REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

REVA Academy for Corporate Excellence (RACE)

An Interactive Web Solution for Electronic Health Records Segmentation and Prediction

Sudeep Mathew
SRN: R19MBA09
Date: 27/08/2022

MBA in Business Analytics

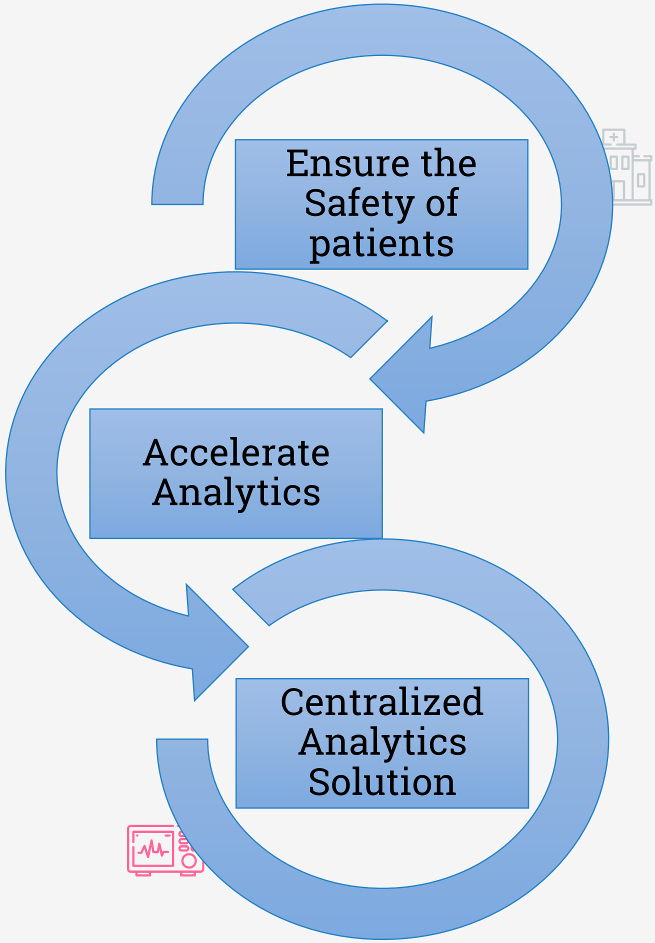
Capstone Project Presentation
Year: II

race.reva.edu.in

Agenda

Introduction

Electronic Health Record (EHR)
Is a source of meaningful insights to the
Patients health



REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,

Literature Review

Year	Author	Description
2021	Irine	Discussed about various NLP Application in the EHR dataset
2020	Aurelie, Macio	Works indicated that traditional classification model suitable best for the EHR text data classification
2016	Ziyi Liu	Indicated that structured data is not enough to get good accuracy but instead combining unstructured data will yield higher accuracy
2018	Bo jin	LSTM sequential model created for predicting the risk of heart failure
2019	Lutz	Mentioned that natural grouping is present in EHR data and hierarchical clustering provides higher quality clusters than kmeans
2021	Hubbard	Developed a machine learning model for predicting the risk of type 2 diabetics patients
2020	Mantas	LDA approach for segmenting patients EHR data

Business Problem

- **Early identification and prevention of disease,** and thereby **ensuring patient care** have been crucial steps for clinical research. Companies find it **difficult to analyze and interpret patients' electronic health records**.

- The medical or clinical team does not have a way to **explore the data and segment patients**.

- The prevention of the occurrence of a serious adverse event like the **probability of occurrence of death** must be prevented. Continuous monitoring of patients' EHR records and predictive analytics reduce the risk to patient's life.

Proposed Solution

- Web Interactive Application
- Exploratory Data Analytics Tool
- Patients Chest Heart Failure Segmentation Tool
- Patient Serious Adverse Event Prediction

Project Methodology

CSV
CSV
CSV

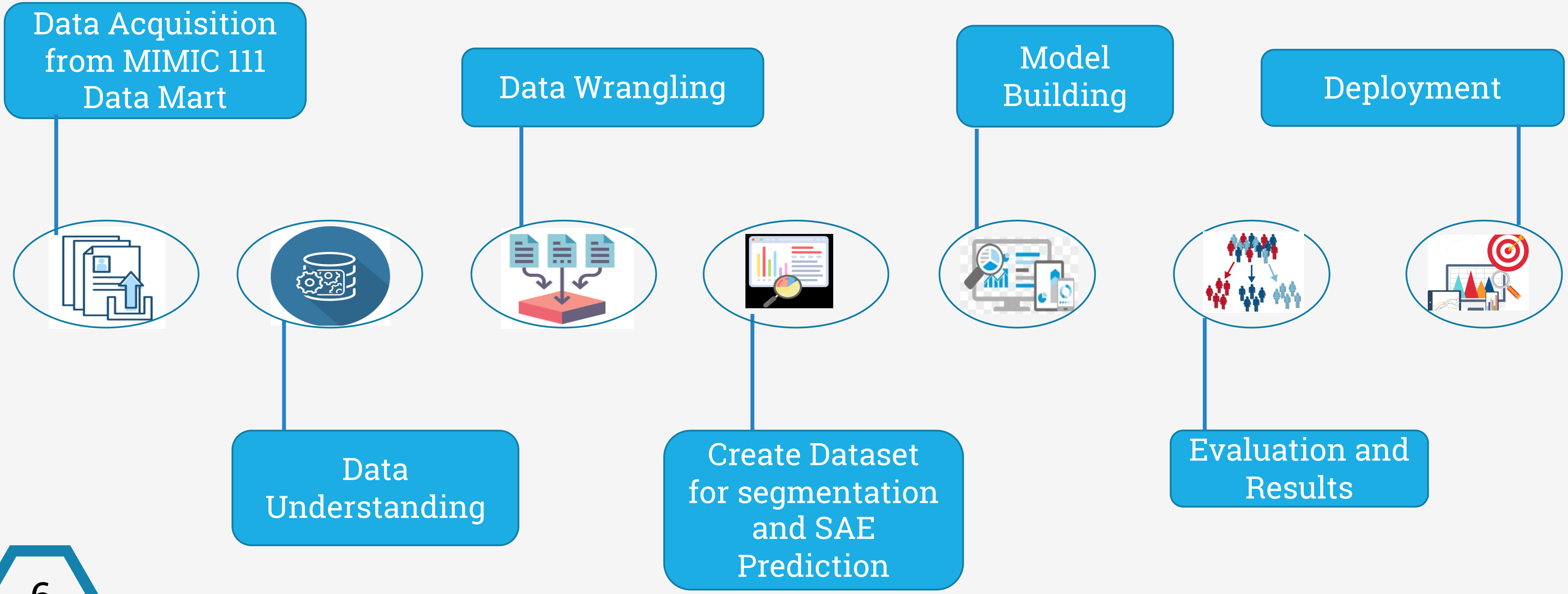
Exploration
Descriptive Analytics
Correlation
Statistics

Chest Heart Failure Segmentation

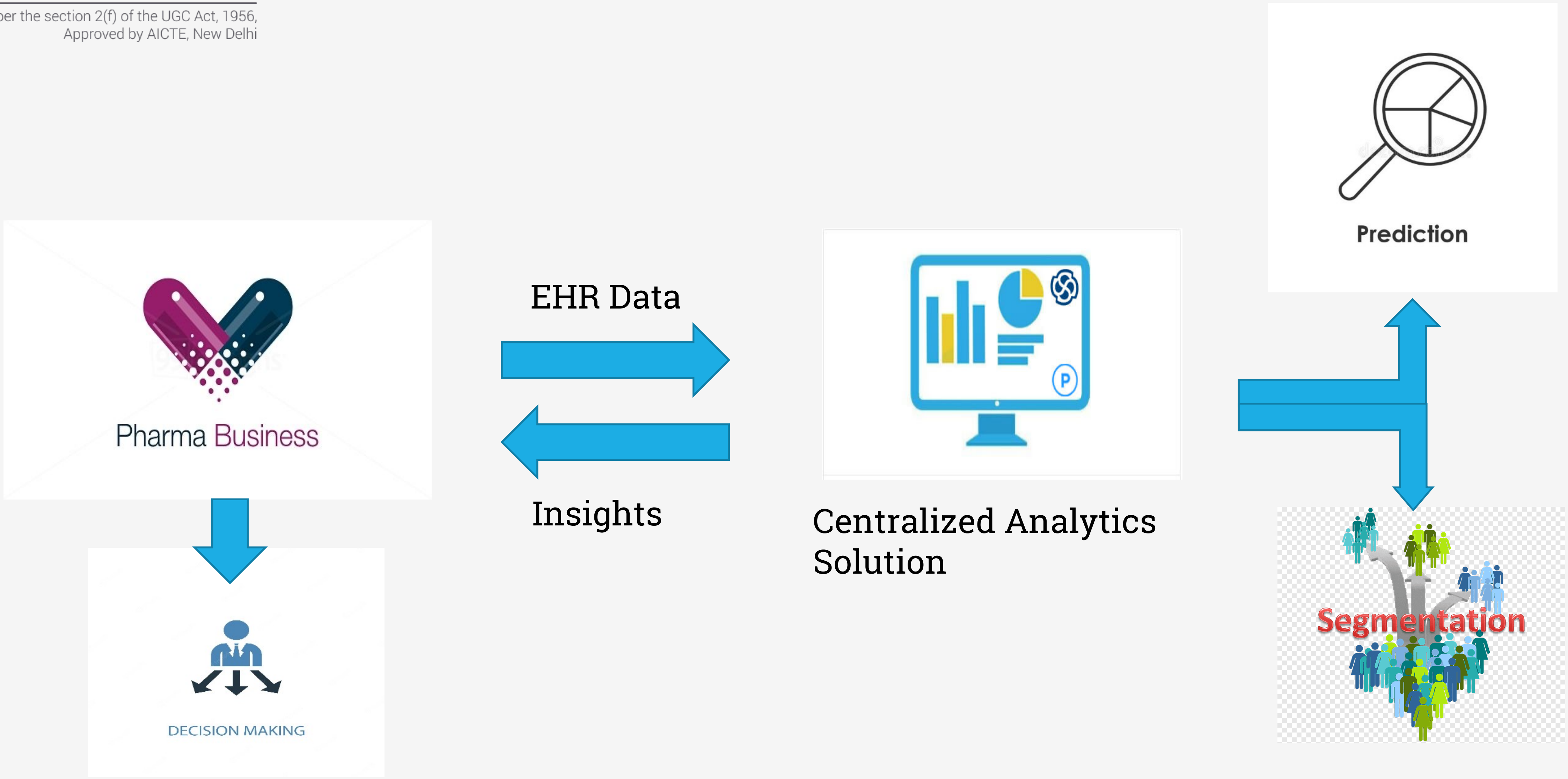
SAE Prediction
NLP Model
Diagnosis Text Data

EHR Datasets → Exploration → Segmentation → Prediction

Approach


Data Acquisition from MIMIC 111 Data Mart
Data Understanding
Data Wrangling
Create Dataset for segmentation and SAE Prediction
Model Building
Evaluation and Results
Deployment

Business Understanding



Data Understanding

Data has collected from MIMIC 111 Data Mart and it consists of 46000 patients electronic health records

Datasets	Description
Patients	Demographic data for unique patients
Admission	Consists of unique records
D_ICD_Diagnosis	Standard coding datasets for diagnosis
DIAGNOSIS_ICD	The standard dataset contains coded information
Prescription	Dataset related to the drug administrated to the patient

Data Understanding

Dataset
Patient
Admissions
Diagnosis ICD
Prescription
Diagnosis

Fields	Description
Subject_id	Unique id for all patients
Gender	Gender for each patients
DOB	Date of birth of the patients
DOD	Date of Death of the patients
DOD_HOSP	Date of death if the death at the hospital
EXPIRE_FLAG	Determine if the patients died or alive

Fields	Description
SUBJECT_ID	Unique id for all patients
HADM_ID	Unique id for every hospital admissions for each patients
ADMITTIME	Date and time of admissions
DISCHTIME	Date and time of discharge
DEATHTIME	Date of death if the death at the hospital
ADMISSION_TYPE	Admission type whether it is elective or emergency
ADMISSION_LOCATION	Location of the Admission
ETHNICITY	Ethnic of the patient
DIAGNOSIS	Diagnosis of the patients disease
HOSPITAL_EXPIRE_FLAG	Whether the patient dies in hospital or not

Fields	Description
SUBJECT_ID	Unique id for all patients
HADM_ID	Unique id for every hospital admissions for each patients
STARTDATE	Date and time of ICU admission date and time
ENDDATE	ICU end date time
DRUG	Drug name given to patient

Fields	Description
SUBJECT_ID	Subject_id for each patients
HADM_ID	Hospital Admission id for patients who are admitted in the hospital
ICD9_CODE	Dictionary code corresponds to patients diagnoses

Fields	Description
ICD9_CODE	Standard code for the diagnosis
SHORT_TITLE	Short title for each diagnosis
LONG_TITLE	Long Title for each diagnosis

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

Data Preparation

Segmentation

Patient EHR datasets

SQL

CSV

CHF Nos Dataset

Scaling

Missing Data Removal

Count of diagnosis

Drug administrated days

Is kidney

Is hyper

No of drugs

Ethnic Group

Age Group

Classification

Diagnosis Dataset

Text Data

Removal of special character

Stop word Removal

Word level count Vectorizer → Char level tfidf vectorizer → Word level tfidf vectorizer → Ngram level tfidf vectorizer

Data Wrangling

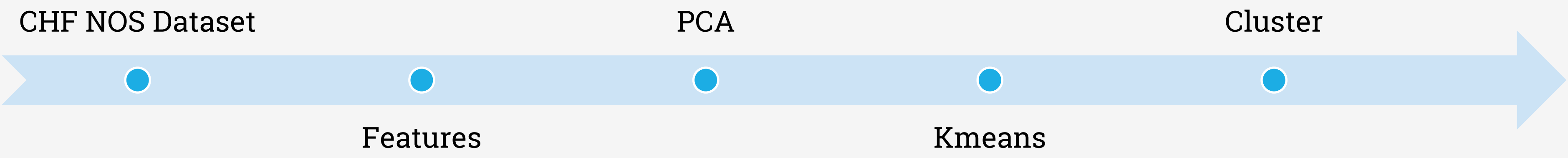
Data Pre-Processing

Feature Engineering

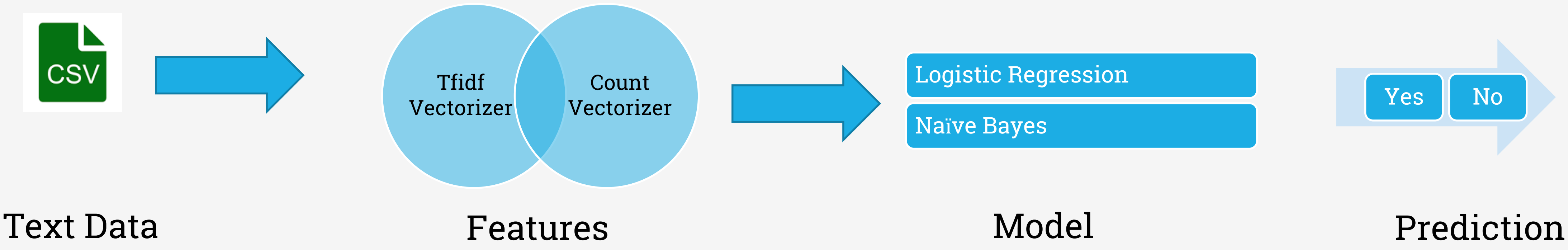
REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

Modeling

Segmentation

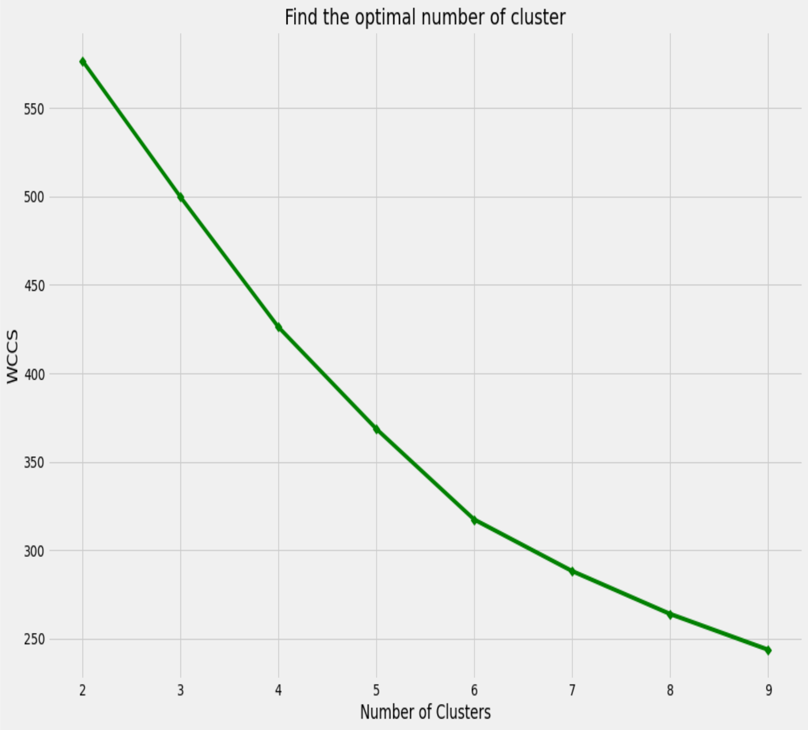


SAE Classification



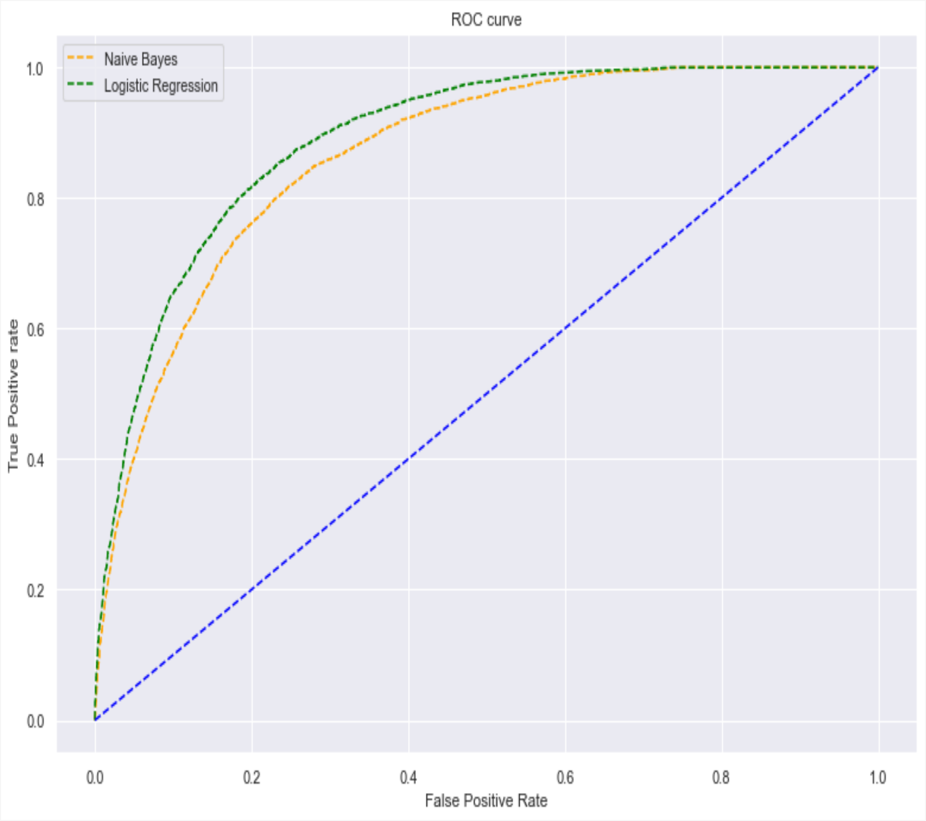
Model Evaluation

Segmentation



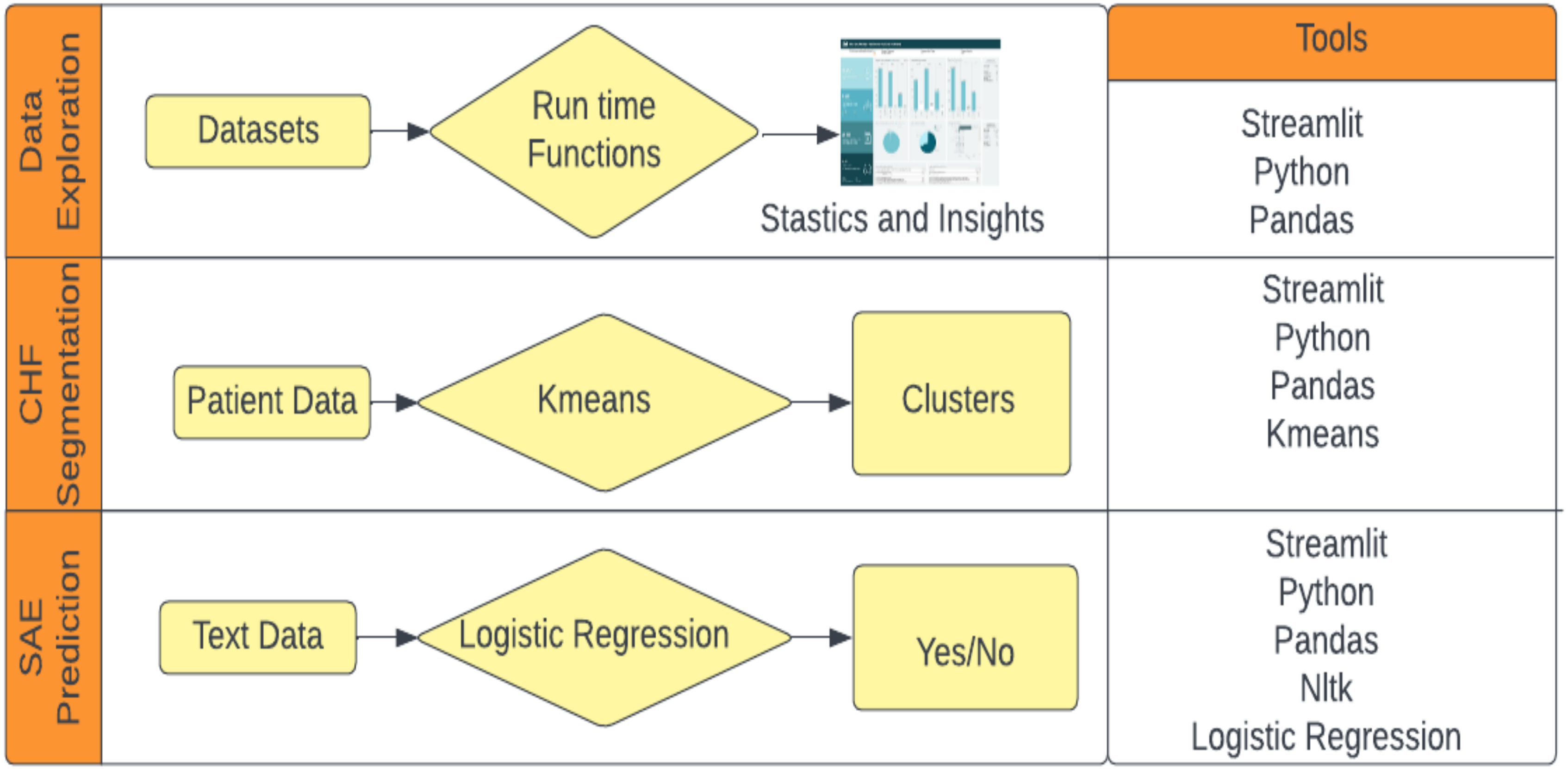
Six Clusters were optimal

SAE Classification



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

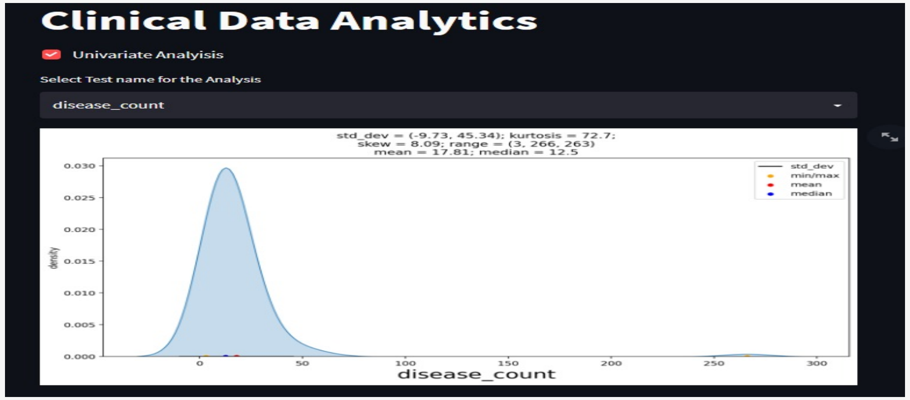
Deployment

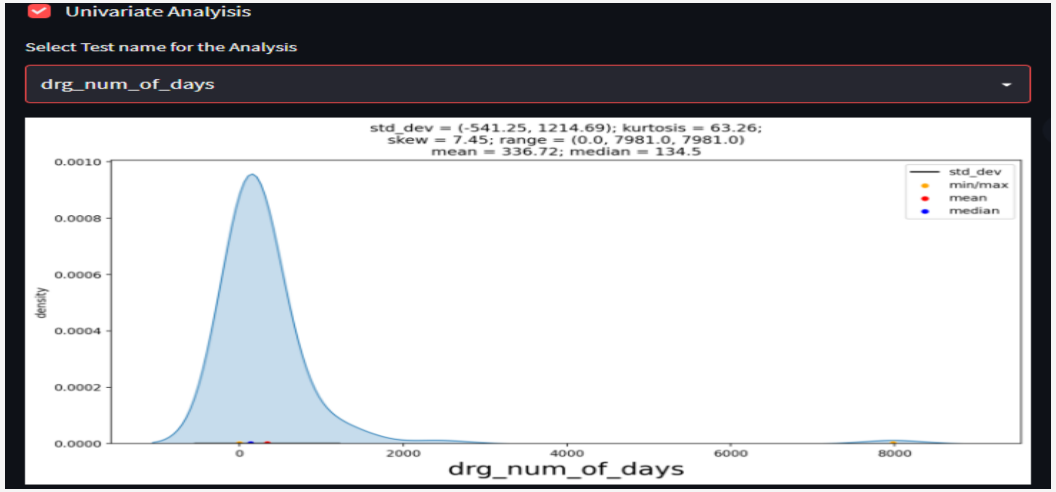

Data Exploration
Datasets
Run time Functions
Stastics and Insights
Tools
Streamlit
Python
Pandas
CHF Segmentation
Patient Data
Kmeans
Clusters
Streamlit
Python
Pandas
Kmeans
SAE Prediction
Text Data
Logistic Regression
Yes/No
Streamlit
Python
Pandas
Nltk
Logistic Regression

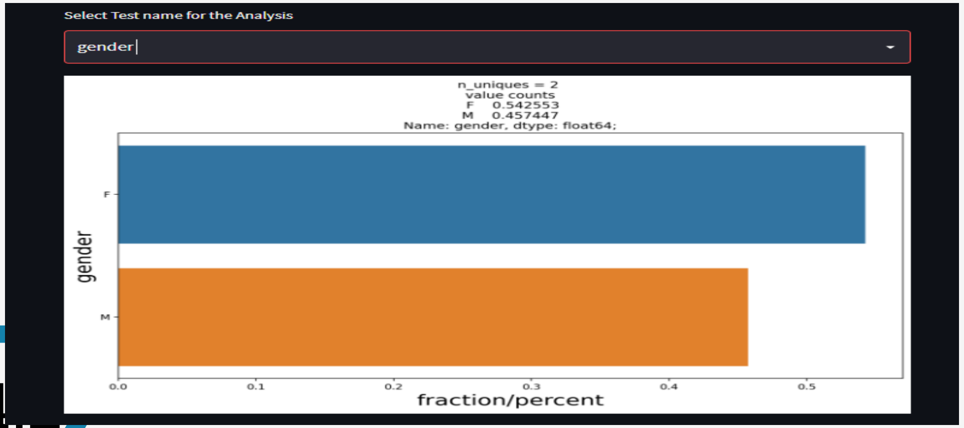
REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

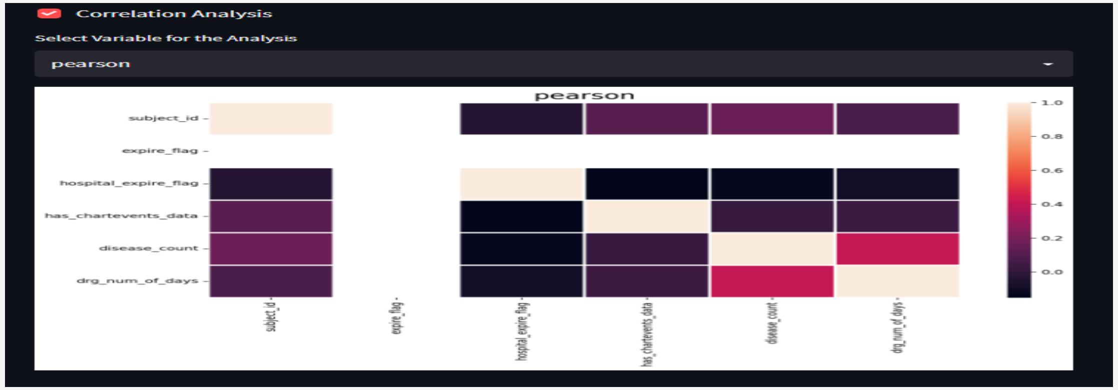
Results and Insights

EDA Application



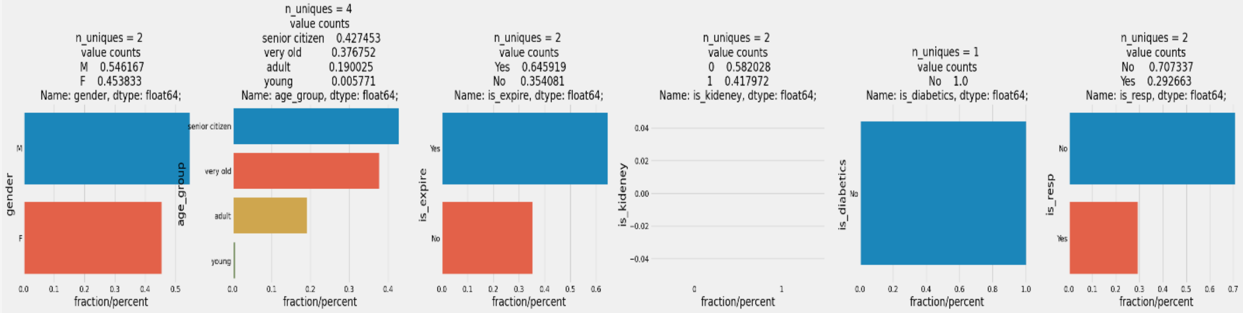


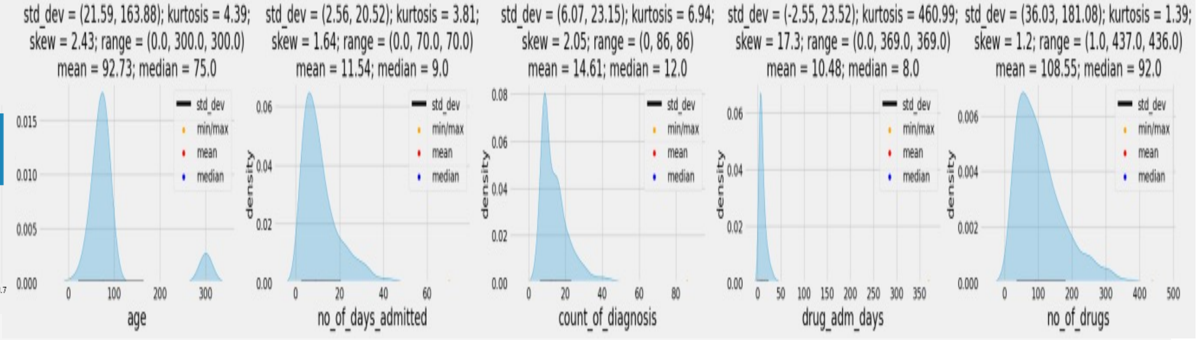




Results and Insights

Segmentation Cluster1

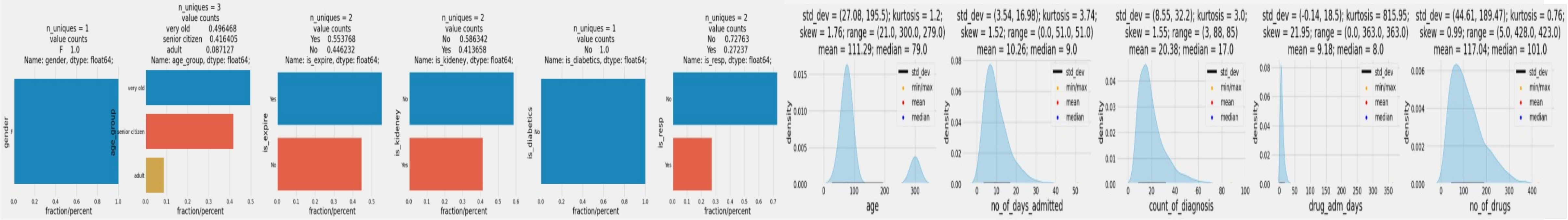




- Males - 54 % and Females 45 % and different age groups are present and 65 % of People are died and no people had diabetic and 70 % of people has respiratory disease
- No of days admitted in the hospital less, no of days mean is 14 and dug administrated days mean is 10 but no of drugs given to them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

Results and Insights

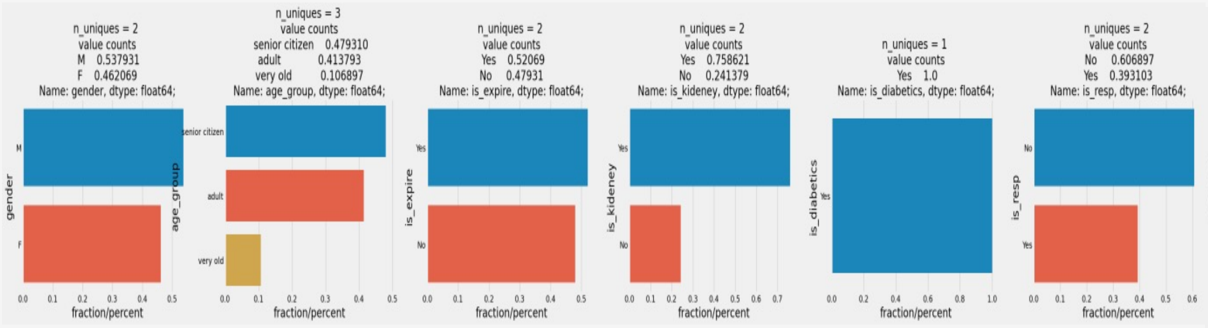
Segmentation Cluster2

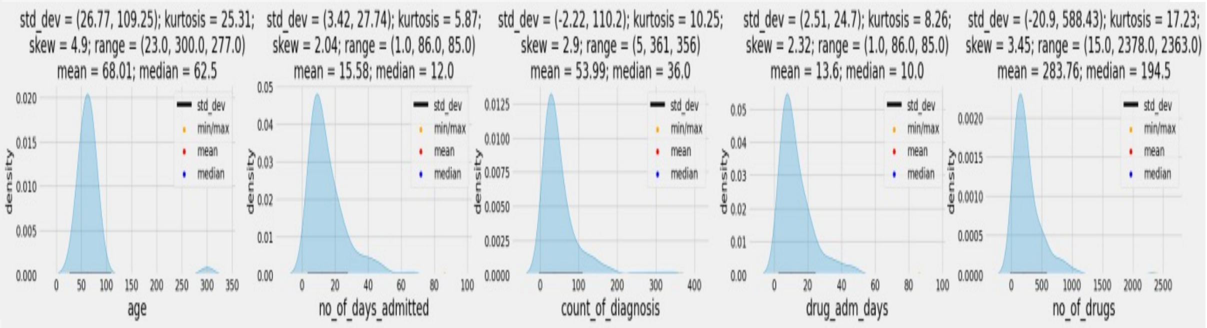


- Only Females very old age people present in this cluster and patient are expired and not expired with almost same distribution
- kidney issues presence is lower
- No people has diabetic and very few people had respiratory issues
- People are not admitted to hospital often and drug administrated days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

Results and Insights

Segmentation Cluster3



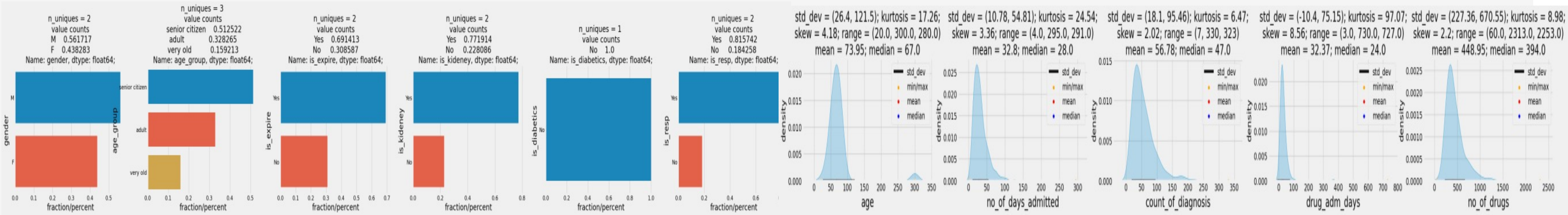


- Both Females and males are equally distributed and most of the patients are adults and senior citizen and very few very old age people and people are died in this cluster is 50 % lesser than not died people
- Most of the people has kidney issues and all the people has diabetic issue and most people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

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Results and Insights

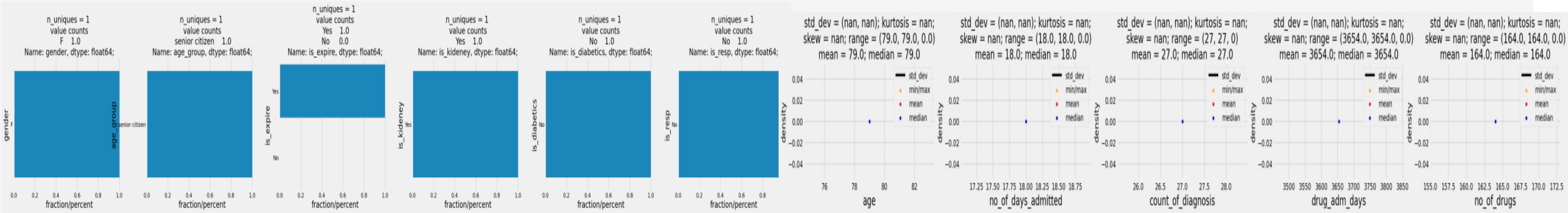
Segmentation Cluster4



- 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

Results and Insights

Segmentation Cluster5

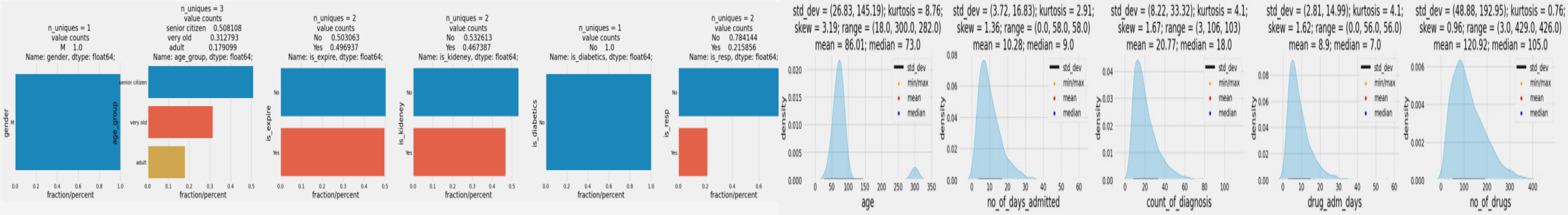


- Only Females present in this cluster and all are senior citizen and all are expired and every one suffered from kidney issue
- No of days admitted is less and count of diagnosis is more and drug administrated days are very high

REVA UNIVERSITY
Bengaluru, India
Established as per the section 2(f) of the UGC Act, 1956, Approved by AICTE, New Delhi

Results and Insights

Segmentation Cluster6



All are males in this cluster and every one 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

Results and Insights

Classification Results

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

The Logistic Regression model produces higher AUC score 89 % and the model integrated with web application for predicting serious adverse event

Conclusion

- This work is intended to provide a business solution to the health care industry and to ensure the safety of the patients the proposed solution is to help the clinician and medical monitors to bring the EHR data to the app and gains insights and statistics
- 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 during the conduction of clinical trial

References

1) Adler Perotte, R. R. (2015). *Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis.*
2) Ahmed Alsayat, H. E.-S. (2016). *Efficient genetic K-Means clustering for health care knowledge discovery.*
3) 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.*
4) Aurelie Mascio, Z. K. (2020). *Comparative Analysis of Text Classification Approaches in Electronic Health Records.*
5) Berger, M. L. (2016). Opportunities and challenges in leveraging electronic health record data in oncology. .
6) Bittar, A. V. (2020). Text Classification to Inform Suicide Risk Assessment in Electronic Health Records.
7) BO JIN, C. C. (2018). *Predicting the Risk of Heart Failure With EHR Sequential Data Modeling.*
8) Churpek MM, Y. T. (2014). Using electronic health record data to develop and validate a prediction model for adverse outcomes in the wards.
9) Eichler, H. G.-D. (2019). Data rich, information poor: can we use electronic health records to create a learning .
10) Estiri, H. K. (2019). A clustering approach for detecting implausible observation values in electronic health records data. .
11) Gabriele Spini, M. v. (2019). *Private Hospital Workflow Optimization via Secure k -Means Clustering.*
12) Hubbard, R. A. (2021). Studying pediatric health outcomes with electronic health records using Bayesian clustering and trajectory analysis. .
13) Irene Li, J. P. (2021). *Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review.*
14) Jose Roberto Ayala Solaresa, b. F. (2019). *Deep learning for electronic health records: A comparative review of multiple deep neural architectures.*
15) LÜTZ, E. (2019). *Unsupervised machine learning to detect patient subgroups in electronic health records.*
16) Mantas, J. (2020). Unsupervised machine learning for the discovery of latent clusters in COVID-19 patients using electronic health records.
17) Mascio, A. K. (2020). Comparative analysis of text classification approaches in electronic health records.
18) Pai, M. M. (2021). Standard electronic health record (EHR) framework for Indian healthcare system. Health Services and Outcomes Research Methodology.
19) Wang, Y. Z. (2020). Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records.
20) Yadav, P. S. (2018). Mining electronic health records (EHRs) A survey. ACM Computing Surveys.
21) Ziyi Liu, J. Z. (2016). Machine Learning for Multimodal Electronic Health Records-based.

REVA
UNIVERSITY
Bengaluru, India

Established as per the section 2(f) of the UGC Act, 1956,
Approved by AICTE, New Delhi

Thank you!