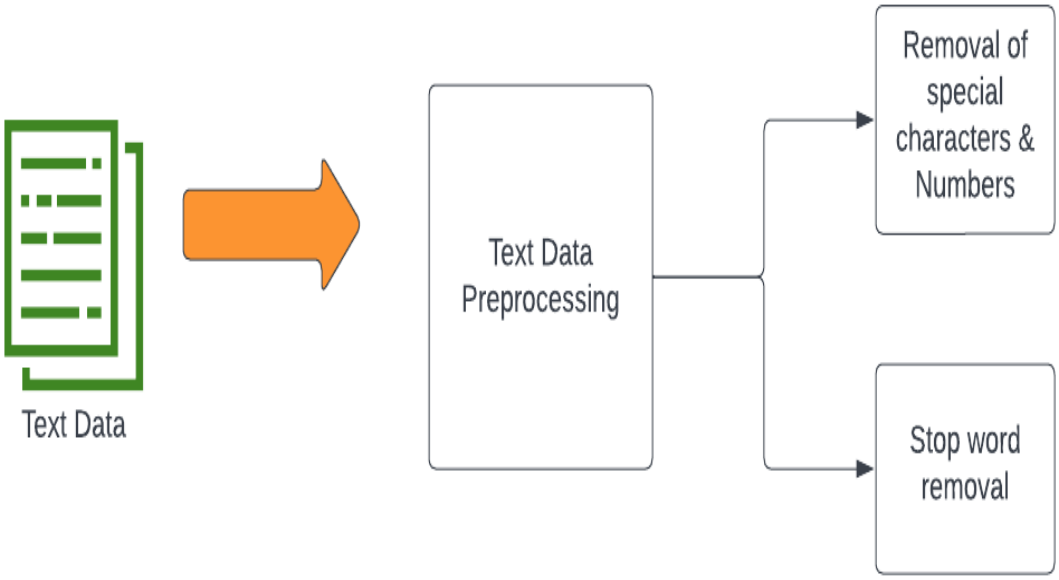


Figure No. 8.2.2 Data Pre-Processing Flow for Classification

8.3 Feature Engineering

Feature engineering one of the crucial steps in data science projects. The features or independent variables are input for the data model to predict the label or dependent variable in supervised learning. Featuring engineering is the process of creating new variables by modifying the existing variables programmatically. By adding new features have advantage and disadvantage in machine learning. If the new features are good predictors, then that will improve the accuracy of the model and also adding more features may cause dimensionality reduction.

8.3.1 Feature Engineering for segmentation

Various features were created in prior to the segmentation modeling. The tables No. 8.3.1 table list down the features were created for segmentation model.

Table No. 8.3.1 Features of Segmentation

No	**Features of Segmentation**	
	Name	***Description***
1	Count of diagnosis	Aggregated diagnosis count for each patient
2	Drug administrated days	Aggregated total days of drugs given for each patient
3	No of drugs	Total number of drugs given to each patient
4	Age Group	Grouped patients bases on the age such as >90 as very old, 60 -80 as senior citizen, >18 as adult, <18 as young
5	Ethnic Group	All the non-white people grouped as other as others as white
6	Is hyper	Whether hypertension present for patient 0 as not present and 1 as present
7	Is kidney	Whether kidney diseases present for patient 0 as not present and 1 as present
8	Is diabetic	Whether diabetic present for patient 0 as not present and 1 as present
9	Is resp	Whether respiratory diseases present for patient 0 as not present and 1 as present

8.3.2 Feature Engineering in SAE Prediction

The cleaned text diagnosis data is converted into input features for the SAE prediction model. Count vectorizer and tfidf vectorizer are used for creating features for the text data. Count vectorizer is method to convert text to numerical data by considering count of the word in each sentence. On other hand tfidf is better than count vectorizer because it is not only focuses on the frequency of words present in the corpus but also provides the importance of the words. In Figure No. 8.3.1 displays the features created for building text classification model.

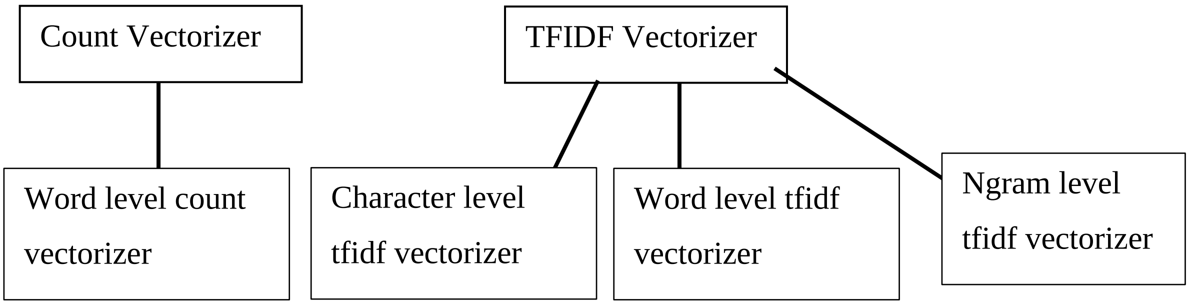

Count Vectorizer
TFIDF Vectorizer
Word level count vectorizer
Character level tfidf vectorizer
Word level tfidf vectorizer
Ngram level tfidf vectorizer

Figure No. 8.3.1 Text Feature Engineering

Chapter 9: Modeling

The application consists of exploratory data analysis tool helps the users to automate the data exploration by loading the data into the app through csv files. Data exploration helps the user to interactive with data and to gain information and statistics.

The second part of the Application is for patient data segmentation. By seeing the patients raw data may not interpretable and segmentation is one option which helps the user to classify patients based on the clusters. Finally, the app consists of serious adverse event prediction using patient's diagnosis data. This helps the users to get the probability of the occurrences of adverse events. Figure No. 9.1 depicts the three different features that included in the design of the application

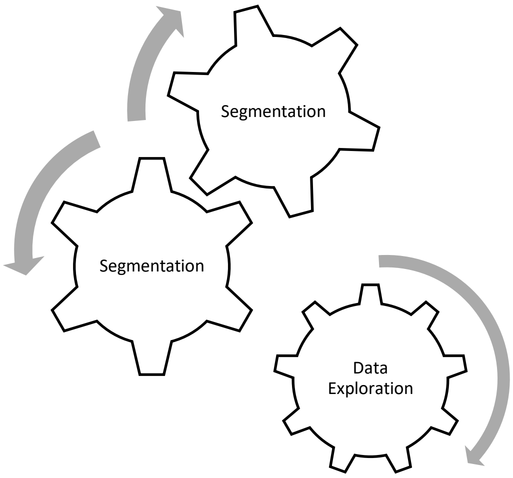


Figure No. 9.1 Design of the Application

9.1: Designing of Data Exploratory Tool

The design of the Application is for exploratory data for patient's health records. It should efficiently join variety of patients data was the initial steps while developing the application.

This was handled by joining the primary keys of all the datasets that discussed in the data understanding section. Python *pandas* package is used extensively for merging and joining various dataset in the application. The application designed and developed in a way that the users have the ability to drag and drop the csv files for the patient's demographics, admission, and diagnosis and prescription datasets to automatically generate visualization and charts that depicts meaningful insights to the user

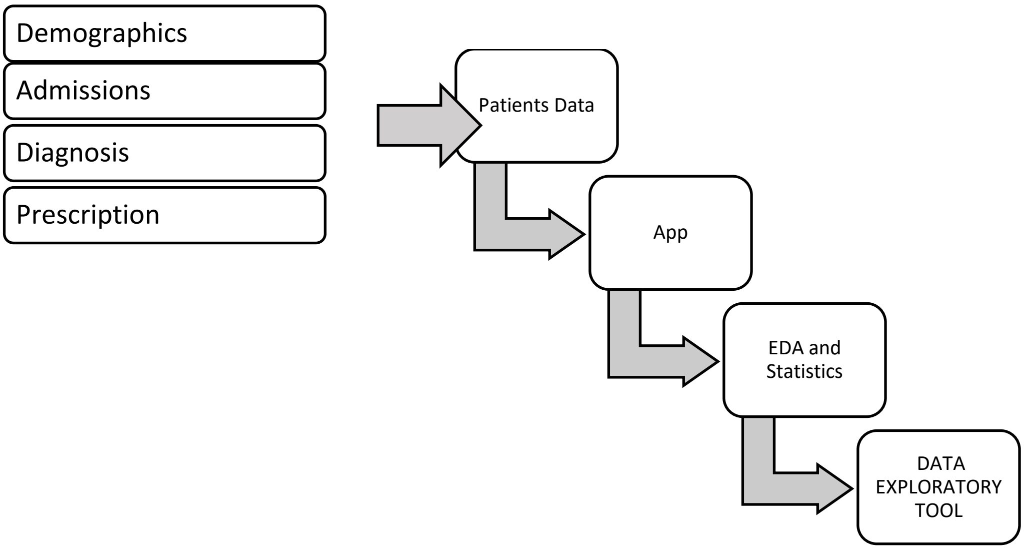


Figure No. 9.1.1 Design of Data Exploratory Tool

In the Figure No. 9.1.1 illustrates the design of the data exploration tool which is one of the primary components of the application.

The second primary component of the application is used for patient's data segmentation for chest heart failure diseases. The segmentation or clustering is an unsupervised machine learning techniques that classifies patients into different clusters of homogeneous and while separating the heterogeneous data. This will arrange the data of the same group together and helps the users to observer and take decision for each distinctive cluster

9.2 Designing of CHF NOS Patient's segmentations

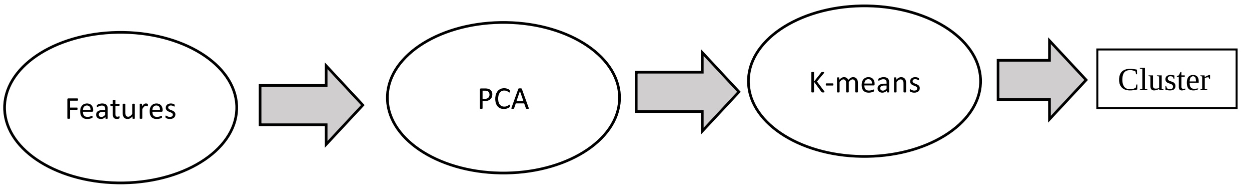


Figure No. 9.2.1 Design of Segmentation Application

The Figure No. 9.2.1 shows the workflow of patient's segmentation feature of the App. The input features contain patient's demographic, admission, prescription and diagnosis datasets were joined given to PCA model for the dimensionality reduction and after words the principal components are transferred to the K-means model for the segmentation of the patients. This model helps the business to identify the risk levels of the patients to take valid decision for each cluster.

9.3 Designing of Serious Adverse Event Classification

The occurrence of serious adverse events such as death in the clinical must be prevented. The classification model developed using text data classification. Patient's historic diagnosis text data collected and joined together for all the patients in the database and multiple classification techniques trained on the dataset to obtain best fitted model. The output of the model is a binary classification that tells whether the probability of the serious adverse can occur in future

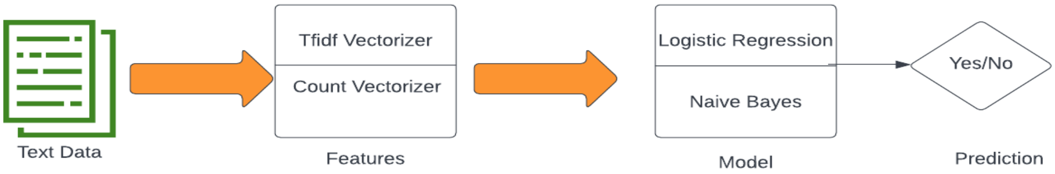


Figure No. 9.3.1 Design of Text Classification Model

Figure No. 9.3.1 depicts the workflow of patient SAE prediction application. Patient's diagnosis text data converted to tfidf and count vectorizer as features for training the model. Term frequency-inverse document frequency is a text vectorizer that transforms the text into a usable vector. It combines 2 concepts, Term Frequency (TF) and Document Frequency (DF).
Count vectorizer is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text.
Converted features then passed as an input for the model such as logistic regression and a naïve bayes which in this case is supervised binary classification techniques. The term binary classification is referred because the outcome of the model is binary or yes/no classification.
In statistics, the (binary) logistic model (or logit model) is a statistical model that models the probability of one event (out of two alternatives) taking place by having the log-odds (the logarithm of the odds) for the event be a linear combination of one or more independent variables ("predictors"). Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong

(naive) independence assumptions between the features. Abstractly, naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector representing some *n* features (independent variables), it assigns to this instance probabilities. Features are trained on both model and best resulted model considered for deployment.

Chapter 10: Model Evaluation

After training the model, the evaluation of the performance of the models is essential in order to pick the best performance model for deploying the solution. Each modeling techniques have different types of evaluation model. In contract to supervised learning where we have the ground truth to evaluate the model's performance, clustering analysis does not have solid evaluation metric that we can use for evaluating the outcome of the model. Elbow method and silhouette analysis where the two types of evaluation metrics used in k-means clustering algorithm. Unlike clustering for SAE prediction, we have the outcome of the patients in the test dataset to evaluate the performance of the model. There are various performance metrics are widely used in the supervised classification model in which the most common ones are F1 ratio, AUC score, AUC Curve plotting, Log ratio etc. In this project Elbow method is used for clustering and AUC curve is used for classification models as evaluation metrics.

Elbow method gives us an idea on what a good *k* number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.

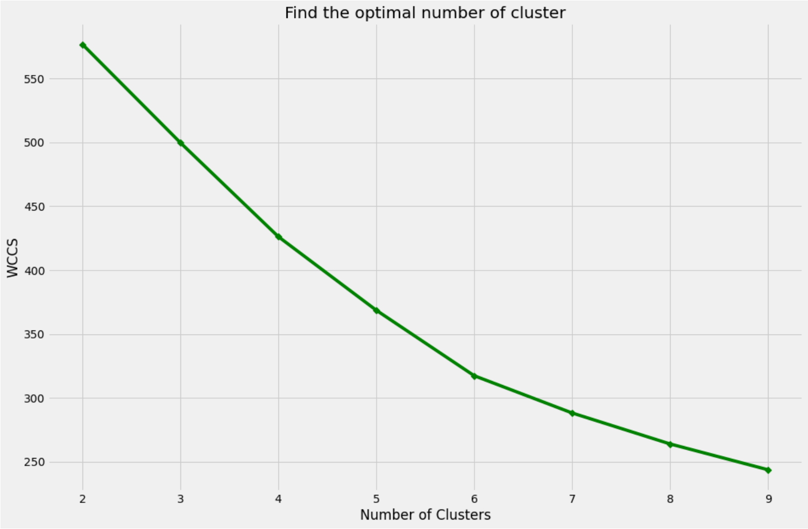


Figure No. 10.1 Segmentation Model Evaluation

From Figure No. 10.1 the y axis shows the WCCS (Within-Cluster Sun of Square) and x axis plot the number of clusters the ideal number of clusters that is used in this project is six as the WCCS point is steadily releases till cluster number six and afterwards the decentness of the cluster of reduced slightly which implies the optimal cluster night be six in this project.

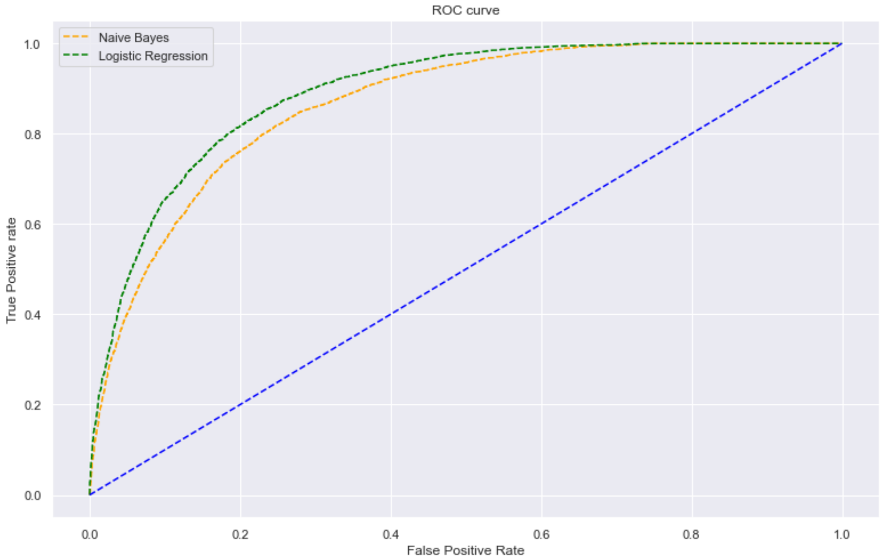


Figure No. 10.2 SAE Classifier AUC Curve

Figure No. 10.2 shows the ROC curve is plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis. An excellent model has AUC near to the 1 which means it has a good measure of separability. A poor model has an AUC near 0 which means it has the worst measure of separability. In this work two models performance was good and per figure the logistic regression shows the higher the curve and so it is considered as the best model for the classification problem.

Chapter 11: Deployment

Development of the model is not enough in providing a solution it is just a phase of stage. The usability of solution is something that business looking for as there are lot of solution builds but all of that cannot be used. Hence it is important to consider that all solution that provided must be in user friendly format that it must easily interactive with the users of the app. A web solution is build using streamlit framework by integrating the model and all other widgets for the usability of the app. The developed models were promoted for deployment by using python streamlit framework which is light weight framework used to construct python-based web dashboard and analytical widget for the data science solution. Application consists of three distinctive features which are exploratory data analysis app, chest heart failure clustering model build on k-means algorithm and logistic regression model for SAE classification are exported as model package for the deployment.

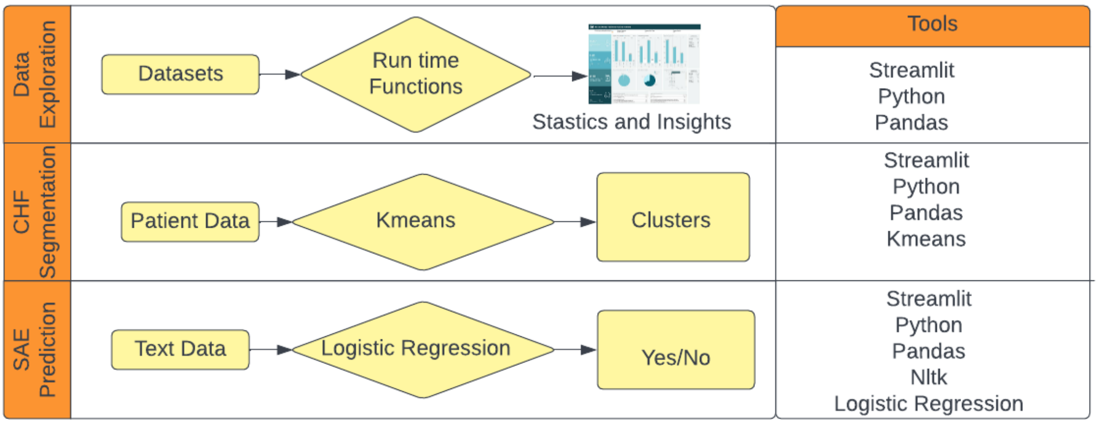


Figure No. 11.1 Deployment Flow and Framework

Python: Python is used for the building of the application. Several functions build in python for handling data wrangling and data transformation primarily for dataset exploration. Various dataset was joined together using pandas merging and aggregating functionality.

Python used extensively as building the models for segmentation as well as classification model. K-means clustering and logistic regression is built on python and exported as files for the deployment. Exported models integrated with streamlit package for building the application.

Streamlit: *Streamlit* framework provides and widgets and graphs to display effectively and integrate all the building blocks of the apps into a web solution and the streamlit server is used for the deploying the App.

Pandas: *Pandas* library is utilized for the data wrangling and data transformation phases for the deployment of the app. Several functions that build on top of panda's library that helps the data to interact with different widgets in streamlit.

Nltk: *Nltk* is used for creating text classification model which is mainly used for data cleaning steps such like removal characters, stop words.

Keras: *Keras* library provides python classes and objects for various machine learning model building and *keras* provided k-means and Logistic regression classes for building the model and later exported as a package for the deployment of the application

Chapter 12: Analysis and Results

When the application loads initially it is displaying the homage which has the navigations and widgets to each section of the app. In addition, home page has the options to upload multiple csv files for the analysis. And Figure No. 12.1 is the home page for the users once the user visits web page.

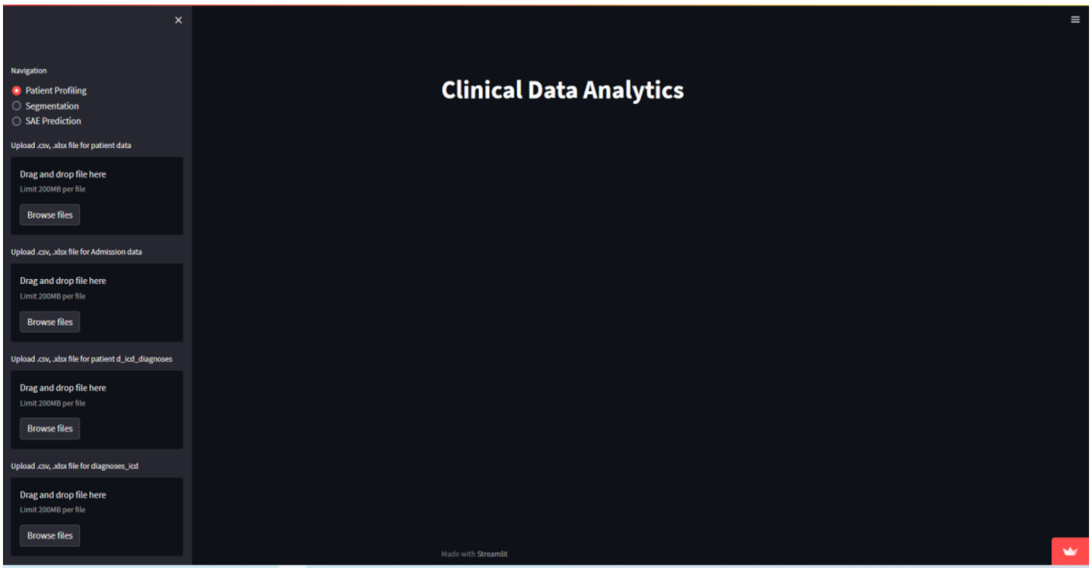
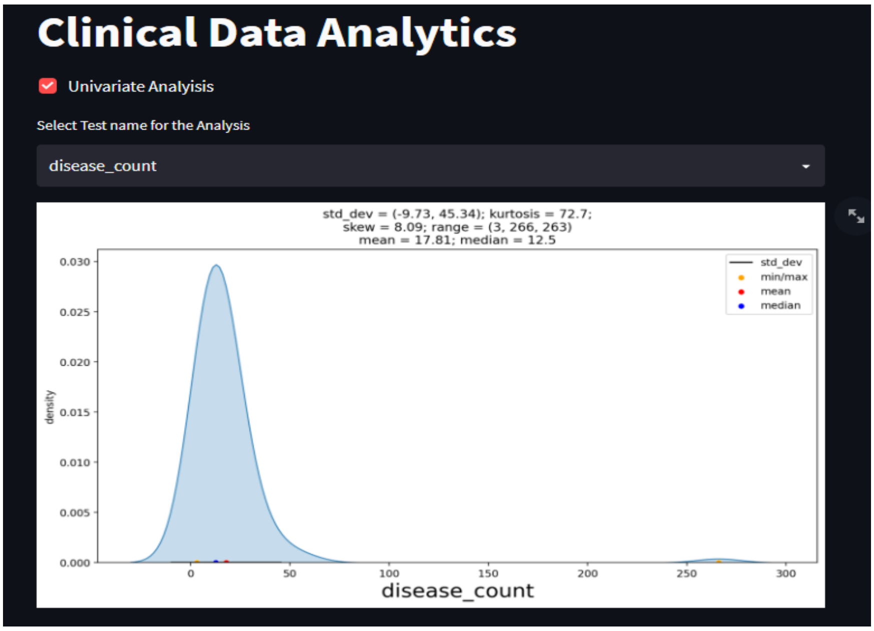


Figure No. 12.1 Home Pages

In Figure No. 12.2 Patient Profiles option, it allows the users to interact with various features of the data to get univariate analysis and multivariate analysis for continuous and categorical datasets


Clinical Data Analytics

☑ Univariate Analyisis

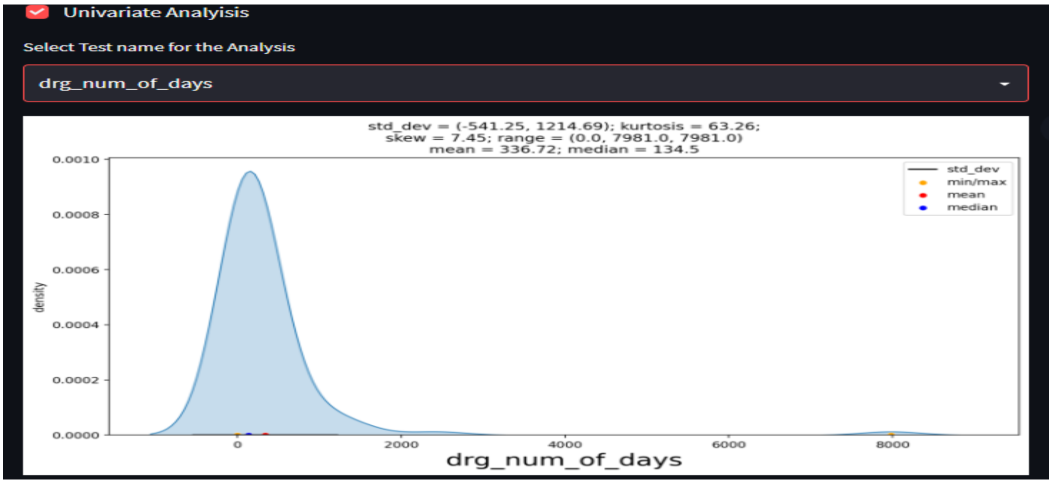
Select Test name for the Analysis

disease_count

std_dev = (-9.73, 45.34); kurtosis = 72.7;
skew = 8.09; range = (3, 266, 263)
mean = 17.81; median = 12.5
density
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
50
100
150
200
250
300
disease_count
std_dev
min/max
mean
median

Figure No. 12.2 Univariate Analyses

In Figure No. 12.2 disease count is plotted as histogram and it gives various statics along with the plot and the mean of the disease count of all patients are 17.81 and there are few outliers such patients has disease count of near to 300.


☑ Univariate Analyisis

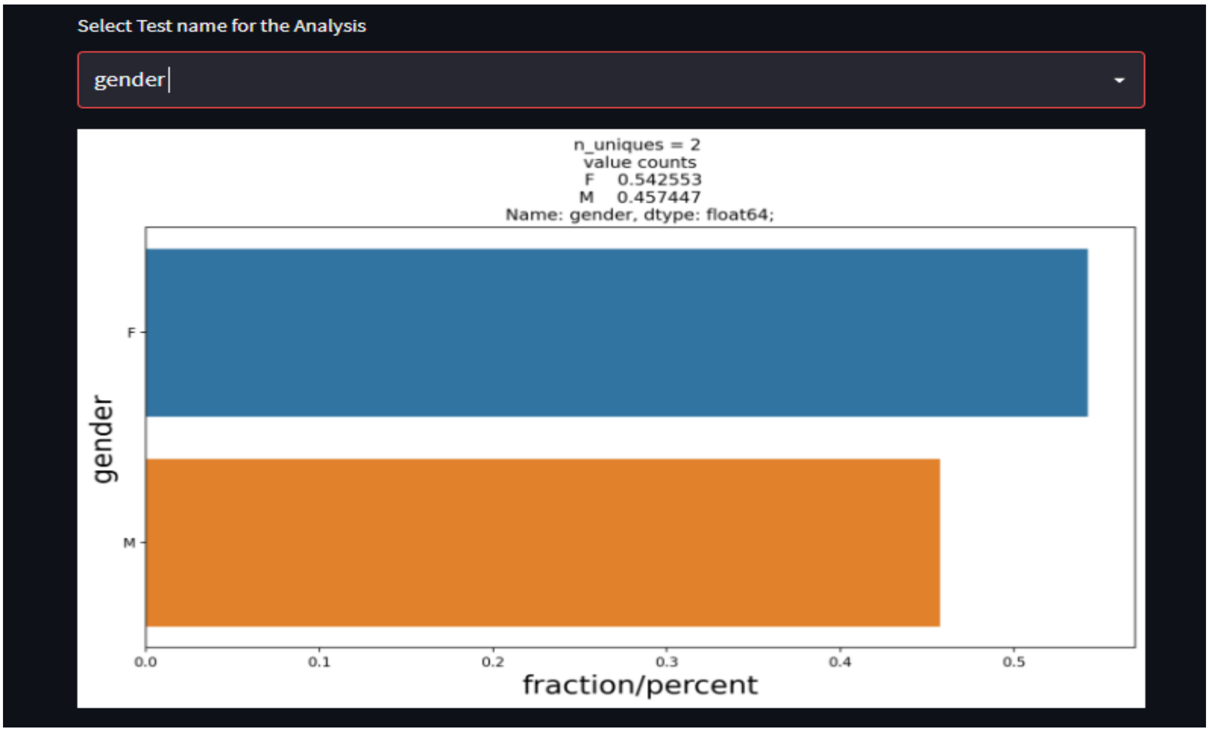
Select Test name for the Analysis

drg_num_of_days

std_dev = (-541.25, 1214.69); kurtosis = 63.26;
skew = 7.45; range = (0.0, 7981.0, 7981.0)
mean = 336.72; median = 134.5
density
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
2000
4000
6000
8000
drg_num_of_days
std_dev
min/max
mean
median

Figure No. 12.3 Drug Administrated Days Histogram

Figure No. 12.3 plots the histogram for the number of days administrated to the patients. The mean of the drug given days are 336 however it is observed that outliers present in the data that for some patients the drug given is more than 8000.



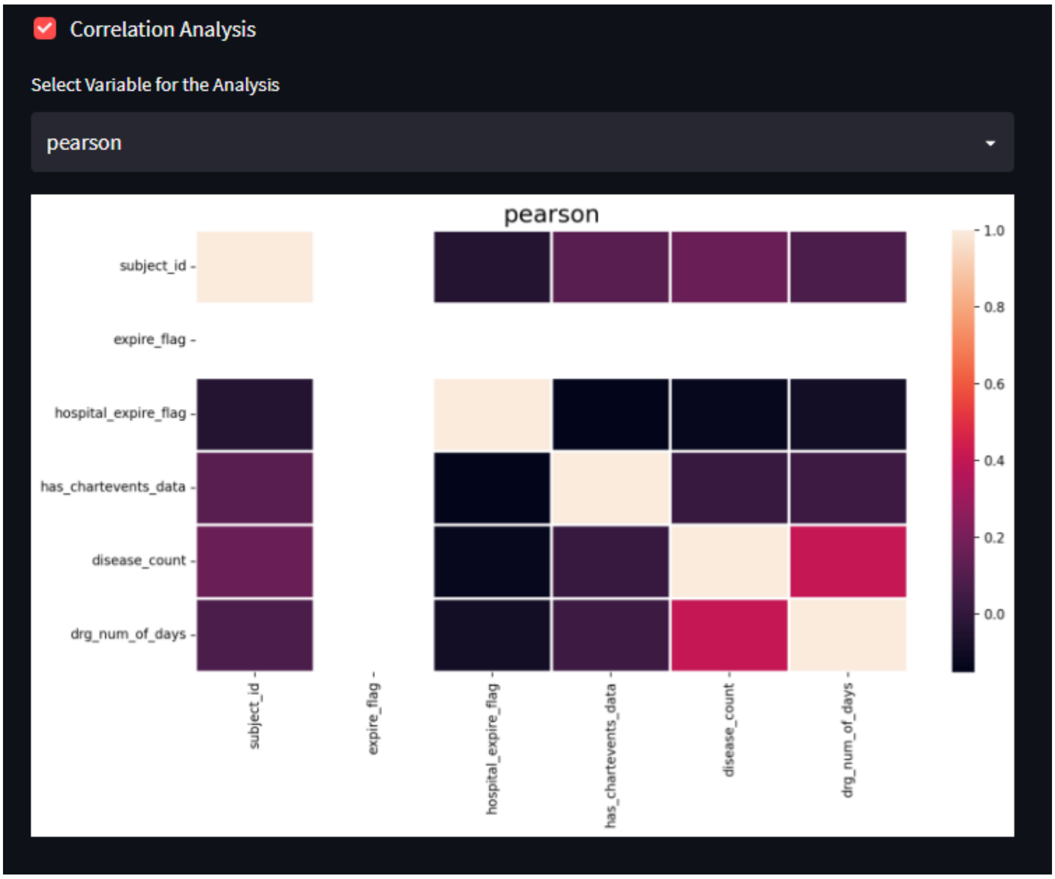
Select Test name for the Analysis

gender

```
n_uniques = 2
value counts
F    0.542553
M    0.457447
Name: gender, dtype: float64;
```

Figure No. 12.4 Gender Bar Plot

In Figure No. 12.4 Gender is a categorical feature and it is best to display in bar plot with density bar diagram which results the distribution of females and males in the dataset. In the data females are more as they are 54% compare to males 46 %.



☑ Correlation Analysis

Select Variable for the Analysis

pearson

Figure No. 12.5 Correlation Plot

In Figure No. 12.5 Plots the correlation graph of all the continuous features and disease count has a positive correlation to the number of drugs administrated which is possible as the number disease increases the drug given to the patients also increases.

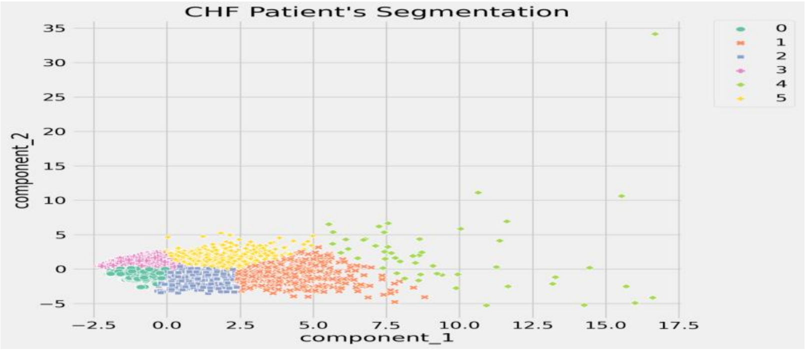


Figure No. 12.6 Segmentation Plot

Figure No. 12.6 display the six different clusters in multiple colors that each cluster has distinctive characteristics and business can take decision on each cluster.

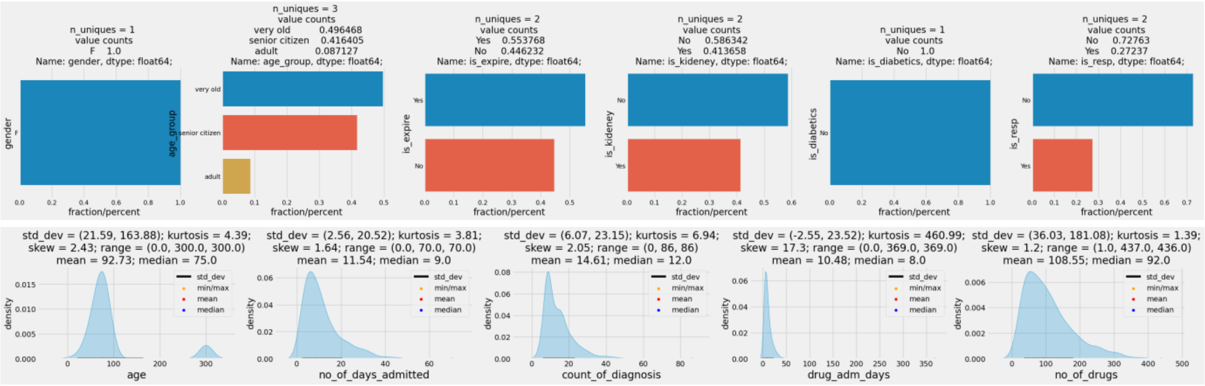


Figure No. 12.7 Cluster 1 Results

Figure No 12.7 details are mentioned below

- Males - 54 % and Females 45 % and different age groups are present
- 65 % of People are died and no people had diabetic and 70 % of people has respiratory disease
- No of days admitted in the hospital less, no of days mean is 14 % and dug administrated days mean is 10 % but no of drugs given to them is huge
- Even though no of days admitted is less however they have more number of drugs

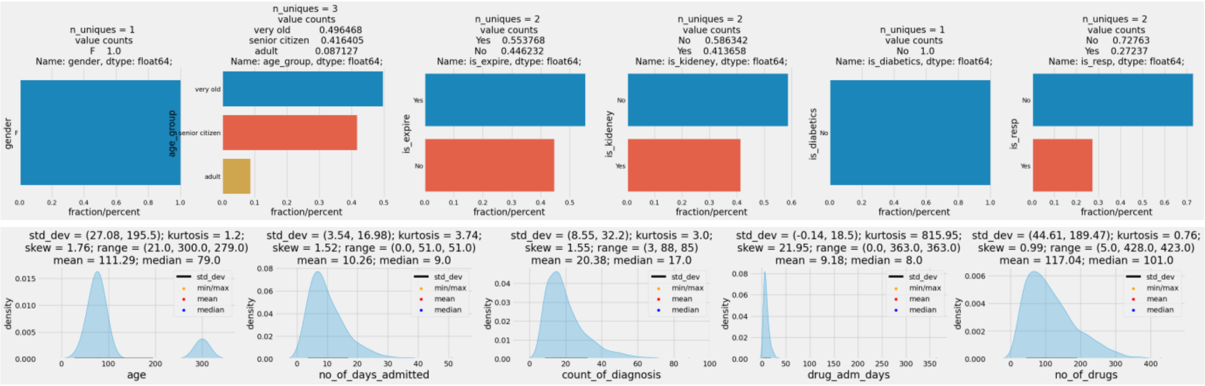


Figure No. 12.8 Cluster 2 Results

Figure No 12.8 details are mentioned below

- Only Females very old age people present in this cluster and patient are expired and not expired with almost same distribution
- kidney issues presence is lower but almost equal distribution
- No people has diabetic and very few people had respiratory issues
- People are not admitted to hospital often and drug administrated days are less
- Though patients are not admitted often they have consumed more drugs

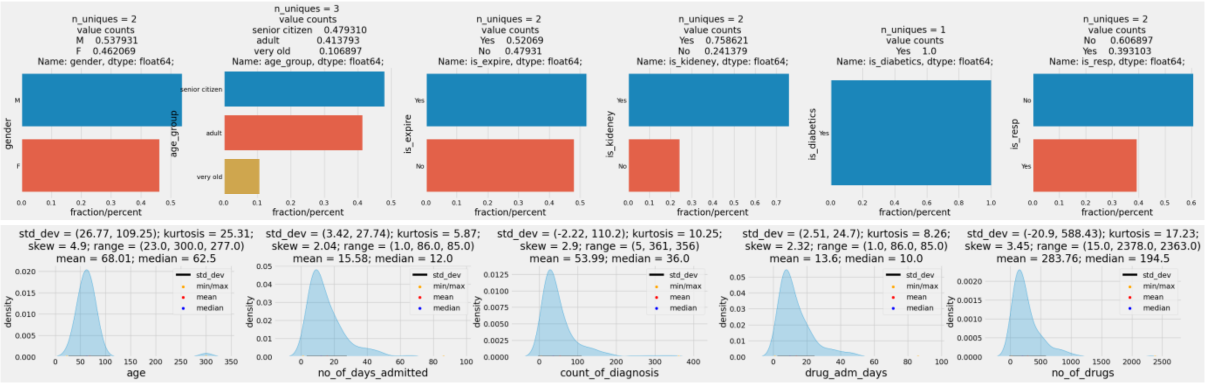


Figure No. 12.9 Cluster 3 Results

Figure No 12.9 details are mentioned below

- Both Females and males are equally distributed and most of the patients are adults and senior citizen and very few very old age people and people are died in this cluster is 50 % lesser than not died people
- Most of the people has kidney issues and all the people has diabetic issue, and most people has respiratory issues
- No of day's admitted is less and count of diagnosis is more and drug administrated days are more

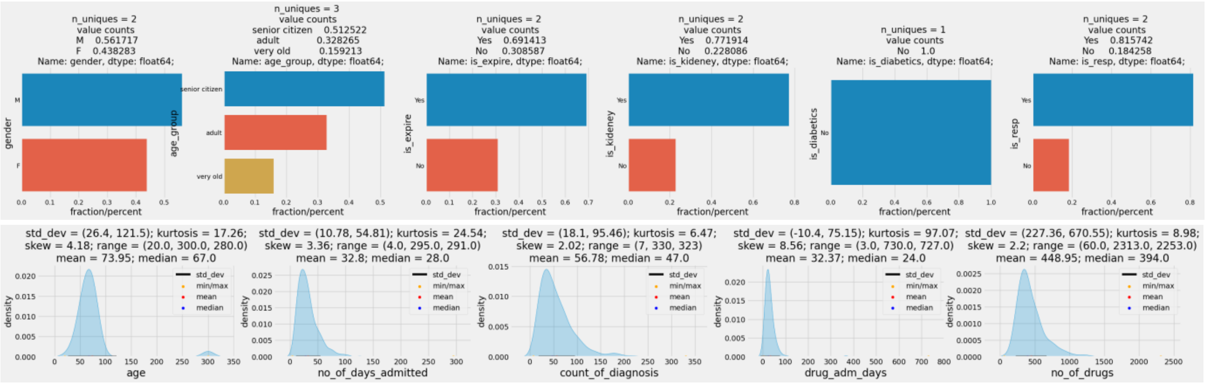


Figure No. 12.10 Cluster 4 Results

Figure No 12.10 details are mentioned below

- Both Male and Females are equally distributed and majority patients are senior citizen
- Most of them expired during the treatment and most of them have kidney issues and none of them had diabetic issue however majority suffered from the respiratory issues
- No of days admitted is more and count of diagnosis is more and patients consumed more drugs in this clusters

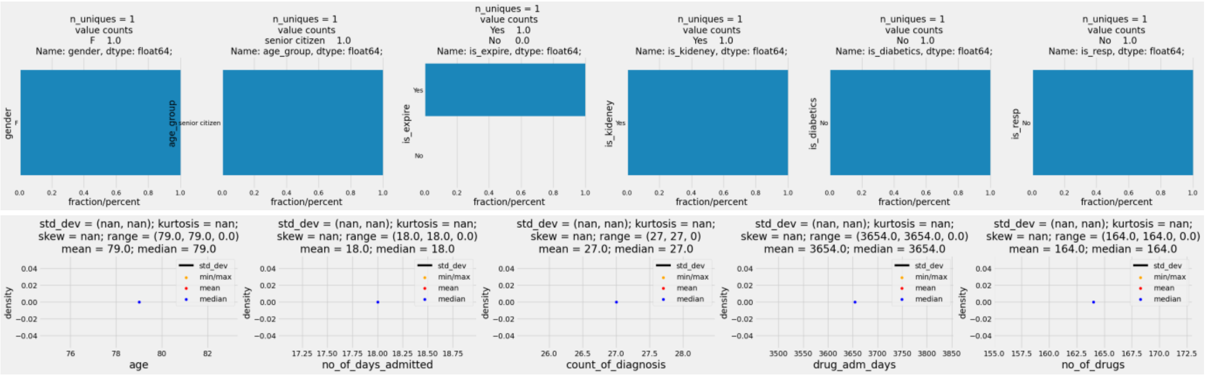


Figure No. 12.11 Cluster 5 Results

Figure No 12.11 details are mentioned below

- Only Females present in this cluster, and all are senior citizen and all are expired and every one suffered from kidney issue

- No of days admitted is less and count of diagnosis is more and drug administrated days are very high

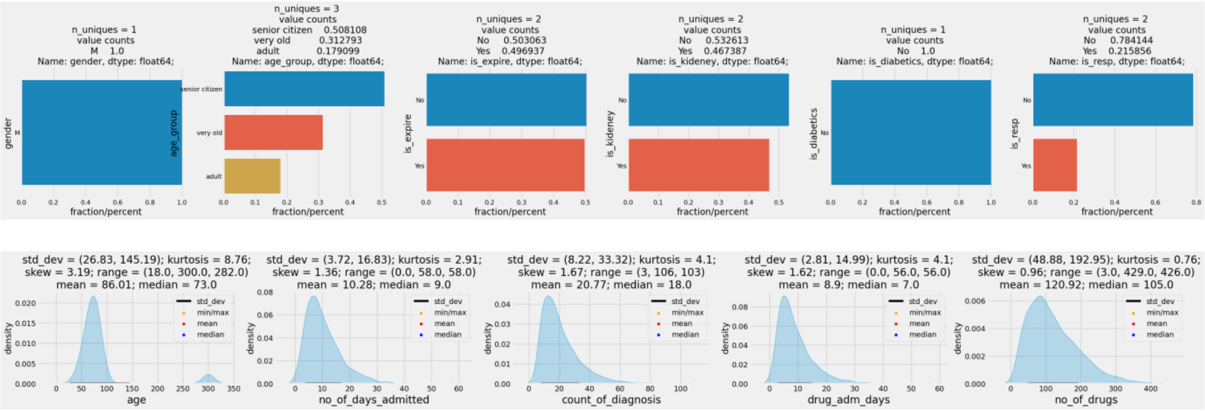


Figure No. 12.12 Cluster 6 Results

Figure No 12.12 details are mentioned below

- All are males in this cluster and everyone are senior citizen and both died and not died people are equally presented

- Most of them do not have kidney issue and none of them had diabetics and most of them had respiratory problem

Segmentation model helps the business to cluster the patients into different homogenous groups and each cluster has distinctive features that have been discussed earlier. By grouping patients into multiple clusters helps the business to act up on each patient separately. This helps in arranging treatments for each patient, diagnosis disease and prescription of medicine etc.

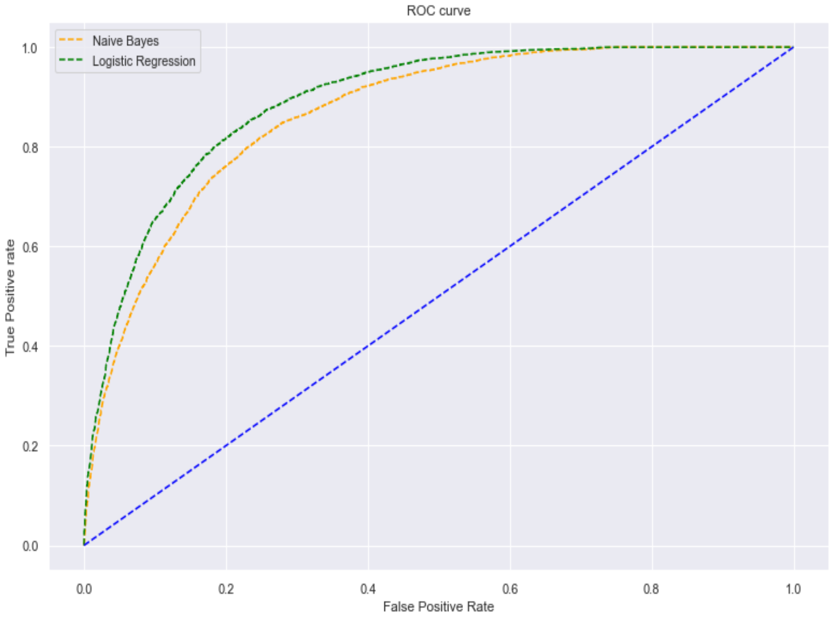


Figure No. 12.13 SAE Model Results

Table 12.1 Model Evaluation Metrics

Machine Learning Model	AUC Score
Logistic Regression	89%
Naïve Bayes	86%

Table No. 12.1 represents the model and the respective AUC score in the evaluation of test data. And Figure No. 12.13 shows the ROC curve of the logistic regression and naïve bayes models respectively. The occurrence of SAE in the duration of clinical trials must be prevented and this helps the business to ensure the safety of clinical trial participants. This model developed recommends the business the possibility of the occurrences of such events by utilizing patient's historic medical records text data to predict the probability of the occurrence. The test results show that the logistic regression model achieved this by 89 % of the time accurately.

Chapter 13: Conclusions and Future Scope

This work intended to provide a business solution to the health care industry and to ensure the safety of the patients and the proposed solution is to help the clinician and medical monitors to bring the EHR data to the app and gains insights and statics and in addition app facilitates the feature to segment the patients for chest heart failure and finally app recommends the predictability of the occurrence of serious adverse events in the duration of clinical trial.

In the Figure No 13.1 shows that including phenotyping models in conjunction with machine learning models to segment the patient population in the future work.

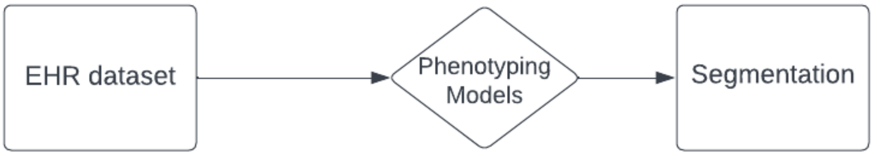


Figure No. 13.1 Future Works

In future the project indent to expand to add more sophisticated segmentation techniques for multiple diseases for applying phenotyping which should help the business to target the clusters to take decision on each cluster. In addition, the text classification for SAE prediction is implemented with generalized models and in next phase this model can be experimented with neural network with LSTM architecture to see whether sequential modeling helps to improvise the performance of the classification.

Bibliography

Adler Perotte, R. R. (2015). *Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis.*

Ahmed Alsayat, H. E.-S. (2016). *Efficient genetic K-Means clustering for health care knowledge discovery.*

Andrew J. Steele, S. C. (2018). *Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease.*

Aurelie Mascio, Z. K. (2020). *Comparative Analysis of Text Classification Approaches in Electronic Health Records.*

Berger, M. L. (2016). Opportunities and challenges in leveraging electronic health record data in oncology. .

Bittar, A. V. (2020). Text Classification to Inform Suicide Risk Assessment in Electronic Health Records.

BO JIN, C. C. (2018). *Predicting the Risk of Heart Failure With EHR Sequential Data Modeling.*

Churpek MM, Y. T. (2014). Using electronic health record data to develop and validate a prediction model for adverse outcomes in the wards.

Eichler, H. G.-D. (2019). Data rich, information poor: can we use electronic health records to create a learning .

Estiri, H. K. (2019). A clustering approach for detecting implausible observation values in electronic health records data. .

Gabriele Spini, M. v. (2019). *Private Hospital Workflow Optimization via Secure k -Means Clustering.*

Hubbard, R. A. (2021). Studying pediatric health outcomes with electronic health records using Bayesian clustering and trajectory analysis. .

Irene Li, J. P. (2021). *Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review.*

Jose Roberto Ayala Solaresa, b. F. (2019). *Deep learning for electronic health records: A comparative review of multiple deep neural architectures.*

LÜTZ, E. (2019). *Unsupervised machine learning to detect patient subgroups in electronic health records.*

Mantas, J. (2020). Unsupervised machine learning for the discovery of latent clusters in COVID-19 patients using electronic health records.

Mascio, A. K. (2020). Comparative analysis of text classification approaches in electronic health records.

Pai, M. M. (2021). Standard electronic health record (EHR) framework for Indian healthcare system. Health Services and Outcomes Research Methodology.

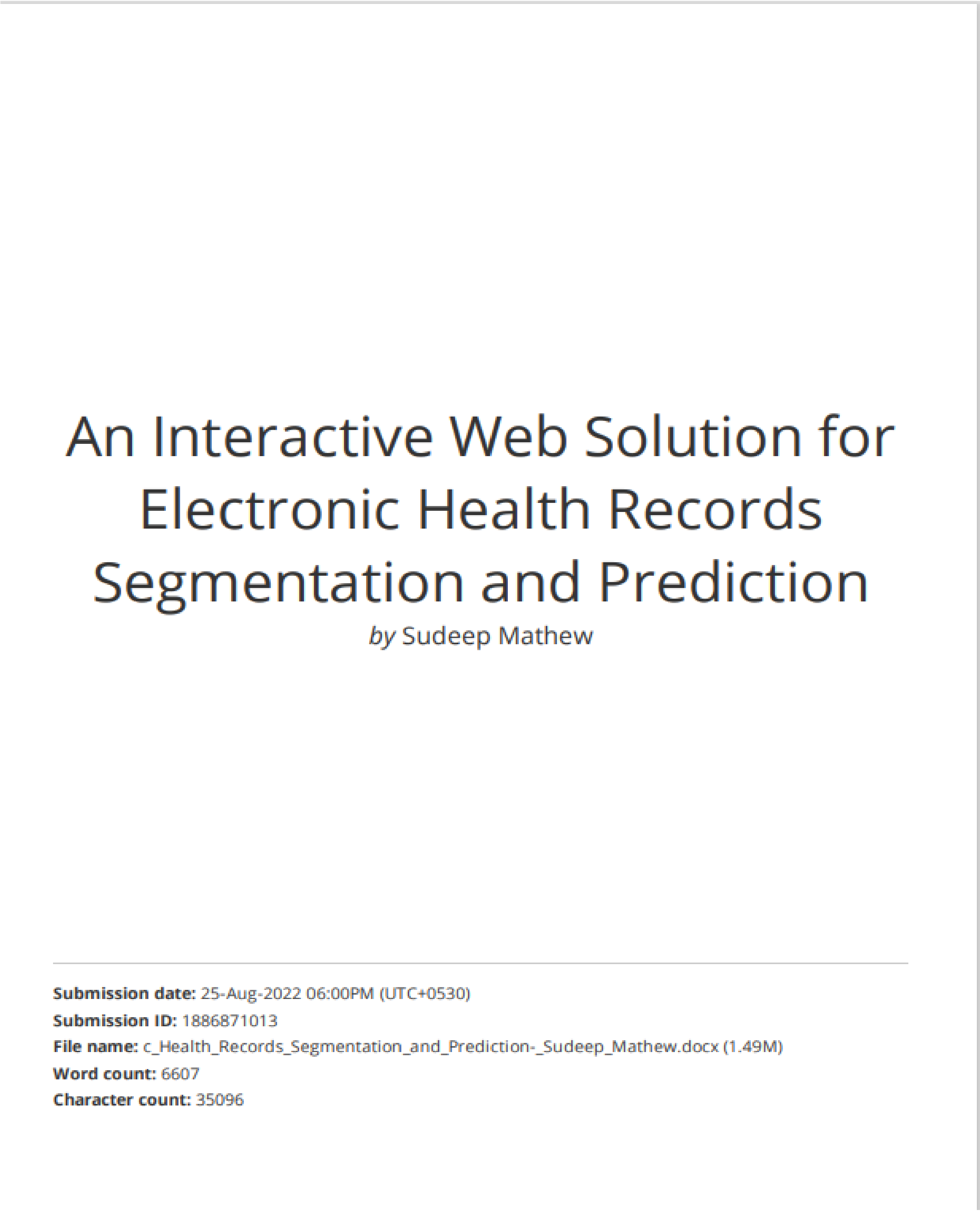
Wang, Y. Z. (2020). Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records.

Yadav, P. S. (2018). Mining electronic health records (EHRs) A survey. ACM Computing Surveys.

Ziyi Liu, J. Z. (2016). Machine Learning for Multimodal Electronic Health Records-based.

Appendix

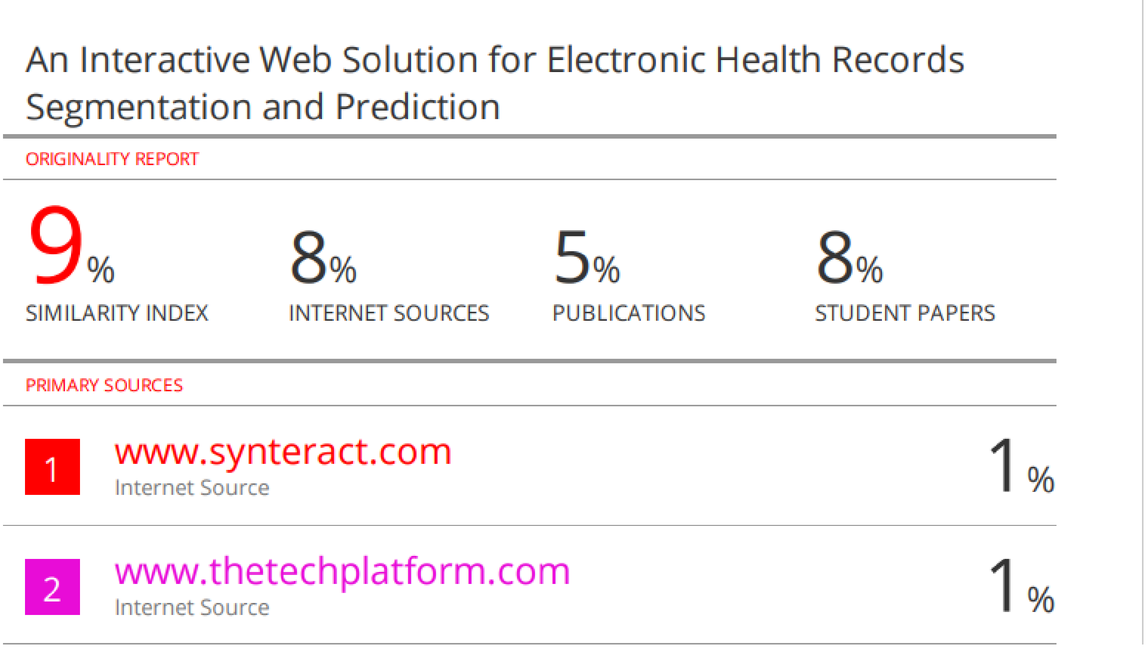
Plagiarism Report



An Interactive Web Solution for Electronic Health Records Segmentation and Prediction

by Sudeep Mathew

Submission date: 25-Aug-2022 06:00PM (UTC+0530)
Submission ID: 1886871013
File name: c_Health_Records_Segmentation_and_Prediction-_Sudeep_Mathew.docx (1.49M)
Word count: 6607
Character count: 35096



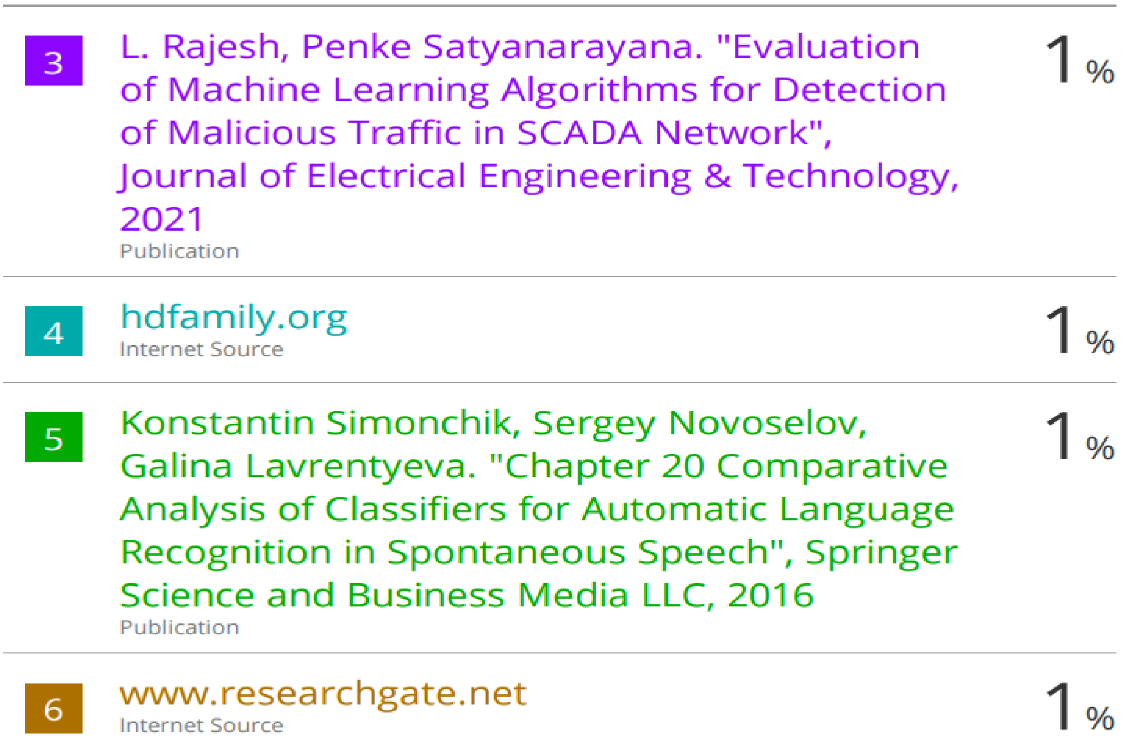
An Interactive Web Solution for Electronic Health Records Segmentation and Prediction

ORIGINALITY REPORT

9%	8%	5%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

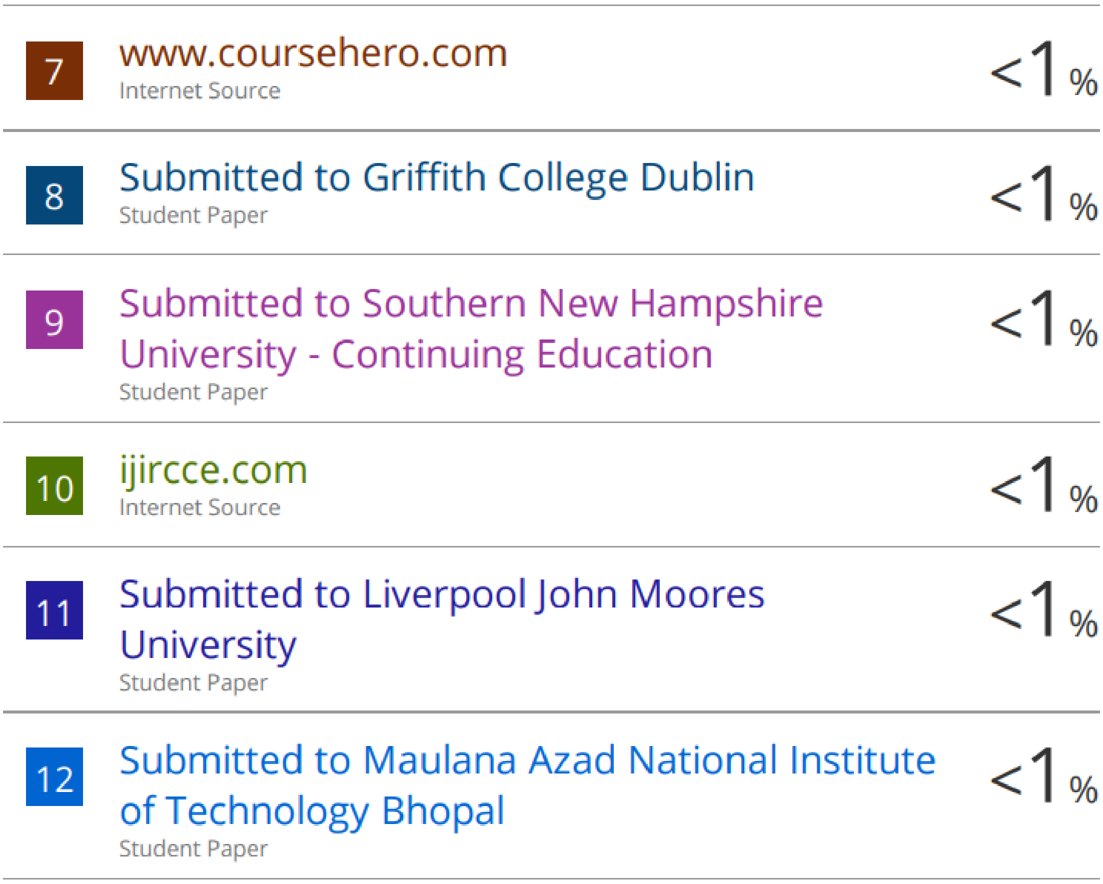
PRIMARY SOURCES

1	www.synteract.com Internet Source	1 %
2	www.thetechplatform.com Internet Source	1 %

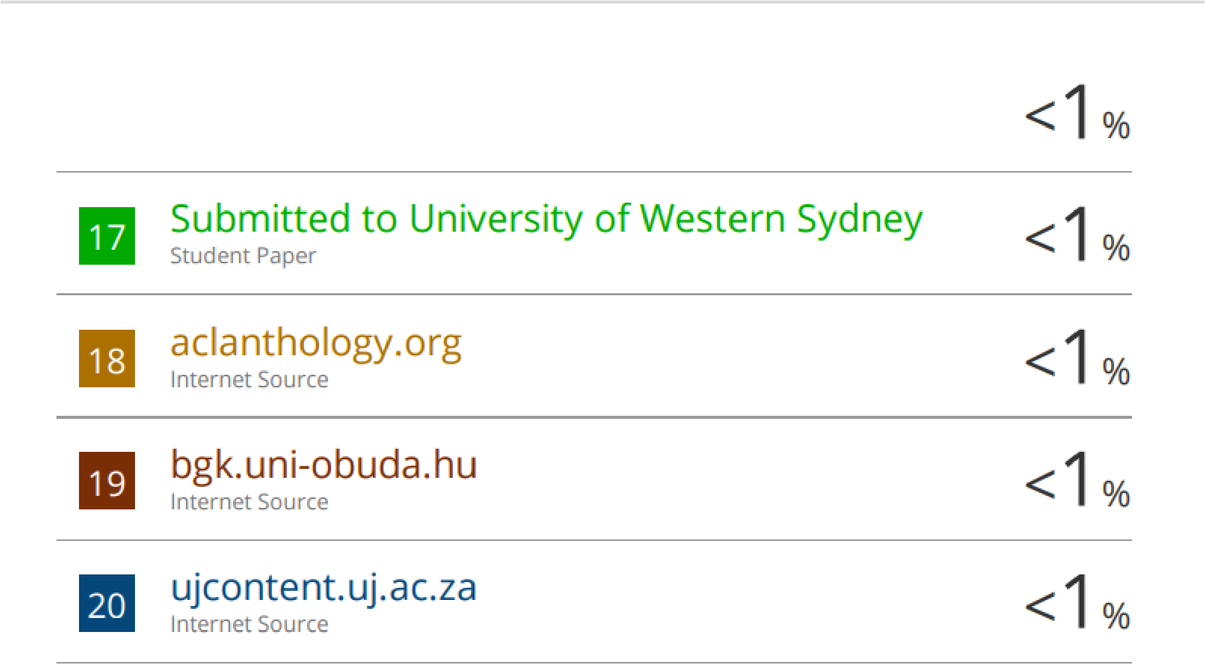


3	L. Rajesh, Penke Satyanarayana. "Evaluation of Machine Learning Algorithms for Detection of Malicious Traffic in SCADA Network", Journal of Electrical Engineering & Technology, 2021 Publication	1 %
4	hdfamily.org Internet Source	1 %
5	Konstantin Simonchik, Sergey Novoselov, Galina Lavrentyeva. "Chapter 20 Comparative Analysis of Classifiers for Automatic Language Recognition in Spontaneous Speech", Springer Science and Business Media LLC, 2016 Publication	1 %
6	www.researchgate.net Internet Source	1 %

13	Submitted to University of Nicosia Student Paper	<1 %
14	Bo Jin, Chao Che, Zhen Liu, Shulong Zhang, Xiaomeng Yin, Xiaopeng Wei. "Predicting the Risk of Heart Failure With EHR Sequential Data Modeling", IEEE Access, 2018 Publication	<1 %
15	academic.oup.com Internet Source	<1 %
16	www.ncbi.nlm.nih.gov Internet Source	



7	www.coursehero.com Internet Source	<1 %
8	Submitted to Griffith College Dublin Student Paper	<1 %
9	Submitted to Southern New Hampshire University - Continuing Education Student Paper	<1 %
10	ijircce.com Internet Source	<1 %
11	Submitted to Liverpool John Moores University Student Paper	<1 %
12	Submitted to Maulana Azad National Institute of Technology Bhopal Student Paper	<1 %



		<1 %
17	Submitted to University of Western Sydney Student Paper	<1 %
18	aclanthology.org Internet Source	<1 %
19	bgk.uni-obuda.hu Internet Source	<1 %
20	ujcontent.uj.ac.za Internet Source	<1 %

Publication in a Conference Presented

Paper Submitted:

Sudeep Mathew, Mithun Dolthody Jayaprakash, Rashmi Agarwal, "**An Interactive Web Solution for Electronic Health Records Segmentation and Prediction**", EAI ICISML 2022, EAI International Conference on Intelligent Systems and Machine Learning, Date of submission: 19-October-2022.

An Interactive Web
Solution for Electronic

Github Link

https://github.com/sudeepmathew/EHR-Patients-data-segementation-prediction

An Interactive Web Solution for Electronic Health Records Segmentation and Prediction

Sudeep Mathew[1], Mithun Dolthody Jayaprakash[2] and Rashmi Agarwal[3]

[1,2,3]REVA Academy for CorporateExcellence,
REVA University,Bengaluru, India

Abstract. A vast variety of patient data has been collected and monitored through Electronic Health Records (EHR) using various tools in the clinical research industry and it is a concern for healthcare providers to ensure the safety of the patients who are participating in the clinical trials. It is evident that need for a centralized analytics solutions for EHR datasets that deliver insights and predictability.

The paper focuses on the healthcare industry, which can benefit immensely by allowing medical practitioners to gain insights into the EHR data. The paper aims to provide a platform to explore and gain descriptive statistics and to provide patient segmentation and recommendation.

The objective of the paper is to start data acquisition and data understanding and then create a web interface for data exploration and segmentation and classification. In the data modeling phase, the objective is to create machine learning models for segmentation and classification.

The first step is data acquisition from the *MIMIC-III v1.4* (Clinical database) data mart. In the data understanding phase, the relationship of multiple tables is evaluated. In the data wrangling phase, SQL and Python are used to combine different tables to create a single dataset for analyzing the data and modeling the data. The combined dataset is then used for k-means clustering techniques for obtaining chest heart failure patients clusters. In the following phase, the diagnosis text data is extracted from the diagnosis dataset and performed text cleaning by removing punctuation, numbers, and stopwords. The cleaned text data is used for data modeling and for that TFIDF (Term Frequency Inverse Document Frequency) vectors and count vectors are created and then multiple classification techniques are applied for predicting the occurrences of death and the best model is considered for the model deployment.

In the model evaluation phase, it is observed that six clusters were optimal while training the model and it is incorporated into the application for predicting the segments of the patients based on the risk levels. Few machine learning models were trained on patient's historic diagnosis text data and the logistic regression model indicated 89 % of AUC score in test data and is deployed into the application for the prediction.

Finally, a web interface is created using the python *streamlit* framework which allows the users to bring raw EHR datasets to explore the data. The created models for segmentation and classification are deployed with the web application and thus will provide a recommendation to the business.

Keywords: Natural Language Processing, EHR, Segmentation, Serious Adverse Event Prediction

Keywords: Natural Language Processing, EHR, Segmentation, Serious Adverse Event Prediction

1 Introduction

Electronic health records (EHRs) contain patient diagnostic records, physician records, and records of hospital departments. For heart diseases, we can receive huge unstructured data from EHR time series. By analyzing and mining, we can identify the links between diagnostic events and ultimately predict the probability of the occurrence of a serious adverse event. The adoption of EHR datasets and the increase of digitized information about patient data revolutionize the emergence of clinical research in oncology research [1]. One of the applications of EHR data is an improvising learning system for clinical research and which helps in various applications of patient selection, dosing, drug target, etc. as discussed [2]. Standardizing electronic health records in the Indian health record system is implemented by [3]. The comprehensive techniques for modeling EHR data are provided by [4].

The web app is developed by aiming to help the medical or clinical team to monitor the safety of the patients during the clinical trials by allowing the users to explore the data through visualization and statistics, clustering, and the probability of the occurrence of SAE.

Clustering or segmentation techniques are helpful to find out the underlying hidden association between each data point. This helps the business to decide on each cluster. Detection anomalies or outliers in EHR data was implemented by [5]. The primary objective is to apply unsupervised clustering techniques to EHR data the result indicated that clustering techniques produced high sensitivity and specificity.

The occurrence of serious adverse events or SAE is one of the primary concerns that pharmaceutical companies face during the clinical trial. This application intended to solve this problem by providing the probability of the occurrence of SAE by analyzing patient's diagnosis data. In one of the works, a prediction model was implemented to detect the occurrence of an adverse event such as cardiac arrest by utilizing patient data [6].

The paper primarily focuses on the clinical research industries team which helps to improve overall the patient's safety concerns and address key issues during the clinical trial.

2 Background

2.1 Application of ML on HER Data

One of the works by Ziyi was referring to the challenges and the perspectives of machine learning multimodal in electronic health records. And this work suggests that including structured data is not enough to achieve a good result instead this study seeks to use machine learning and deep learning models on structured and unstructured EHR datasets [7]. The machine learning models in electronic health records could outperform conventional survival models for predicting mortality in coronary disease. These works include multiple machine learning models such as the cox model, and random forest in the 80000 patients EHR dataset and the output indicated that it outperforms conventional models [8]. Adeler proposed a risk prediction model for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. This work was to combine the time series model and cox proportional model to get a range of risk prediction models and the evaluation of the model is based on discriminatory statistics [9]. Predicting the risk of heart failure with EHR sequential data modeling is developed and used patient's diagnosis data with the LSTM sequential model used and the evaluation is based on the utility and efficacy of the proposed solution [10].

2.2 Unsupervised Techniques on HER Data

Lutz Has performed an unsupervised machine learning model to detect patient subgroups in electronic health records. This project used agglomerative hierarchical clustering and *k-means* clustering on patient's lab and coded datasets. The results indicated that natural grouping is present in the dataset and hierarchical clustering provides higher quality clusters than *k-means* clustering [11]. One of the works implemented by Gabriele is for private hospital workflow optimization through *k-means* clustering. This work is to optimize work allocation-based staff members, patients, hospitals, and locations to cluster the staff members utilizing the frequency of the facing time [12]. *k-means* clustering is used for healthcare knowledge discovery is one of the works and the objective is used to discover hidden patterns by applying *k-means* clustering and a self-organizing map (MAP)[13]. One of the works is an unsupervised machine learning model used for the discovery of latent disease clusters using electronic health records [14]. Prediction of health outcomes for pediatric patients is one of the works implemented and the approach of the project was to implement a Bayesian model and clustering for predicting the risk of type 2 diabetics for children between the ages of 10 and 14 [15]. Another work implemented using an unsupervised LDA approach to cluster patient subgroups into multiple clusters using patient's health records [16].

2.3 Text Analytics on HER Data

Natural language processing is used in the field of unstructured text data and this work explores the possibilities of applying NLP models in EHR datasets. The author discussed various applications of NLP such as classification models, question answering, phenotyping, knowledge graphs, medical dialogue, etc. [17]. A comparative analysis of text classification approaches in electronic health records indicated that text classification in traditional approaches could exceed the performance of contextual embedding models such as BERT [18]. A work mentioning the application of deep learning models in an electronic health record by developing various deep neural network models. The proposed solution is a deep learning model for predicting the health risk of the patient [19]. On other hand, a work implemented by Mascio suggests different text classification approaches to patient text data [20]. This work focused on various traditional machine learning models and compared them with contextual embedding BERT and identified that traditional models performed well on the text data. Bittar tried to implement suicide risk assessment using text data and the work implemented to predict the tendency to commit suicide by extracting text features from the clinical notes and the data trained SVM model [21].

Various machine learning techniques on electronic health records were discussed above on segmentation as well as text classification. This paper seeks to predictability of death of the patient based on the historic text data as well segmentation of HER data.

3 Methodology

In contrast to the above-mentioned methods, we develop a *k-means* clustering technique to group Congestive Heart Failure (CHF) patients based on risk levels. We also developed a Serious Adverse Event (SAE) prediction, model for predicting the probability of death using the text classification technique. We present the details of our approach in the following session.

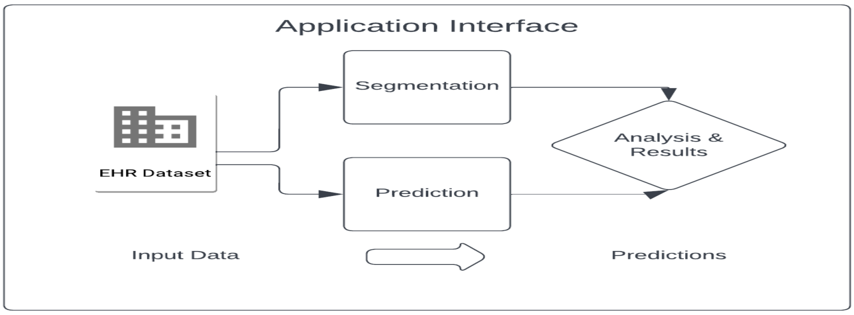


Fig. 1. Application Interface

In Fig. 1 the details of the application are represented. In the following sessions, the details of the application describing.

3.1 K-means Clustering on CHF Data

The input data for the clustering or segmentation technique is combined data of patient demographic, admission, diagnosis, and drug administration. The following flowchart in Fig. 1. depicts the proposed methods for the *k-means* clustering technique.

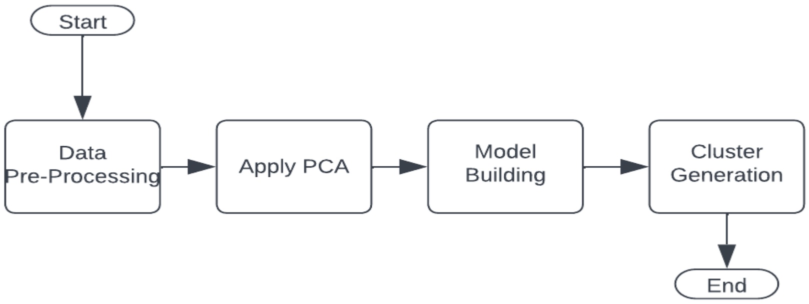


Fig. 2. Methodology for K-means Clustering

The input features containing patient's demographic, admission, prescription, and diagnosis datasets were joined and transferred to Principal Component Analysis (PCA) model for dimensionality reduction, and afterward, the principal components are transferred to the *k-means* model for the segmentation of the patients. This model helps the business to identify the risk levels of the patients to take a valid decision for each cluster.

Data Pre-Processing and Feature Engineering. The output of the data wrangling created a single data frame for the data analysis and data modeling. The primary diagnosis of the patients with this CHF disease formed a data frame in the previous data wrangling steps. To perform segmentation, data scaling and missing records removal were the two methods performed before the data modeling phase. Data scaling is the approach to normalizing the range of independent features of the dataset. And finally, all the missing values are removed from the data frame by the missing value removal method in python. Table No.1 lists all the features that were generated.

Table 1. Features Of Segmentation

No	Name	Description
1	Count of diagnosis	Aggregated diagnosis count for each patient
2	Drug administrated days	Aggregated total days of drugs given for each patient
3	No of drugs	Total number of drugs given to each patient
4	Age Group	Grouped patients based on age such as >90 as very old, 60 -80 as a senior citizen, >18 as an adult, <18 as young
5	Ethnic Group	All the non-white people grouped as other as others as white
6	Is hyper	Whether hypertension present for patient 0 as not present and 1 as present
7	Is kidney	Whether kidney diseases present for patient 0 as not present and 1 as present
8	Is diabetic	Whether diabetic present for patient 0 as not present and 1 as present
9	Is resp	Whether respiratory diseases present for patient 0 as not present and 1 as present

Applying PCA Technique. Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms a set of in a dataset into a smaller number of features called *principal components* while at the same time trying to retain as much information in the original dataset as possible. Preprocessed data was transferred to the PCA model and the output was four components.

Model Building. The *k-means* algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known as the clustering or segmentation technique. Below are the stages of clustering.

1. First provided the number of the cluster that we need to generate from the algorithm
2. Next, choose K data points at random and assign to each cluster
3. The cluster centroid is computed
4. Iterate the below steps until the ideal centroid is met, which the assigning of data points similar into the same clusters and heterogenous into other clusters

The sum of the squared distance between data points and cluster centroid is calculated first and allocated to the data points similar in the same clusters. Clustering is an unsupervised approach where the hidden association in the data can be extracted. The input PCA components are transferred to the *k-means* algorithm and as an outcome, the ideal clusters were obtained.

3.2 SAE Text Data Classification

The input data for SAE classification is extracted from the patient diagnosis dataset and it consists of text data features. The text data classification process flow chart is in Fig. 3. provided below.

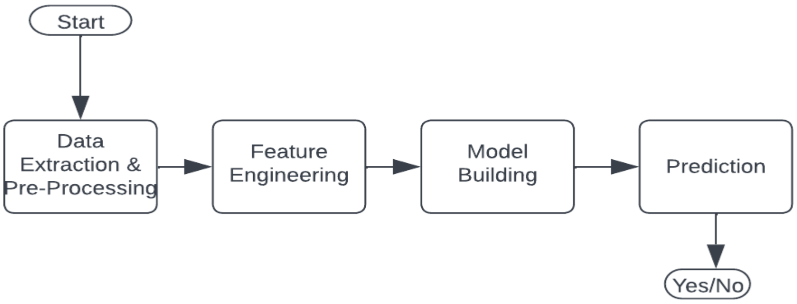


Fig. 3. Methodology for SAE Classification

Text Data Pre-Processing and Feature Engineering. The data frame after the wrangling step contains one feature containing patient's diagnosis text data and the label for the serious adverse event is a binary field called "expire flag" which tells whether the patient had observed any serious adverse event or not. The input text data had gone through various text pre-processing techniques before data modeling. Text pre-processing steps include text normalization, removing special characters and numbers, and stop ward removal. These steps are applied to each patient's diagnosis text data for better results in modeling. Text normalization is the process of converting all the text data into lower cases. Removal of unwanted special characters and integers is an important step in text pre-processing hence those should not need for text prediction. Finally, stop words are non-important words in the text and are the most occurred words and which may not provide the overall semantic meaning of the text. The cleaned text diagnosis data is converted into input features for the SAE prediction model. Count vectorizer and Term Frequency Inverse Document Frequency (TFIDF) vectorizer are used for creating features for the text data. Count vectorizer is a method to convert text to numerical data by considering the count of the word in each sentence. On the other hand, TFIDF is better than the count vectorizer because it not only focuses on the frequency of words present in the corpus but also provides the importance of the words. Fig. 4. displays the features created for building the text classification model.

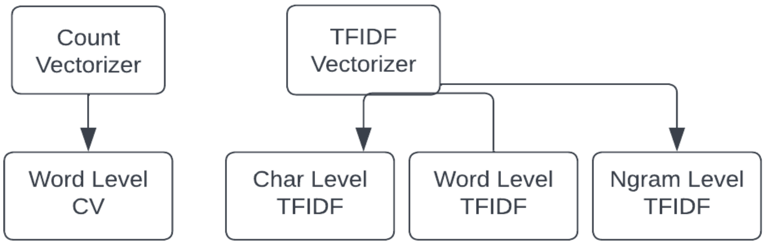


Fig. 4. Text Feature Engineering

Model Building. The classification model was developed using text data classification. Patient's historic diagnosis text data was collected and joined together for all the patients in the database and multiple classification techniques were trained on the dataset to obtain the best-fitted model. The output of the model is a binary classification that tells whether the probability of a serious adverse can occur in the future.

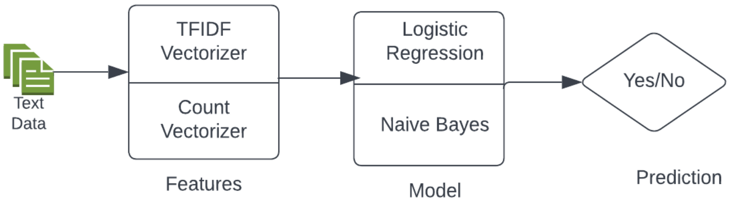


Fig. 5. Design of Text Classification Model

Fig. 5. depicts the workflow of the patient SAE prediction application. Patient's diagnosis text data converted to TFIDF and count vectorizer as features for training the model. Term frequency-inverse document frequency is a text vectorizer that transforms the text into a usable vector. It combines 2 concepts, Term Frequency (TF) and Document Frequency (DF). A count vectorizer is used to transform a given text into a vector based on the frequency (count) of each word that occurs in the entire text. Converted features are then passed as input for the model such as logistic regression and a naïve Bayes which in this case is supervised binary classification techniques. The term binary classification is referred to because the outcome of the model is binary or yes/no classification. In statistics, the (binary) logistic model (or logit model) is a statistical model that models the probability of one event (out of two alternatives) taking place by having the log-odds (the logarithm of the odds) for the event be a linear combination of one or more independent variables ("predictors"). Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying the Bayes theorem with strong (naive) independence assumptions between the features. Abstractly, naive Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector representing some *n* features (independent variables), it assigns to this instance probabilities. Features are trained on both models and the best-resulting model is considered for deployment.

4 **Data Analysis**

The data set is collected from MIMIC III data mart v.1.4 and it consists of 46000 patient's data [22]. The data contains patient's admission to the hospital. The data used in this work consists of patient's demographics, admission, diagnosis, and prescription datasets. Table No. 2 lists the datasets and the description.

Table 2. Datasets And Description

No	Dataset	Description
1	Patients	Contains the demographic data for each patient's
2	Admissions	Consists of unique records of patient's admission to the hospital
3	D_icd_diagnosis	Standard 1cd9_code and label for different diagnoses which is a standard dataset
4	Diagnoses_Icd	Diagnosis contains icd9_code for each patient's visits
5	Prescriptions	Data is related to the drugs administrated to each patient during the admission to the hospital

The Patients. Admission, D_icd_diagnosis, Diagnosis_ICD, and prescription dataset is combined using SQL for data wrangling and extracted chest heart failure dataset for segmentation. Diagnosis dataset contained the patient's diagnosis data which were used for SAE classification.

5 Deployment

The usability of the solution is something that businesses looking for as there are a lot of solution builds but all of them cannot be used. Hence it is important to consider that all solutions that are provided must be in user-friendly format that it must easily interact with the users of the app. A web solution is build using *streamlit* framework by integrating the model and all other widgets for the usability of the app. The developed models were promoted for deployment by using python *streamlit* framework which is lightweight framework used to construct a python-based web dashboard and analytical widget for the data science solution. Application consists of three distinctive features which are exploratory data analysis app, a chest heart failure clustering model build on the *k-means* algorithm and a logistic regression model for SAE classification are exported as a model package for the deployment.

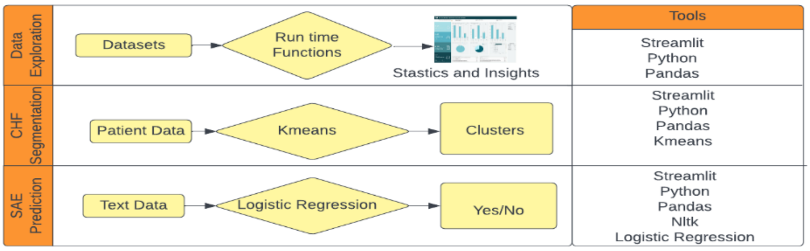


Fig. 6. Deployment Flow and Tools Used

5.1 Software Used

Python. Python is used for the building of the application. Several functions build in python for handling data wrangling and data transformation primarily for dataset exploration. Various dataset was joined together using pandas merging and aggregating functionality.

Streamlit. *Streamlit* framework provides widgets and graphs to display effectively and integrate all the building blocks of the apps into a web solution and the streamlit server is used for deploying the App.

Nltk. *Nltk* is a set of packages used for creating a text classification model which is mainly used for data cleaning steps such as the removal of characters, and stop words.

Keras. *Keras* library provides python classes and objects for various machine learning model building and Keras provided *k-means* and Logistic regression classes for building the model and later exported as a package for the deployment of the application.

6 Evaluation and Results

The evaluation of the performance of the models is essential to pick the best performance model for deploying the solution. Each modeling techniques have different types of the evaluation model. In contrast to supervised learning where we have the ground truth to evaluate the model's performance, clustering analysis does not have a solid evaluation metric that we can use for evaluating the outcome of the model. Elbow method and silhouette analysis where the two types of evaluation metrics used in the *k-means* clustering algorithm. Unlike clustering for SAE prediction, we have the outcome of the patients in the test dataset to evaluate the performance of the model. There are various performance metrics are widely used in the supervised classification model in which the most common ones are the F1 ratio, AUC score, AUC Curve plotting, Log ratio, etc. In this paper, Elbow method is used for clustering and the AUC curve is used for classification.

The Elbow method gives us an idea on what a good *k* number of clusters would be based on the Sum of Squared distance (SSE) between data points and their assigned clusters' centroids. AUC - ROC curve is a performance measurement for classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.

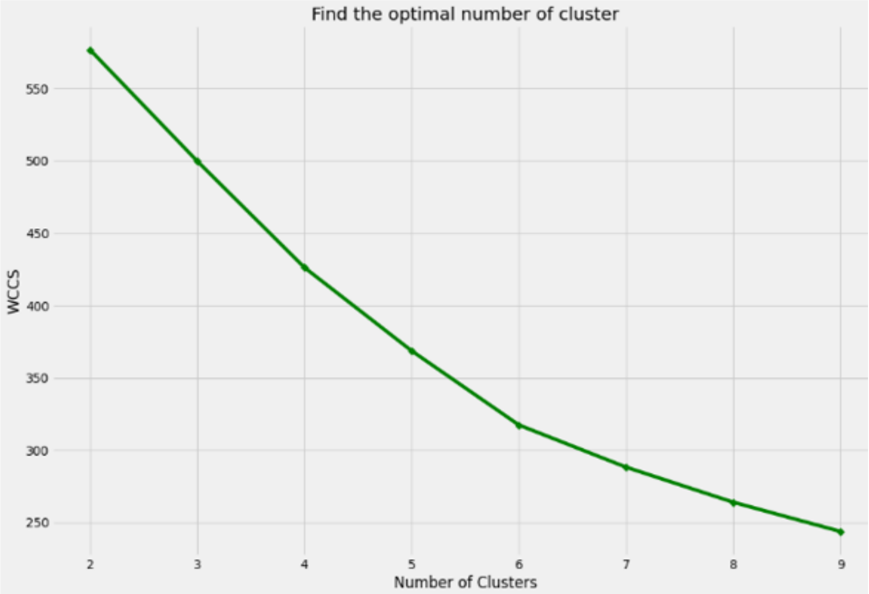


Fig. 7. WCCS Plot for Clustering

From Fig. 7. the y-axis shows the WCCS (Within-Cluster Sun of Square) and the x-axis plot the number of clusters the ideal number of clusters that are used in this work is six as the WCCS point is steady releases till the cluster number six and afterward the decentness of the cluster of reduced slightly which implies the optimal cluster might be six.

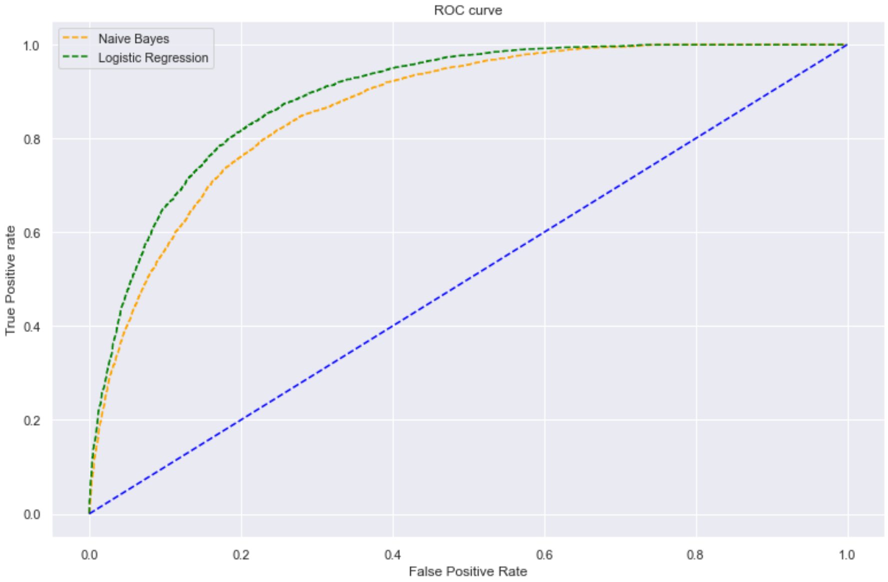


Fig. 8. SAE Classifier AUC Curve

Fig. 8. shows the ROC curve plotted with TPR against the FPR where TPR is on the y-axis and FPR is on the x-axis. An excellent model has AUC near the 1 which means it has a good measure of separability. A poor model has an AUC near 0 which means it has the worst measure of separability. In this work Logistic Regression and Naïve Bayes model's performance was good and per the figure, the logistic regression shows the higher the curve and so it is considered the best model for the classification problem.

7 Conclusion

In this paper, we have presented an efficient way of *k-means* clustering technique that can be used to identify groups and associations for chest heart failure. The outcome of the clustering technique indicated that it succeeded in grouping six clusters for grouping the data effectively. In addition, SAE classification techniques are implemented by which patient diagnosis text data is used for modeling and the outcome flags the serious threat to the life of the patient. The logistic Regression algorithm showed 89 % of the AUC score in successfully classifying the data.

References

[1] M. L. Berger, M. D. Curtis, G. Smith, J. Harnett, and A. P. Abernethy, “Opportunities and challenges in leveraging electronic health record data in oncology,” *Future Oncol*, vol. 12, no. 10, pp. 1261–1274, May 2016, doi: 10.2217/FON-2015-0043.

[2] H. G. Eichler *et al.*, “Data Rich, Information Poor: Can We Use Electronic Health Records to Create a Learning Healthcare System for Pharmaceuticals?,” *Clin Pharmacol Ther*, vol. 105, no. 4, p. 912, Apr. 2019, doi: 10.1002/CPT.1226.

[3] M. M. M. Pai, R. Ganiga, R. M. Pai, and R. K. Sinha, “Standard electronic health record (EHR) framework for Indian healthcare system,” *Health Serv Outcomes Res Methodol*, vol. 21, no. 3, pp. 339–362, Sep. 2021, doi: 10.1007/S10742-020-00238-0/FIGURES/9.

[4] P. Yadav, M. Steinbach, V. Kumar, and G. Simon, “Mining electronic health records (EHRs): A survey,” *ACM Comput Surv*, vol. 50, no. 6, Jan. 2018, doi: 10.1145/3127881.

[5] H. Estiri, J. G. Klann, and S. N. Murphy, “A clustering approach for detecting implausible observation values in electronic health records data,” *BMC Med Inform Decis Mak*, vol. 19, no. 1, Jul. 2019, doi: 10.1186/S12911-019-0852-6.

[6] M. M. Churpek, T. C. Yuen, S. Y. Park, R. Gibbons, and D. P. Edelson, “Using Electronic Health Record Data to Develop and Validate a Prediction Model for Adverse Outcomes on the Wards,” *Crit Care Med*, vol. 42, no. 4, p. 841, 2014, doi: 10.1097/CCM.0000000000000038.

[7] Z. Liu, J. Zhang, Y. Hou, X. Zhang, G. Li, and Y. Xiang, “Machine Learning for Multimodal Electronic Health Records-based Research: Challenges and Perspectives,” Nov. 2021, doi: 10.48550/arxiv.2111.04898.

[8] A. J. Steele, S. C. Denaxas, A. D. Shah, H. Hemingway, and N. M. Luscombe, “Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease,” *PLoS One*, vol. 13, no. 8, Aug. 2018, doi: 10.1371/JOURNAL.PONE.0202344.

[9] A. Perotte, R. Ranganath, J. S. Hirsch, D. Blei, and N. Elhadad, “Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis,” *J Am Med Inform Assoc*, vol. 22, no. 4, pp. 872–880, Jul. 2015, doi: 10.1093/JAMIA/OCV024.

[10] B. Jin, C. Che, Z. Liu, S. Zhang, X. Yin, and X. Wei, “Predicting the Risk of Heart Failure with EHR Sequential Data Modeling,” *IEEE Access*, vol. 6, pp. 9256–9261, Jan. 2018, doi: 10.1109/ACCESS.2017.2789324.

[11] E. Lütz, “Unsupervised learning to detect patient subgroups in electronic health records,” *DEGREE PROJECT COMPUTER SCIENCE AND ENGINEERING*, 2019.

[12] G. Spini, M. van Heesch, T. Veugen, and S. Chatterjea, “Private Hospital Workflow Optimization via Secure k-Means Clustering,” *J Med Syst*, vol. 44, no. 1, pp. 1–12, Jan. 2020, doi: 10.1007/S10916-019-1473-4/TABLES/5.

[13] M. Zubair, M. Asif Iqbal, A. Shil, E. Haque, M. Moshiul Hoque, and I. H. Sarker, “An Efficient K-means Clustering Algorithm for Analysing COVID-19,” *Advances in Intelligent Systems and Computing*, vol. 1375 AIST, pp. 422–432, Dec. 2020, doi: 10.48550/arxiv.2101.03140.

[14] Y. Wang *et al.*, “Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records,” *J Biomed Inform*, vol. 102, Feb. 2020, doi: 10.1016/J.JBI.2019.103364.

[15] R. A. Hubbard, J. Xu, R. Siegel, Y. Chen, and I. Eneli, “Studying pediatric health outcomes with electronic health records using Bayesian clustering and trajectory analysis,” *J Biomed Inform*, vol. 113, Jan. 2021, doi: 10.1016/J.JBI.2020.103654.

[16] W. Cui, D. Robins, and J. Finkelstein, “Unsupervised Machine Learning for the Discovery of Latent Clusters in COVID-19 Patients Using Electronic Health Records,” *Stud Health Technol Inform*, vol. 272, pp. 1–4, 2020, doi: 10.3233/SHTI200478.

[17] I. Li *et al.*, “Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review,” Jul. 2021, doi: 10.48550/arxiv.2107.02975.

[18] A. Mascio *et al.*, “Comparative Analysis of Text Classification Approaches in Electronic Health Records,” pp. 86–94, Jul. 2020, doi: 10.18653/V1/2020.BIONLP-1.9.

[19] J. R. Ayala Solares *et al.*, “Deep learning for electronic health records: A comparative review of multiple deep neural architectures,” *J Biomed Inform*, vol. 101, p. 103337, Jan. 2020, doi: 10.1016/J.JBI.2019.103337.

[20] A. Mascio *et al.*, “Comparative Analysis of Text Classification Approaches in Electronic Health Records,” May 2020, Accessed: Sep. 22, 2022. [Online]. Available: http://arxiv.org/abs/2005.06624

[21] A. Bittar, S. Velupillai, A. Roberts, and R. Dutta, “Text classification to inform suicide risk assessment in electronic health records,” *Stud Health Technol Inform*, vol. 264, pp. 40–44, Aug. 2019, doi: 10.3233/SHTI190179.

[22] A. E. W. Johnson *et al.*, “MIMIC-III, a freely accessible critical care database,” *Sci Data*, vol. 3, May 2016, doi: 10.1038/SDATA.2016.35.