# Hospital Readmission Classification of Diabetes Patients using Data Mining Techniques

**ABSTRACT:** Hospital readmissions increase the healthcare costs and negatively influence hospitals' reputation. Diabetes disease is long-lasting and a chronic disease that occurs when there is an increase in the level of blood sugar (glucose) in a person's body. The symptoms of diabetes are easy to miss due to which it is also known as a silent killer. The cost of hospital readmission of diabetic patients requires a major portion of the hospital's medical expenditure. In this project, I have used data mining techniques namely association and cluster analysis to investigate and extract the useful information from diabetes patient's hospital readmission dataset. The objectives of the project are to perform association rule mining in order to investigate the association between various indicators /attributes that leads to the readmission of diabetic patients. Furthermore, the proposed approach will be used to find clusters that could be derived using various factors effecting hospital admission cost by using other indicators like admission_type, time_in_hospital, num_of_procedures, num_of_medications, etc. the developed system can help hospital management to find the indicators that affect the health of diabetic patients and to manage the hospitalization of diabetic patients by monitoring the factors effecting the patient's readmission. Additionally, it can be helpful for hospitals management to perform decision making in various aspects of hospital managements and patients care and safety. The results of the proposed approach can be successfully used by hospital management and professions in predicting the factors affecting the patients' readmissions in early stages. The results obtained can helps healthcare professionals, and patient's guardians to allows prompting great attention to patients with high risk of readmission, which leverages the healthcare system and saves healthcare expenditures.

**KEYWORDS:** *Diabetes; Data Mining; Association Rule Mining; Apriori; Cluster Analysis, K-Means*

## 1 INTRODUCTION

Diabetes disease is long-lasting and a chronic disease that occurs when there is an increase in the level of blood sugar (glucose) in a person's body. The symptoms of diabetes are easy to miss due to which it is also known as a silent killer. Diabetes disease can cause various other diseases in a person such as heart disease, blood pressure, kidney failure, blindness, etc. [1]. Diabetes is the fourth leading cause of death in the world [2]. Furthermore, it is one of the most expensive chronic diseases in the United States (US) [3]. The cost of hospital readmission of diabetic patients requires a major portion of the hospital's medical expenditure, and it also negatively effects the reputation of hospital [8]. The risk of readmission of hospitalized patients is much higher than those with negative diabetes. However, the reduction in the hospital readmission rates of diabetic patients potentially reduce the medical and healthcare costs [5].

There is a need to reduce the diabetic patients' hospital readmission rate in order to reduce hospital inpatient costs. The manual investigation and analysis is very time consuming, prune to errors, and biased most of the times. Therefor, there is a need to perform analysis of patient's data by some automatic and precise means. However, the solution to this problem is the analysis of data using data mining (DM) and machine learning (ML) techniques. These DM and ML techniques require large amount of data to perform automatic decision making [6-7]. Consequently, there is a need to develop an automated system or model that can predict the readmission of diabetic patients in a precise and accurate way with less time consumption, human resources, and economically.

On the other hand, these days hospitals are generating large amount of electronic healthcare records (EHR) on daily basis. But these EHRs are just the bundle of enormous amount of information which is seldomly used to extract knowledge from these EHRs. Thus, with the help of DM and ML algorithms and their

applications it is possible to automatically extract useful, novel, and hidden patterns from large amount of data sets which will ultimately be useful to convert the information of EHRs into knowledge [9], [11]. The extracted set of knowledge from EHRs can be further used to help decision makers of hospital management and professionals for taking necessary actions in making decisions for the readmission of patients in hospitals.
In this project, I have used data mining techniques namely association and cluster analysis to investigate and extract the useful information from diabetes patient's hospital readmission dataset. The objectives of the project are to perform association rule mining in order to investigate the association between various indicators /attributes that leads to the readmission of diabetic patients. Furthermore, the proposed approach will be used to find clusters that could be derived using various factors effecting hospital admission cost by using other indicators like admission_type, time_in_hospital, num_of_procedures, num_of_medications, etc. the developed system can help hospital management to find the indicators that affect the health of diabetic patients and to manage the hospitalization of diabetic patients by monitoring the factors effecting the patient's readmission. Additionally, it can be helpful for hospitals management to perform decision making in various aspects of hospital managements and patients care and safety.

### 1.1 MOTIVATION

The readmission risk of diabetes patients is much higher than others. However, the cost of hospitalization of diabetic patients accounts for the large amount of hospital and medical expenditures. Therefore, the early prediction of the hospital readmission likelihood of a diabetic patient will reduce the readmission rate of diabetic patients. Ultimately, the medical expenses of hospitals on the hospitalization of diabetic patients will be reduced. Furthermore, this will be effectively used to find future directions in patient monitoring and management which might lead to improvements in patient health and safety.

The remainder of the paper is divided into seven sections. Section 2 presents the survey of the related literature. Section 3 describes the problem statement and expected outcomes of the research work. Section 4 illustrates the description of the dataset used in this research work. Section 5 presents the proposed experimental setup and Section 6 provide the description of the results of the proposed research methodology. Finally, the Section 7 concludes the findings of the research work.

## 2 RELATED LITERATURE

In this section, I have presented a detailed literature survey on the related work. Several studies have used the Diabetes 130-US hospitals for years 1999-2008 Data Set to classify individuals into two groups of hospital readmission namely Yes and No. I have briefly provided the description of the algorithms and methods used and results of each relevant study. [10] have proposed a prediction model that classified the hospital readmission individuals into two groups (Yes or No). They have used the clinical dataset of patient's characteristics over the period of two years having 1211 (12.9%) records of patient's readmission out of total 9381 samples. They have used associative rule mining and five different data mining (DM) classifiers. Association rule mining were used to find the most dominant factors effecting the readmission of diabetic patients' readmission. The results of the proposed methodology indicated that random forest classifier was the most optimal DM classifier used in this study for predicting readmission of diabetic patients with precision-recall curve value of 0.296.

[11] have used machine learning classifiers to predict the hospital readmission of diabetic patients after being discharged from hospital. The result of the study indicated that the most optimal algorithm is random

forest which can effectively classify the patient’s readmission as Yes or No with an accuracy of 89.8%. [8] have proposed a balanced approach between data engineering and neural networks for the prediction of diabetic patient’s hospital readmission. The results of the study indicated that the combination of data engineering and convolutional neural networks (CNN) outperform other ML techniques when applied to real world problems.

[12] have developed a tool to investigate a predict the risk of readmission of diabetic patients within 30 days of their discharge from hospital. The dataset used in this research work is collected from an urban academic medical center consisting of a unit of 44203 discharges. The dataset is randomly divided into training and test dataset with a ratio of 60:40 of the entire dataset. The training dataset comprises of 26402 discharges and test dataset contains 17801 discharges. They have used multivariate logistic regression (MLR) with generalized estimation equation to develop the Diabetes Early Readmission Risk Indicator (DERRI). The results of this research proposed ten most dominant factors effecting the readmission of diabetic patients namely employment status; living within 5 miles of the hospital; preadmission insulin use; burden of macrovascular diabetes complications; admission serum hematocrit, creatinine, and sodium; having a hospital discharge within 90 days before admission; most recent discharge status up to 1 year before admission; and a diagnosis of anemia. The DERRI tool is found to be an effective approach to predict the 30-day readmission of diabetic patients with an accuracy of 70%.

[13] have proposed a model for classifying the EHRs of diabetic patients into hospital readmission of Yes or No. the dataset used in this research is obtained from a reputed hospital in the National Capital Region in India. This dataset comprises of patients of age greater than 18 years who are discharged from hospital from 2012 to 2015. The dataset contains 9381 samples. Tin this research work three data mining and machine learning techniques are used namely logistic regression, naïve bayes, and decision trees. The performance of the classifiers is evaluated with different versions of dataset such after feature selection, imputation of missing values, data balancing. The results of the study indicated that the value of area under curve (AUC) increases from 56%-68% to 83%-86% by using the preprocessed dataset. The results of the study revealed that the preprocessing of the dataset has a very significant effect on the accuracy of patient’s hospital readmission prediction model.

This section presents review of the several research articles on the related topic which illustrates that several studies have been done to predict the readmission of patients. Different data mining and machine learning techniques have been used to address the problem of readmission. As the survey in the related literature indicated that all the studies focused on the prediction and classification of hospital readmission. A lot of work has been done to solve this problem. However, there is a need to investigate the factors effecting the patient’s hospital readmission. Therefore, the focus of this research work is to investigate on this research gap.

# 3 PROBLEM STATEMENT

Diabetes is one of the most chronic and expensive diseases in the US. The cost of hospital readmission of diabetic patients requires a major portion of the hospital’s medical spending. There is a need to investigate and find out the factors effecting hospital admission.

## 3.1 EXPECTED OUTCOMES

The expected outcomes of the proposed project are:

- To perform association rule mining in order to investigate the association between various indicators /attributes that leads to the readmission of diabetic patients.
    - Which will help us find the indicators that affect the health of diabetic patients.
    - Which will be effectively used by hospital professionals to manage the hospitalization of diabetic patients by monitoring the factors effecting the patient's readmission.
- To find clusters that could be derived using various factors effecting hospital admission cost by using other indicators like admission_type, time_in_hospital, num_of_procedures, num_of_medications, etc.
    - Which will be a novel point of this research work that can be helpful for hospitals management to perform decision making in various acpects of hospital managements and patients care.

## 4 DATASET DESCRIPTION

The "Diabetes 130-US hospitals for years 1999-2008 Data Set" is used in this research project. This dataset is acquired from the UCI machine learning repository [5]. This dataset is provided by the Center for Clinical and Translational Research at Virginia Commonwealth University in the United States. This dataset presents the clinical care records of 130 US hospitals over the time period of 10 years from 1999 to 2008. The dataset contains 55 attributes indicating the patient and hospital outcomes such as patient number, gender, age, race, admission type, time in hospital, the medical specialty of admitting physician, HbA1c test result, diagnosis, number of medications, diabetic medications, and emergency visits in the year before the hospitalization, etc. The dataset comprises 101,766 clinical records. This dataset has been collected and prepared with the objective to analyze cases of readmissions and other consequences related to diabetic patients. Table 1 shows the overall summary of the Diabetes 130-US hospitals for the years 1999-2008 dataset.

Table 1: Summary of the Diabetes 130-US hospitals for years 1999-2008 dataset

| | | | |
|---|---|---|---|
| Number of Instances | 101,766 | Number of Attributes | 55 |
| Data Set Characteristics | Multivariate | Attribute Characteristics | Integer |
| Area | Life | Missing Values? | Yes |
| Associated Tasks | Clustering, Association analysis | Date Donated | 2014-05-03 |

The detailed description of the attributes is given in Table 2.

**Table 2:** Description of the attributes in Diabetes 130-US hospitals for years 1999-2008 dataset

| Sr. # | Attribute | Description | Data Type |
|---|---|---|---|
| 1 | encounter_id | Encounter ID Unique identifier of an encounter | int64 |
| 2 | patient_nbr | Patient number Unique identifier of a patient | int64 |
| 3 | race | Race Values: Caucasian, Asian, African American, Hispanic, and other | object |
| 4 | gender | Gender Values: male, female, and unknown/invalid | object |
| 5 | age | Age Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100) | object |
| 6 | weight | Weight Weight in pounds | object |

|  |  |  |  |
|---|---|---|---|
| 7 | admission_type_id | Admission type Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available | int64 |
| 8 | discharge_disposition_id | Discharge disposition Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available | int64 |
| 9 | admission_source_id | Admission source Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital | int64 |
| 10 | time_in_hospital | Time in hospital Integer number of days between admission and discharge | int64 |
| 11 | payer_code | Payer code Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay Medical | object |
| 12 | medical_specialty | Medical specialty Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon | object |
| 13 | num_lab_procedures | Number of lab procedures Number of lab tests performed during the encounter | int64 |
| 14 | num_procedures | Number of procedures Numeric Number of procedures (other than lab tests) performed during the encounter | int64 |
| 15 | num_medications | Number of medications Number of distinct generic names administered during the encounter | int64 |
| 16 | number_outpatient | Number of outpatient visits Number of outpatient visits of the patient in the year preceding the encounter | int64 |
| 17 | number_emergency | Number of emergency visits Number of emergency visits of the patient in the year preceding the encounter | int64 |
| 18 | number_inpatient | Number of inpatient visits Number of inpatient visits of the patient in the year preceding the encounter | int64 |
| 19 | diag_1 | Diagnosis 1 The primary diagnosis (coded as first three digits of ICD9); 848 distinct values | object |
| 20 | diag_2 | Diagnosis 2 Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values | object |
| 21 | diag_3 | Diagnosis 3 Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values | object |
| 22 | number_diagnoses | Number of diagnoses Number of diagnoses entered to the system 0% | int64 |
| 23 | max_glu_serum | Glucose serum test result Indicates the range of the result or if the test was not taken. Values: ">200," ">300," "normal," and "none" if not measured | object |
| 24 | A1Cresult | A1c test result Indicates the range of the result or if the test was not taken. Values: ">8" if the result was greater than 8%, ">7" if the result was greater than 7% but less than 8%, "normal" if the result was less than 7%, and "none" if not measured. | object |
| 25 | metformin | 24 features for medications. For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride- pioglitazone, metformin-rosiglitazone, and metformin- pioglitazone.<br>The feature indicates whether the drug was prescribed or there was a change in the dosage. Values: "up" if the dosage was increased during the encounter, "down" if the dosage was decreased, "steady" if the dosage did not change, and "no" if the drug was not prescribed | object |
| 26 | repaglinide |  | object |
| 27 | nateglinide |  | object |
| 28 | chlorpropamide |  | object |
| 29 | glimepiride |  | object |
| 30 | acetohexamide |  | object |
| 31 | glipizide |  | object |
| 32 | glyburide |  | object |
| 33 | tolbutamide |  | object |
| 34 | pioglitazone |  | object |
| 35 | rosiglitazone |  | object |
| 36 | acarbose |  | object |
| 37 | miglitol |  | object |
| 38 | troglitazone |  | object |
| 39 | tolazamide |  | object |
| 40 | examide |  | object |
| 41 | citoglipton |  | object |
| 42 | insulin |  | object |
| 43 | glyburide-metformin |  | object |
| 44 | glipizide-metformin |  | object |
| 45 | glimepiride-pioglitazone |  | object |
| 46 | metformin-rosiglitazone |  | object |
| 47 | metformin-pioglitazone |  | object |

| | | | |
|---|---|---|---|
| 48 | change | Change of medications Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change” | object |
| 49 | diabetesMed | Diabetes medications Indicates if there was any diabetic medication prescribed. Values: “yes” and “no” | object |
| 50 | readmitted | Readmitted Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission | object |

### 4.1 Data Cleaning

This section presents the steps used to clean the dataset. Firstly, there are many missing values for several attributes in the dataset. Table 3 illustrates the attributes with their number of missing values.

Table 3: Attributes with their number of missing values

| Attribute | Missing Values |
|---|---|
| race | 2273 |
| weight | 98569 |
| payer_code | 40256 |
| medical_specialty | 49949 |
| diag_1 | 21 |
| diag_2 | 358 |
| diag_3 | 1423 |

This indicates that there are very large of missing values for weight, payer_code, and medical_specialty attributes. Therefore, I have simply omitted these attributes from the dataset.

After that, I have dropped the unnecessary / Uninformative attributes from the dataset. The uninformative features in the dataset (21 in total) are discarded, due to either, a huge amount of missing sample values (>50%), or due to the fact that some features are not relevant to the further analysis of the data (Like encounter_id), or if the feature is compeletly unbalanced (>95% of data points have the same value for the feature). After data cleaning, we have 29 coulmns in the cleaned dataset.

### 4.2 EXPLORATORY DATA ANALYSIS (EDA)

This section presents a detailed exploratory data analysis (EDA) done on the diabetes dataset used in this research work. Figure 1 (a) illustrates the distribution of readmission patients which means days to inpatient readmission. This means:

- If the patient was readmitted in less than 30 days "<30"
- if the patient was readmitted in more than 30 days ">30"
- If there is no record "NO"

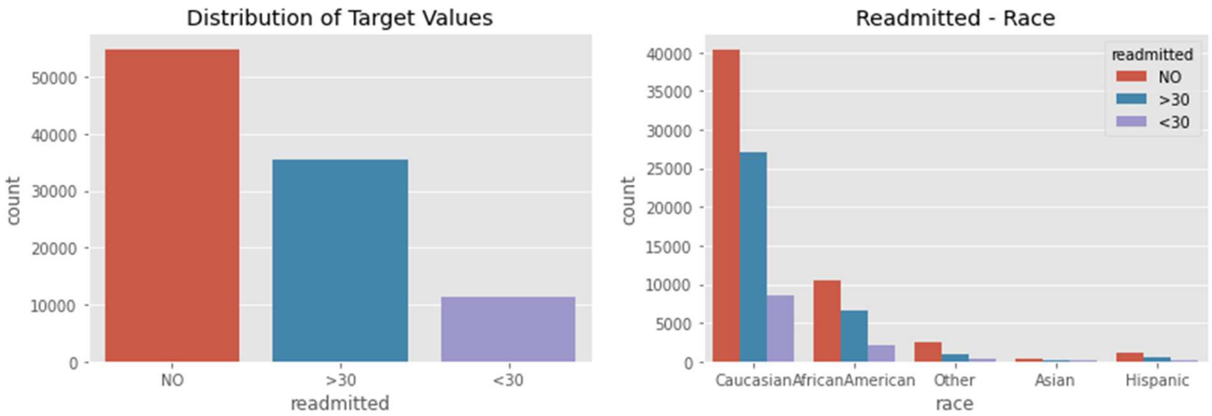


Figure 1: (a) distribution of readmitted attribute, (b) proportion of readmitted patients with respect to Race

Figure 1 (b) illustrated the proportion of readmitted patients with respect to race of the patients. There is Caucasians in 74 percent of all our data. And other 26 percent is divided into African Americans, Hispanics, Asians and Others. Most of the patients are Caucasian, followed by African Americans. Although the Hispanic values are few than Caucasian, we see that the Readmitted Probability almost close to Caucasian. Similarly, Figure 2 (a) and (b) illustrates the distribution of readmission patients with respect to gender and age, respectively.

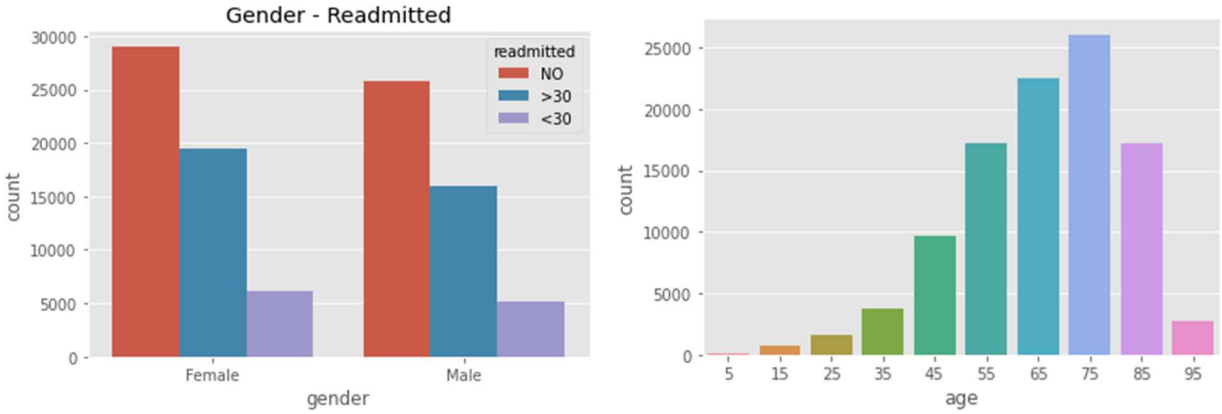


Figure 2: (a) proportion of readmitted patients with respect to Gender (b) proportion of readmitted patients with respect to Age

We see a nearly equal distribution of Gender. Also, we can state that Females are a little more prone than Males. Similarly, if we look at age distribution, we can conclude that elderly population is more suspectable to be readmitted.

Figure 3 illustrates the distribution of readmission patients with respect to time spent in hospital by the patients at the their first admission. It is integer number of days between admission and discharge. Shortly it is "treatment time". This illustrates that most people stayed 2 - 3 days in hospital.

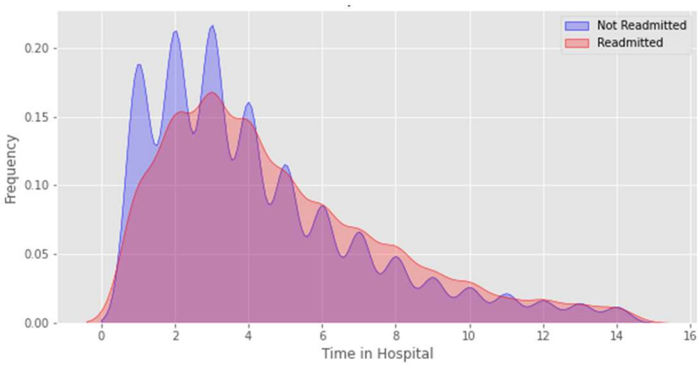


Figure 3: Proportion of readmitted patients with respect to Time spend in hospital

### 4.3 CORRELATION ANALYSIS

The correlation between the attributes of diabetes dataset is given in Figure 4. Figure 4 indicates that the strongest correlations among the predictors are:

- num_medications & time_in_hospital (corr = 0.5)
- change & insulin (corr = 0.5)
- change & diabetesMed (corr = 0.5)
- diabetesMed & insulin (corr = 0.5)

Also, the readmitted variable has in general weak correlation with all predictors.

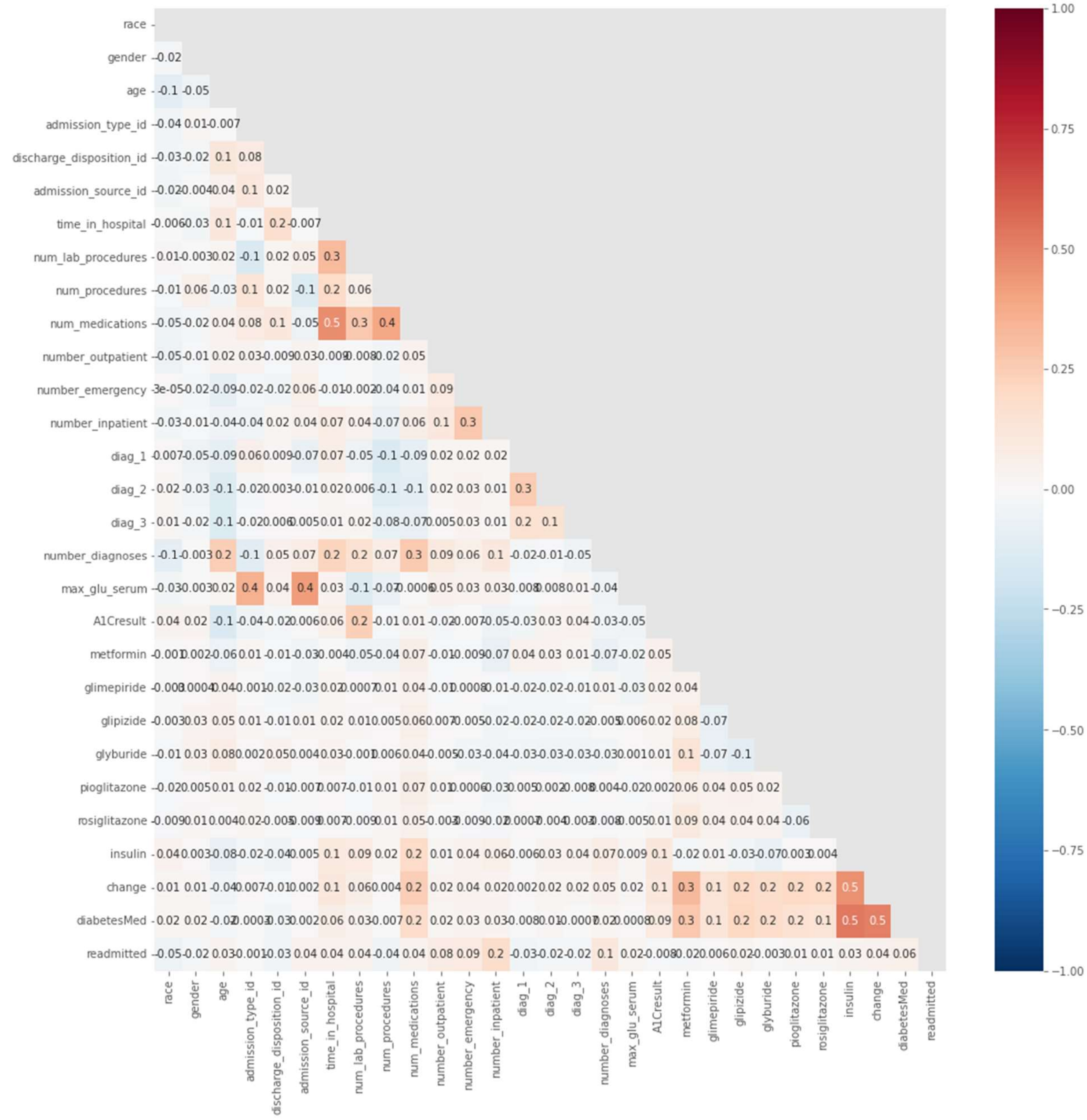


**Figure 4:** Correlation analysis between attributes of the dataset

## 5 ALGORITHMS AND PROPOSED SETUP

The proposed experimental setup is divided into three phases. Firstly, data preprocessing and data cleaning is done in order to improve the quality of the dataset, Secondly, the exploratory data analysis (EDA) will be performed to get an insight into the dataset. In the third phase, I have used association rule mining to find the significant factors effecting the hospital readmission. Apriori algorithms will yield association rules indicating the parameters that leads to the readmission of patients to the hospitals. After that I have used cluster analysis technique to find clusters that could be derived using various factors effecting hospital admission cost by using other indicators like admission_type, time_in_hospital, num_of_procedures, num_of_medications, etc. Finally, in the last phase of this research work, I have performed analysis of the results obtained using association rule mining, and cluster analysis. Figure 5 illustrates the graphic representation of the proposed experimental setup.

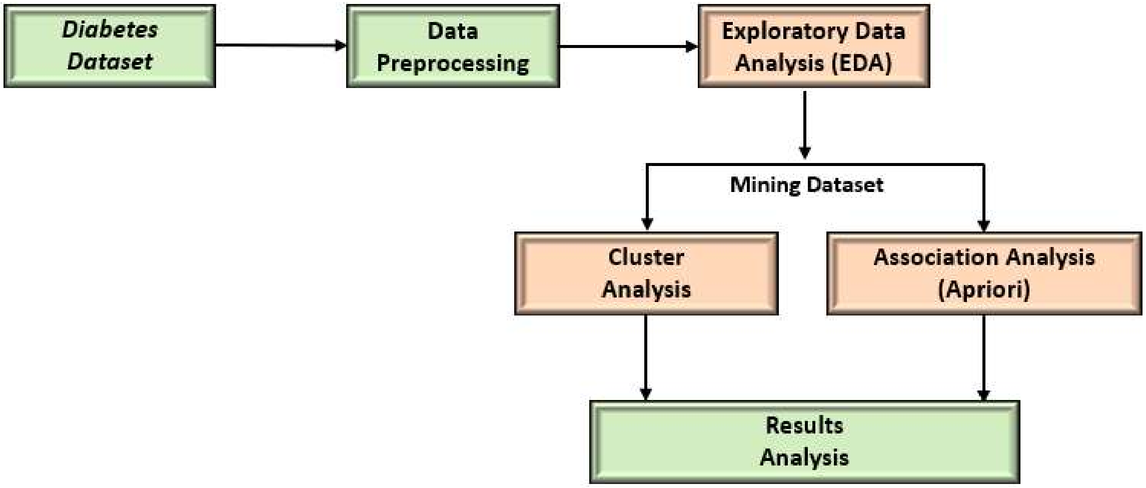


**Fig 5:** Proposed experimental setup

1) **K-Means clustering**: K-Means clustering algorithm will be used on the dataset to identify diabetic patients belonging to two major categories: Readmitted and Not Readmitted. K-means clustering is an unsupervised learning algorithm that divides a dataset into K number of non-overlapping separate clusters or subgroups based on its attributes.
2) **Apriori algorithm**: Apriori will be used for finding the parameters / attributes that effects the hospital readmission of diabetic patients. Apriori algorithms will yield association rules indicating the parameters that leads to the readmission of patients to the hospitals.

### 5.1 ASSOCIATION RULE MINING

Association rule mining is a data mining technique to identify the relations between different set of attributes or mostly known as items in association rule mining [14-15]. There are several algorithms of association rule mining. In this research work, I have used one of the most widely used association rule mining algorithm namely Apriori algorithm in to find the association between different set of factors / attributes that majorly effects the readmission of patients. There are three major components of Apriori algorithm:

**Support**: Support refers fraction of transactions that contain an itemset X. This can be calculated as:

$$\textbf{Support(X) = (Transactions containing (X)) / (Total Transactions)} \tag{1}$$

**Confidence**: confidence refers to the measures how often items in Y appear in transactions that contain X.

$$\textbf{Confidence(A}\rightarrow\textbf{B) = (Transactions containing both (A and B)) / (Transactions containing A)} \tag{2}$$

**Lift**: Lift(X→Y) refers to the increase in the ratio of sale of Y when X is sold. Lift (X → B) can be calculated by dividing Confidence (X →Y) divided by Support(Y). This can be calculated as:

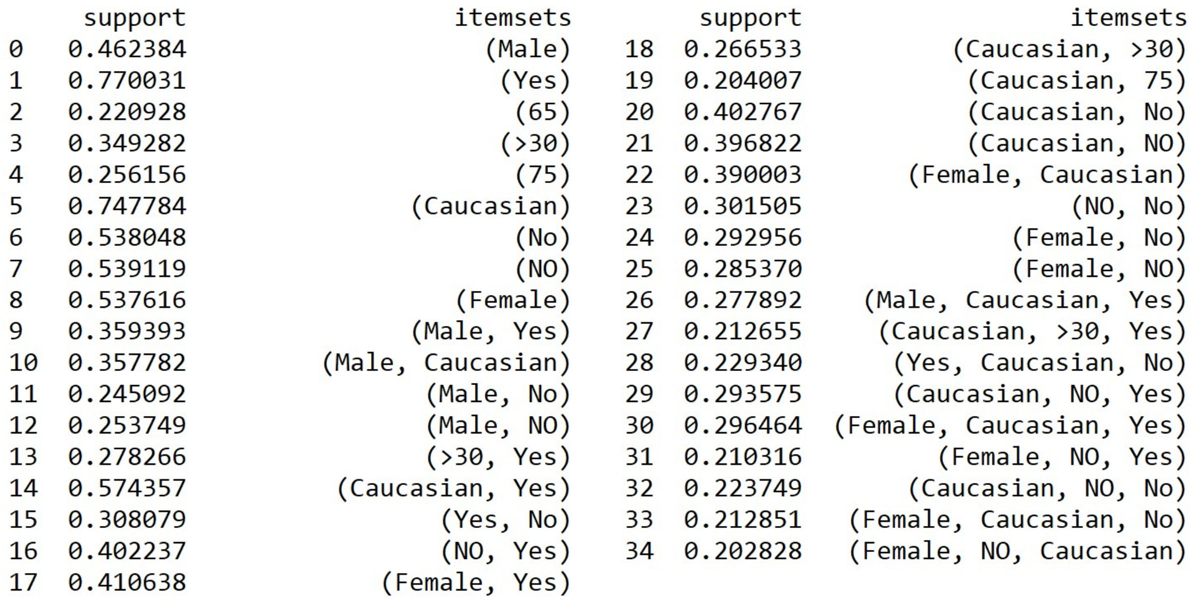
$$\textbf{Lift(X}\rightarrow\textbf{Y) = (Confidence (X}\rightarrow\textbf{Y)) / (Support (Y))} \tag{3}$$

### 5.2 CLUSTER ANALYSIS

Cluster analysis the unsupervised machine learning algorithms. Unsupervised learning algorithms, as the name suggests, make extrapolations from the input attributes of the datasets without using any known, or labelled, outcomes against the input vectors. K-means clustering is one of the simplest and most widely used unsupervised learning algorithm [16-17]. The objective of K-means clustering algorithm is to group similar data points / attributes together and discover underlying patterns. K-means algorithm looks for a fixed number (k) of clusters in a dataset. A cluster refers to a collection of data points aggregated together because of certain similarities. The quality of the clustering algorithm depends on the high inter-cluster dissimilarity and high intra-cluster similarities. The detailed working of the k-means algorithm is given in [18-19].

## 6 RESULTS AND DISCUSSION

Figure 6 illustrates the list of frequent item sets obtained by using Apriori algorithm. This illustrates the frequent-1, frequent-2, frequent-3 itemset with their support value. I have set the minimum support = 0.2.



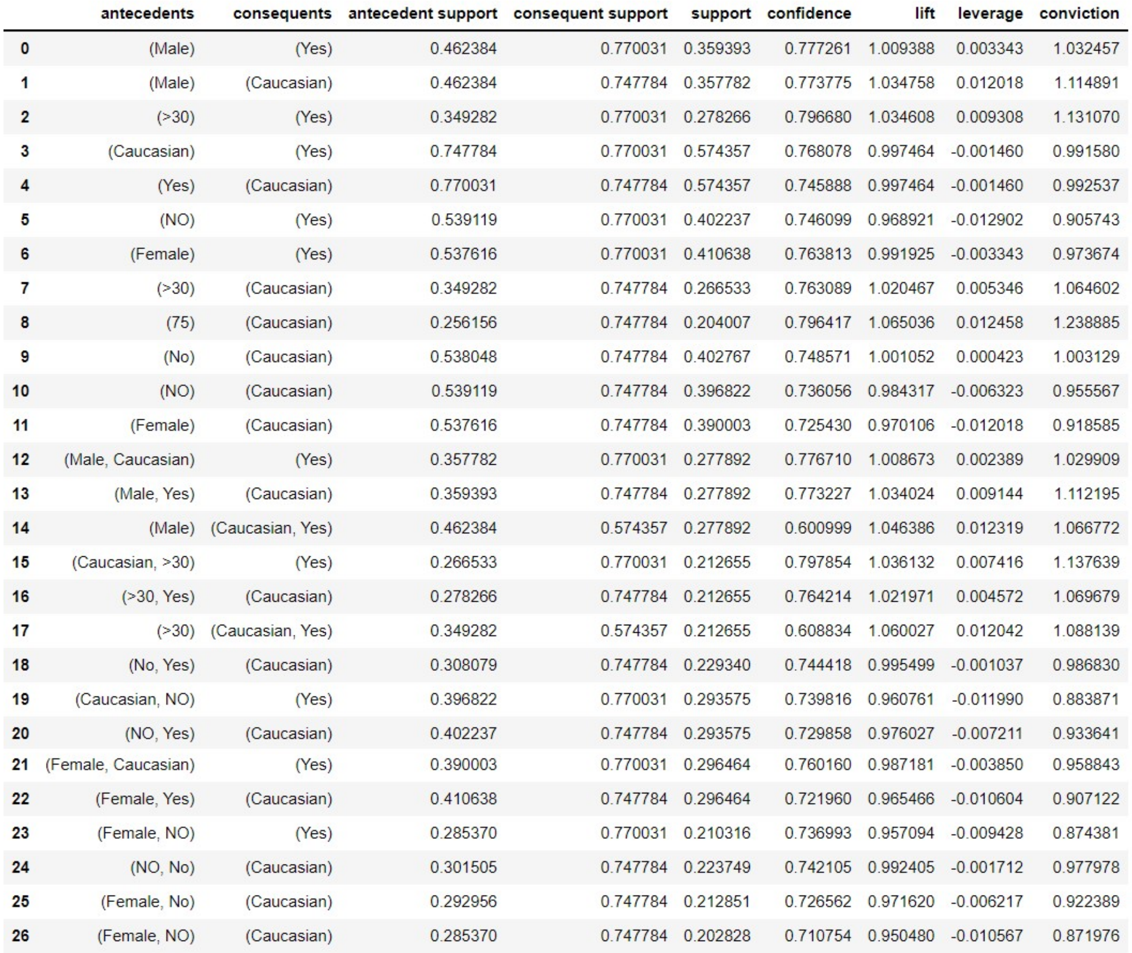
```
     support                    itemsets        support                   itemsets
0    0.462384                     (Male)    18  0.266533           (Caucasian, >30)
1    0.770031                      (Yes)    19  0.204007            (Caucasian, 75)
2    0.220928                       (65)    20  0.402767            (Caucasian, No)
3    0.349282                      (>30)    21  0.396822            (Caucasian, NO)
4    0.256156                       (75)    22  0.390003        (Female, Caucasian)
5    0.747784                (Caucasian)    23  0.301505                   (NO, No)
6    0.538048                       (No)    24  0.292956               (Female, No)
7    0.539119                       (NO)    25  0.285370               (Female, NO)
8    0.537616                   (Female)    26  0.277892     (Male, Caucasian, Yes)
9    0.359393                (Male, Yes)    27  0.212655      (Caucasian, >30, Yes)
10   0.357782          (Male, Caucasian)    28  0.229340       (Yes, Caucasian, No)
11   0.245092                 (Male, No)    29  0.293575       (Caucasian, NO, Yes)
12   0.253749                 (Male, NO)    30  0.296464   (Female, Caucasian, Yes)
13   0.278266                 (>30, Yes)    31  0.210316          (Female, NO, Yes)
14   0.574357           (Caucasian, Yes)    32  0.223749        (Caucasian, NO, No)
15   0.308079                  (Yes, No)    33  0.212851    (Female, Caucasian, No)
16   0.402237                  (NO, Yes)    34  0.202828    (Female, NO, Caucasian)
17   0.410638              (Female, Yes)
```

**Figure 6:** Frequent item sets obtained by Apriori Algorithm with minimum support = 0.2

Figure 7 illustrates the list of association rules obtained by applying Apriori algorithm. There are two parts in association rules name rule antecedents and rule consequents. If we talk about rule 23, as illustrated in figure 7, it is written as:

IF gender = Female AND readmitted = NO THEN diabetesMed = Yes

The support of the rule is 0.210316 and confidence is 0.736562. Similarly, we can evaluate all other association rules. We can find the strong association rules by defining the threshold values for support and confidence of the association rule.



| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction |
|---|---|---|---|---|---|---|---|---|---|
| **0** | (Male) | (Yes) | 0.462384 | 0.770031 | 0.359393 | 0.777261 | 1.009388 | 0.003343 | 1.032457 |
| **1** | (Male) | (Caucasian) | 0.462384 | 0.747784 | 0.357782 | 0.773775 | 1.034758 | 0.012018 | 1.114891 |
| **2** | (>30) | (Yes) | 0.349282 | 0.770031 | 0.278266 | 0.796680 | 1.034608 | 0.009308 | 1.131070 |
| **3** | (Caucasian) | (Yes) | 0.747784 | 0.770031 | 0.574357 | 0.768078 | 0.997464 | -0.001460 | 0.991580 |
| **4** | (Yes) | (Caucasian) | 0.770031 | 0.747784 | 0.574357 | 0.745888 | 0.997464 | -0.001460 | 0.992537 |
| **5** | (NO) | (Yes) | 0.539119 | 0.770031 | 0.402237 | 0.746099 | 0.968921 | -0.012902 | 0.905743 |
| **6** | (Female) | (Yes) | 0.537616 | 0.770031 | 0.410638 | 0.763813 | 0.991925 | -0.003343 | 0.973674 |
| **7** | (>30) | (Caucasian) | 0.349282 | 0.747784 | 0.266533 | 0.763089 | 1.020467 | 0.005346 | 1.064602 |
| **8** | (75) | (Caucasian) | 0.256156 | 0.747784 | 0.204007 | 0.796417 | 1.065036 | 0.012458 | 1.238885 |
| **9** | (No) | (Caucasian) | 0.538048 | 0.747784 | 0.402767 | 0.748571 | 1.001052 | 0.000423 | 1.003129 |
| **10** | (NO) | (Caucasian) | 0.539119 | 0.747784 | 0.396822 | 0.736056 | 0.984317 | -0.006323 | 0.955567 |
| **11** | (Female) | (Caucasian) | 0.537616 | 0.747784 | 0.390003 | 0.725430 | 0.970106 | -0.012018 | 0.918585 |
| **12** | (Male, Caucasian) | (Yes) | 0.357782 | 0.770031 | 0.277892 | 0.776710 | 1.008673 | 0.002389 | 1.029909 |
| **13** | (Male, Yes) | (Caucasian) | 0.359393 | 0.747784 | 0.277892 | 0.773227 | 1.034024 | 0.009144 | 1.112195 |
| **14** | (Male) | (Caucasian, Yes) | 0.462384 | 0.574357 | 0.277892 | 0.600999 | 1.046386 | 0.012319 | 1.066772 |
| **15** | (Caucasian, >30) | (Yes) | 0.266533 | 0.770031 | 0.212655 | 0.797854 | 1.036132 | 0.007416 | 1.137639 |
| **16** | (>30, Yes) | (Caucasian) | 0.278266 | 0.747784 | 0.212655 | 0.764214 | 1.021971 | 0.004572 | 1.069679 |
| **17** | (>30) | (Caucasian, Yes) | 0.349282 | 0.574357 | 0.212655 | 0.608834 | 1.060027 | 0.012042 | 1.088139 |
| **18** | (No, Yes) | (Caucasian) | 0.308079 | 0.747784 | 0.229340 | 0.744418 | 0.995499 | -0.001037 | 0.986830 |
| **19** | (Caucasian, NO) | (Yes) | 0.396822 | 0.770031 | 0.293575 | 0.739816 | 0.960761 | -0.011990 | 0.883871 |
| **20** | (NO, Yes) | (Caucasian) | 0.402237 | 0.747784 | 0.293575 | 0.729858 | 0.976027 | -0.007211 | 0.933641 |
| **21** | (Female, Caucasian) | (Yes) | 0.390003 | 0.770031 | 0.296464 | 0.760160 | 0.987181 | -0.003850 | 0.958843 |
| **22** | (Female, Yes) | (Caucasian) | 0.410638 | 0.747784 | 0.296464 | 0.721960 | 0.965466 | -0.010604 | 0.907122 |
| **23** | (Female, NO) | (Yes) | 0.285370 | 0.770031 | 0.210316 | 0.736993 | 0.957094 | -0.009428 | 0.874381 |
| **24** | (NO, No) | (Caucasian) | 0.301505 | 0.747784 | 0.223749 | 0.742105 | 0.992405 | -0.001712 | 0.977978 |
| **25** | (Female, No) | (Caucasian) | 0.292956 | 0.747784 | 0.212851 | 0.726562 | 0.971620 | -0.006217 | 0.922389 |
| **26** | (Female, NO) | (Caucasian) | 0.285370 | 0.747784 | 0.202828 | 0.710754 | 0.950480 | -0.010567 | 0.871976 |

**Figure 7:** Association rules obtained by Apriori algorithm

Figure 8 illustrates the results of the principal clustering algorithm of unsupervised learning.

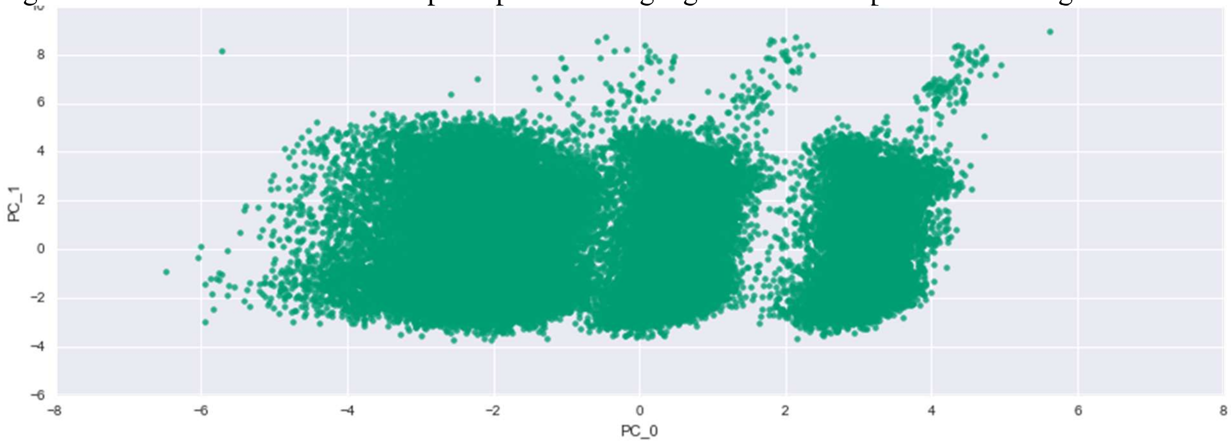


**Figure 8**: projection of cluster on first two principal components

This indicates the projection of dataset onto first two principal components (PC). There are three primary clusters such as one especially large one on the left, and two identically shaped smaller ones on the right. I have used k-means clustering to split the dataset into the clusters identified above. I noted, however, that k-means will tend to extract clusters of similar sizes, which meant that I had to perform the operation twice - once to split the dataset into the large leftward and coupled right clusters, then again to split the coupled clusters in twain. I normalized the features to allow features of different magnitudes to be compared in a similar way.

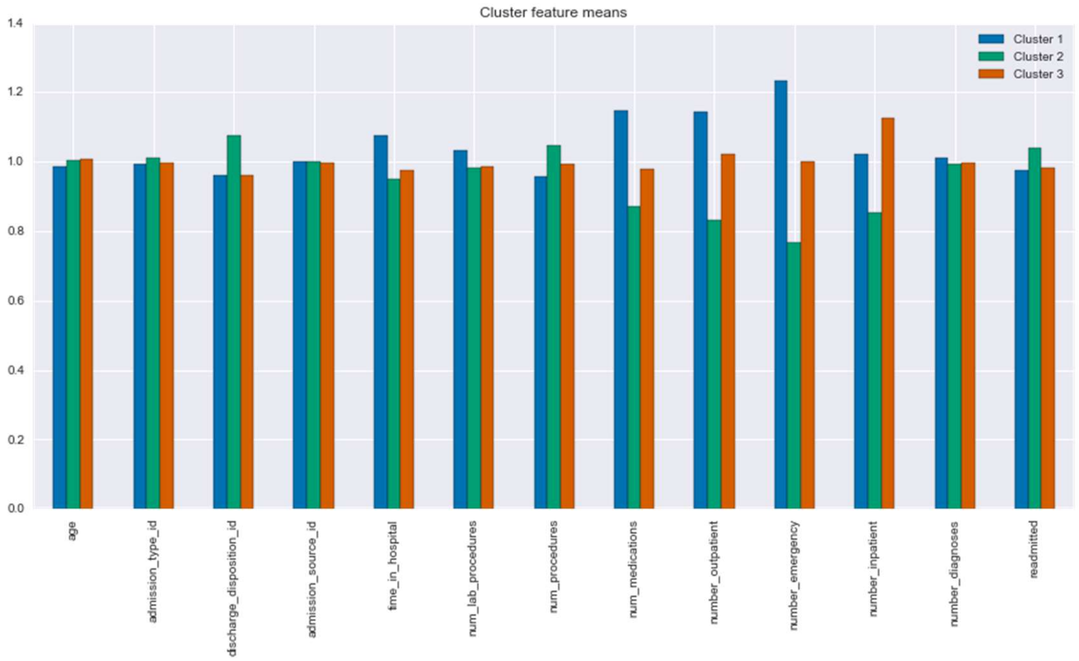


**Figure 9**: Cluster feature means

**Table 4:** Attributes values with respected to clusters

| | Cluster 1 | Cluster 2 | Cluster 3 |
|---|---|---|---|
| age | 0.98635 | 1.005507 | 1.008144 |
| admission_type_id | 0.994391 | 1.010118 | 0.995491 |
| discharge_disposition_id | 0.961746 | 1.075683 | 0.962571 |
| admission_source_id | 1.001827 | 0.999966 | 0.998207 |
| time_in_hospital | 1.074667 | 0.951387 | 0.973945 |
| num_lab_procedures | 1.032297 | 0.983117 | 0.984586 |
| num_procedures | 0.958633 | 1.04644 | 0.994926 |
| num_medications | 1.149239 | 0.870161 | 0.9806 |
| number_outpatient | 1.145165 | 0.833544 | 1.02129 |
| number_emergency | 1.233205 | 0.767615 | 0.99918 |
| number_inpatient | 1.020498 | 0.853859 | 1.125643 |
| number_diagnoses | 1.011049 | 0.992814 | 0.996137 |
| readmitted | 0.975335 | 1.041715 | 0.982951 |

I have drawn following observations from the above given chart in Figure 9 and Table 4:

- Cluster 1 indicates that the patients spent between a third and half a day longer in hospital
- Cluster 1 indicates that the patients had about 5% more lab procedures than those in clusters 2 or clusters 3, and are, on average, using between 15 and 25% more medicaments.
- Cluster 1 indicates that the patients had a record of more encounters (inpatient, emergency, and outpatient).

## 7 CONCLUSION

Hospital readmissions increase the healthcare costs and negatively influence hospitals' reputation. In this project, I have used data mining techniques namely association and cluster analysis to investigate and extract the useful information from diabetes patient's hospital readmission dataset. The objectives of the project are to perform association rule mining in order to investigate the association between various indicators /attributes that leads to the readmission of diabetic patients. Furthermore, the proposed approach will be used to find clusters that could be derived using various factors effecting hospital admission cost by using other indicators like admission_type, time_in_hospital, num_of_procedures, num_of_medications, etc. the results of clustering indicates that patients who spent between a third and half a day longer in hospital have greater possibility to be readmitted. Furthermore, the patients having about 5% more lab procedures than those in clusters 2 or clusters 3, and are, on average, using between 15 and 25% more medicaments have high possibility of hospital readmission and the patients having a record of more encounters (inpatient, emergency, and outpatient) are more likely to be readmitted. The results of the research work can help hospital management to find the indicators that affect the health of diabetic patients and to manage the hospitalization of diabetic patients by monitoring the factors effecting the patient's readmission. The results of the proposed approach can be successfully used by hospital management and professions in predicting the factors affecting the patients' readmissions in early stages. The results obtained can helps healthcare professionals, and patient's guardians to allows prompting great attention to patients with high risk of readmission, which leverages the healthcare system and saves healthcare expenditures.

## BIBLIOGRAPHY

[1] Mostafa Fathi Ganji, Mohammad Saniee Abadeh "Using fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease" Proceedings of ICEE 2010, May 11-13, IEEE 2010.

[2] Gan, Delice. Diabetes atlas. International Diabetes Federation, 2003.

[3] Strack, B., DeShazo, J.P., Gennings, C., Olmo, J.L., Ventura, S., Cios, K.J. and Clore, J.N., 2014. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. BioMed research international, 2014.

[4] UCI Machine Learning Repository. Diabetes 130-US hospitals for years 1999-2008 Data Set. Retrieved May 27, 2019 from https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

[5] Rubin, D.J. Hospital Readmission of Patients with Diabetes. Curr Diab Rep 15, 17 (2015). https://doi.org/10.1007/s11892-015-0584-7

[6] Muhammad Shahbaz, Shahzad Ali, Aziz Guergachi, Aneeta Niazi, Amina Umer (2019). "Classification of Alzheimer's Disease using Machine Learning Techniques", In Proceedings of the 8th International Conference on Data Science, Technology and Applications - Volume 1: DATA, ISBN 978-989-758-377-3, pages 296-303. DOI: 10.5220/0007949902960303

[7] Shahzad Ali, Muhammad Usman, Dawood Saddique, Umair Maqbool, Muhammad Usman Aslam, Shoaib Ejaz, "Prediction of Diabetes Disease using Data Mining Classification Techniques," Proceeding of Al Yamamah University Engineering Forum (YUENG), Riyadh, KSA, 10-11 March 2019.

[8] Hammoudeh, A., Al-Naymat, G., Ghannam, I. and Obied, N., 2018. Predicting hospital readmission among diabetics using deep learning. Procedia Computer Science, 141, pp.484-489.

[9] Rubin, D.J. Correction to: Hospital Readmission of Patients with Diabetes. Curr Diab Rep 18, 21 (2018). https://doi.org/10.1007/s11892-018-0989-1

[10] Duggal, R., Shukla, S., Chandra, S. et al. Predictive risk modelling for early hospital readmission of patients with diabetes in India. Int J Diabetes Dev Ctries 36, 519–528 (2016). https://doi.org/10.1007/s13410-016-0511-8

[11] Neto, C., Senra, F., Leite, J. et al. Different Scenarios for the Prediction of Hospital Readmission of Diabetic Patients. J Med Syst 45, 11 (2021). https://doi.org/10.1007/s10916-020-01686-4

[12] Rubin, D.J., Handorf, E.A., Golden, S.H., Nelson, D.B., McDonnell, M.E. and Zhao, H., 2016. Development and validation of a novel tool to predict hospital readmission risk among patients with diabetes. Endocrine Practice, 22(10), pp.1204-1215.

[13] Duggal, R., Shukla, S., Chandra, S. et al. Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India. Int J Diabetes Dev Ctries 36, 469–476 (2016). https://doi.org/10.1007/s13410-016-0495-4

[14] Abaya, S.A., 2012. Association rule mining based on Apriori algorithm in minimizing candidate generation. International Journal of Scientific & Engineering Research, 3(7), pp.1-4.

[15] Yuan, X., 2017, March. An improved Apriori algorithm for mining association rules. In AIP conference proceedings (Vol. 1820, No. 1, p. 080005). AIP Publishing LLC.

[16] Sinaga, K.P. and Yang, M.S., 2020. Unsupervised K-means clustering algorithm. IEEE access, 8, pp.80716-80727.

[17] El Khediri, S., Fakhet, W., Moulahi, T., Khan, R., Thaljaoui, A. and Kachouri, A., 2020. Improved node localization using K-means clustering for Wireless Sensor Networks. Computer Science Review, 37, p.100284.

[18] Tan, P.N., Steinbach, M. and Kumar, V., 2016. Introduction to data mining. Pearson Education India.

[19] Han, J., Pei, J. and Kamber, M., 2011. Data mining: concepts and techniques. Elsevier.