



**Table of Contents**

[2. ABSTRACT](#h.30j0zll)

[3. INTRODUCTION](#h.1fob9te)

[Healthcare delivery problem](#h.3znysh7)

[b. Data mining problem: Classify risk mortality](#h.2et92p0)

[c.Motivation](#h.3dy6vkm)

[4. LITERATURE REVIEW FOR THE PROBLEM](#h.4d34og8)

[5. PROBLEM FORMULATION](#h.u45ck3745fa)

[a. Data mining plan](#h.81q4bw1efxno)

[b.Pre-Processing](#h.17dp8vu)

[Principal Diagnosis coded in ICD-9-CM (PRINC\_DIAG\_CODE)](#h.26in1rg)

[Secondary Diagnosis coded in ICD-9-CM (OTH\_DIAG\_CODE)](#h.lnxbz9)

[Age (PAT\_AGE)](#h.35nkun2)

[Sex (SEX\_CODE)](#h.1ksv4uv)

[Discharge Disposition (PAT\_STATUS)](#h.2jxsxqh)

[Race (RACE)](#h.3j2qqm3)

[Ethnicity (ETHNICITY)](#h.1y810tw)

[Type of admission (TYPE\_OF\_ADMISSION)](#h.4i7ojhp)

[Admitting diagnosis (ADMITTING\_DIAGNOSIS)](#h.2xcytpi)

[Risk Mortality](#h.1ci93xb)

[Case Table and Nested Case Tables](#h.3whwml4)

[6. IMPLEMENTATION](#h.6d1vajmlbkd1)

[a.Data cleaning using SSIS.](#h.nf9k527p5uc1)

[Importing data from flat files](#h.qsh70q)

[Filter data for heart diseases](#h.3as4poj)

[Cleaning data and separating data into case and nested case tables](#h.1pxezwc)

[Reducing the number of categories in principal diagnosis code](#h.49x2ik5)

[Creating dummy variables](#h.2p2csry)

[Creating Data Source View (DSV)](#h.147n2zr)

[Data Exploration](#h.3o7alnk)

[Distribution of Risk Mortality on the basis of principal diagnosis](#h.23ckvvd)

[Distribution of Risk Mortality based on Type of Admissions](#h.ihv636)

[Distribution of Risk Mortality based on Race](#h.32hioqz)

[Distribution of Risk Mortality based on Ethnicity](#h.1hmsyys)

[Distribution of Risk Mortality based on Sex](#h.41mghml)

[Distribution of Risk Mortality based on Patient Discharge Status](#h.2grqrue)

[b. SQL Server based mining](#h.jz4ejme8qn5a)

[Data Source View (DSV)](#h.z48mizydz8ng)

[Algorithms](#h.uvajxkin95yz)

[Logistic Regression](#h.hnav8y9009lg)

[Results obtained by running different models](#h.gqsgaez71s2h)

[c. Prediction performed on new data](#h.dg7edbywbpmk)

[7. CONCLUSIONS](#h.vx1227)

[REFERENCES](#h.3fwokq0)

# **2. ABSTRACT**

The objective of this data mining project is to classify the risk mortality for patients suffering from circulatory or heart diseases. The data used to classify the risk mortality comes from the *Base 1* file for both Q1 and Q2. We have considered data for patients suffering from circulatory diseases by filtering the data based on the Principal Diagnosis (PRINC\_DIAG\_CODE). The ICD-9-CM values considered for PRINC\_DIAG\_CODE ranges from 390 to 459 (Circulatory Diseases). Based on our initial analysis of the data we found the final dataset to contain around 180,000 records of patients suffering from circulatory diseases. This report, in particular, talks about our approach towards cleaning the data and creating data source views using SQL Server Data Tools.

# **3. INTRODUCTION**

## **Healthcare delivery problem**

Our world has a population of more than 7 billion people today. This growing trend of global population has triggered some of the biggest healthcare challenges that the healthcare industry is facing. One of the major business problems in healthcare domain is the assessment and management of clinical problems in the hospital. Governments and patients evaluate a hospital's quality of care by looking at performance data. In many countries, the data used to compare and evaluate outcomes is frequently based on Diagnosis Related Groups (DRGs) [5]. This involves classifying a patient’s severity of illness or his mortality rate, or reducing the length of stay (LoS) of patients in the hospital based on the initial diagnosis.

For example, by classifying the patients based on the severity of the illness, the doctors can plan a treatment schedule well ahead of time. The hospital can also plan the logistics for the proposed treatment which can help in the efficient treatment of the patient. Similarly, by predicting the length of stay of a patient, hospital can solve the problem of manpower allocation. This can help the hospital in effective scheduling for admission of elective patients. [3]

## **b. Data mining problem: Classify risk mortality**

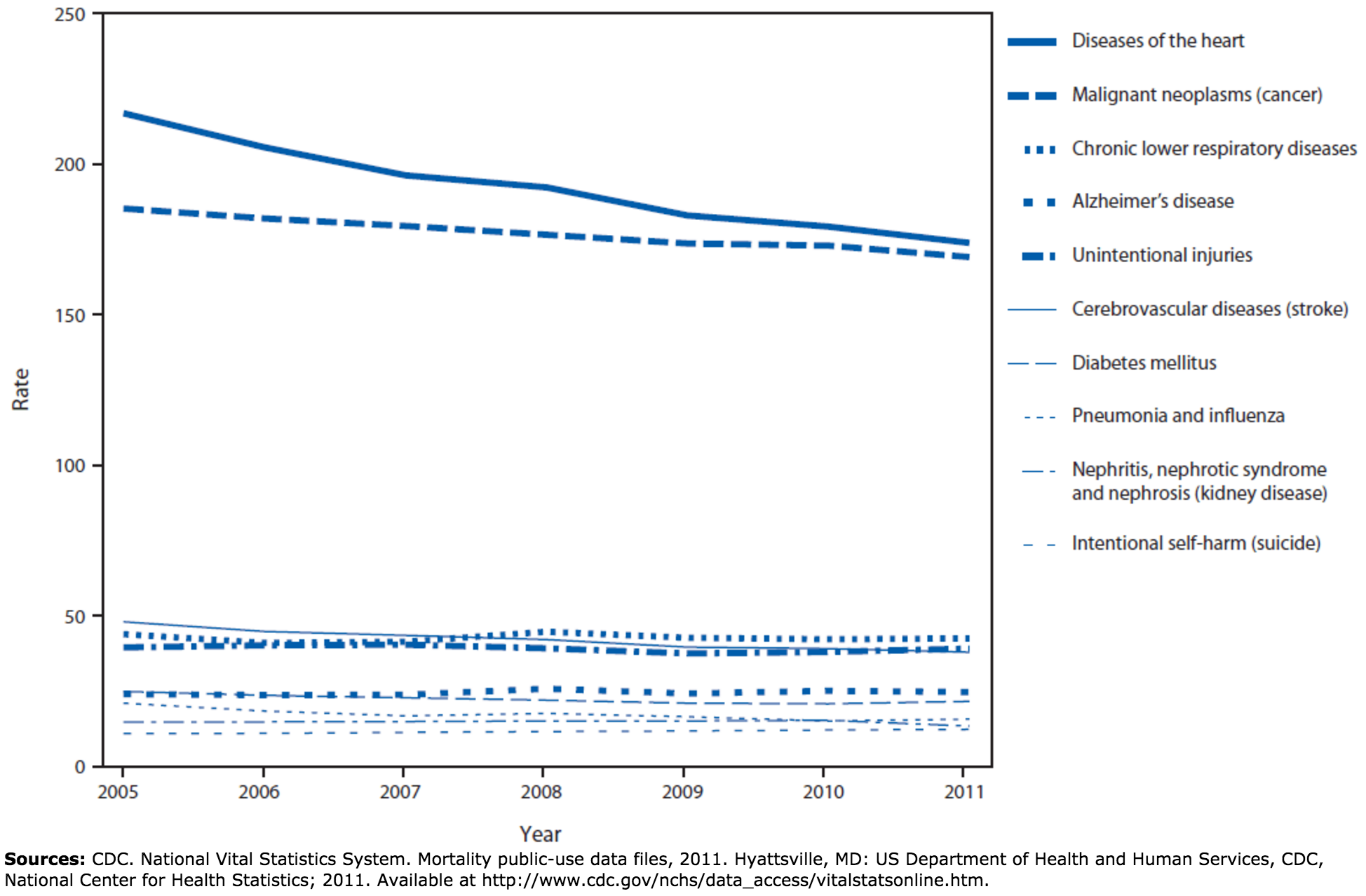
Risk of mortality is defined as the likelihood of dying. The four risk of mortality subclasses are numbered sequentially from 1 to 4 indicating respectively, minor, moderate, major, and extreme severity of illness. Based on our preliminary research we found that the variables listed below impact the risk of mortality: [4]

1. Principal Diagnosis coded in ICD-9-CM (PRINC\_DIAG\_CODE)
2. Secondary Diagnoses coded in ICD-9-CM (OTH\_DIAG\_CODE)
3. Procedures Coded in ICD-9-CM (PRINC\_ICD9\_CODE,OTH\_ICD9\_CODE)
4. Age (PAT\_AGE)
5. Sex (SEX\_CODE)
6. Discharge Disposition (PAT\_STATUS)

## **c.Motivation**

Risk mortality refers to the in-hospital deaths and is one of the most important factors that define the quality of service at hospitals. In most of the cases the risk mortality of the patient is estimated during admission using history of the patient either through intuition or experience of the hospital staff. This helps them provide better services and improve intensive care to higher need patients, but with increasing complexity of diseases and increasing influential factors, it is imperative to have a more effective and reliable scoring system to provide accurate information on the risk mortality of the inpatients. Hospitals are profiled based on risk-standardized rates with increasing emphasis on improving patient management.

In recent years, there has been a significant increase in the number of patients admitted for heart diseases due to degrading food lifestyle and less health awareness. With increasing incidents, there has also been an increase in the risk mortality of inpatients with heart issues. According to a recent survey by CDC, diseases of the heart is one of the leading cause of deaths among patients in the United States.



*Figure 1: Rate of deaths per 100,000 population, by leading cause of death — United States, 2005–2011*

**Every year more than 750,000 Americans have a heart attack and more than 610,000 people die of heart disease**. There following are some of the common heart diseases among the population:

* Rheumatic heart disease
* Coronary artery disease
* Aortic aneurysm
* Pulmonary heart disease

**Coronary heart disease alone kills more than 370,000 people every year in the United States of America** [14]. As seen in Fig 1. The deaths due to the diseases of heart has seen a slight decline, but there has not been a significant drop in the death rates given the technological advancement in the treatment of heart diseases in the last decade. This makes it imperative to find the leading factors that increase the mortality risk for patients of heart disease. This will enable the hospital management to prioritize services for patients with high pre-determined risk of mortality.

Listed below are some of the important factors that affect the risk mortality among patients with diseases of the heart [14]:

1. Age
2. Sex
3. Race
4. Non-cardiac history - diabetes etc.
5. Cardiac history - Previous attacks or heart diseases
6. Vital measures - measure of vital elements in blood
7. Dependencies - Alcohol or drug dependencies
8. Family history or genetic disorder

# **4. LITERATURE REVIEW FOR THE PROBLEM**

To deliver this project we have studied research papers from organizations that are in the forefront of medical research. Our understanding of the healthcare operations, problem formulations and various dependencies are based on research papers from *Journal of the American Medical Informatics Association*, *US National Library of Medicine National Institutes of Health*, *and American Health Information Management Association (AHIMA)* etc.

The healthcare industry has modernized its operations, and is increasingly adopting electronic health records (EHRs). This has led to the deployment of new health information technology systems that constantly create, collect, and manage their information. Therefore, the amount of data available to clinicians and hospital administrators in the healthcare domain is growing at an exponential rate. However, despite such advances in technology and proven success in making data driven decisions in other domains such as retail, marketing etc., healthcare providers often report only minor improvements in decision making capability through the use of existing data.[1]

Costs and risk are not spread evenly across a population in many healthcare systems. Therefore, relatively small number of patients who are classified as high-risk patients tend to consume or utilize more medical resources than their peers. Also, studies are showing that deficits in managing care for these patients could lead to higher expenses. These findings necessitate and warrant systematic efforts that focus on *identifying high risk patients* to ensure that they receive the most efficient and effective care possible. [2]

McKinsey has recently stated that big data analytics has the potential to enable savings of more than $300 billion per year in U.S. healthcare. Also, McKinsey believes big data could help minimize inefficiencies in the following three areas [3]:

1. *Clinical operations*
2. *Research & development*
3. *Public health*

# **5. PROBLEM FORMULATION**

## **a. Data mining plan**

Data Mining plan in our classification problem involves 5 stages

1. Data Cleaning or Pre-Processing
2. Data Integration
3. Data Selection and Transformation
4. Data Mining using SSAS
5. Algorithm selection, validation and testing

## **b.Pre-Processing**

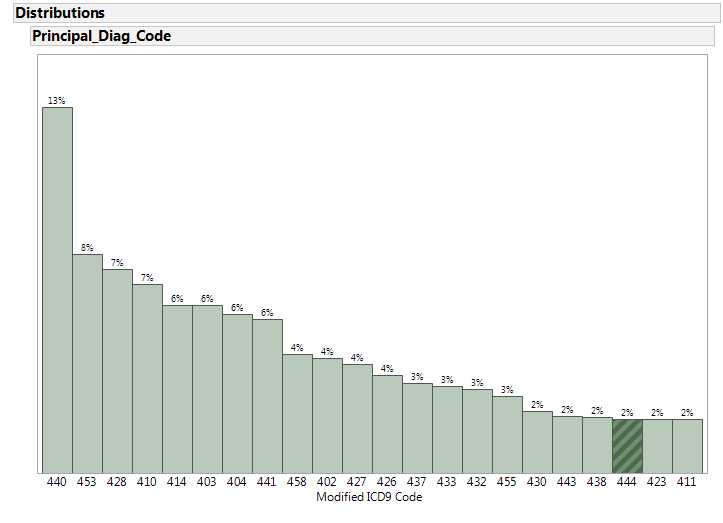
Since our main focus is on risk mortality associated with heart diseases we have included principal diagnosis codes starting from 390 to 459 (inclusive). Based on our exploration of this data we found that there are around 96,000 records in Q1 and around 89000 records in Q2.

### **Principal Diagnosis coded in ICD-9-CM (PRINC\_DIAG\_CODE)**

**Reason:**

The Principal Diagnosis is the condition established after ascertaining the condition that is chiefly responsible for the admission of the patient to the hospital for care.

Based on our research, Principal Diagnosis Code is one of the data elements that is used to determine the risk of mortality [4]. For the intents and purposes of this project, we are considering the ICD-9-CM codes that correspond to diseases related to the circulatory system.

**Data Exploration:**

*Inference:* In order to reduce the number of categories we have considered only the first 3 digits of the ICD 9 code for the principal diagnosis code column in the dataset. Figure 2 depicts the number of records in each of these *modified ICD 9 codes (first 3 digits)*. We can see that 440 (arterial diseases) constitute almost 13% of the 96000 records.

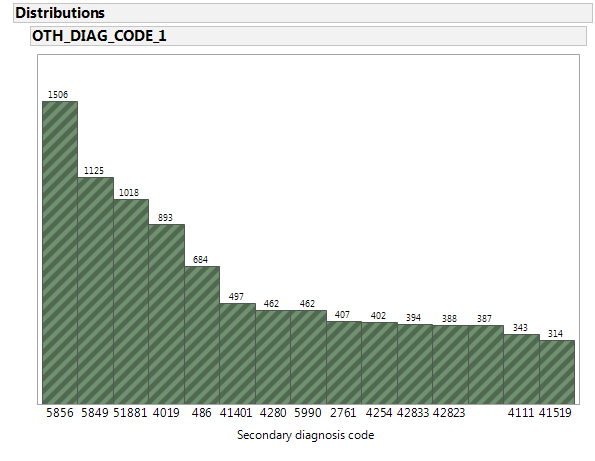
**Data cleansing approach:**

Since the records were filtered based on the ICD 9 codes column to get records pertinent to circulatory diseases, the PRINC\_DIAG\_CODE field does not contain any null or empty values that needs to be cleansed.

### **Secondary Diagnosis coded in ICD-9-CM (OTH\_DIAG\_CODE)**

**Reason:**

Other Diagnoses include all conditions that coexist at the time of inpatient admission or ambulatory surgical service, or develop subsequently, which affect the treatment received and/or length of stay. Based on our research, we believe that pre-existing medical condition could play a major role in determining the risk mortality of the patient.

**Data exploration:**

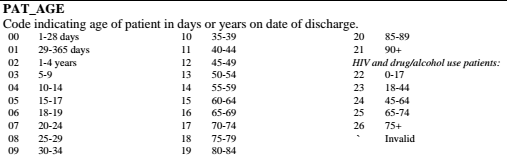
*Figure 3: Subset of distribution of number of records based on other diagnosis code (ICD 9 codes)*

*Inference:* As seen in figure 3, the distribution of records for secondary diagnosis is spread across multiple ICD 9 codes. Since this is just a supporting variable in cases where a principal diagnosis code is not present, we do not intend to clean this field.

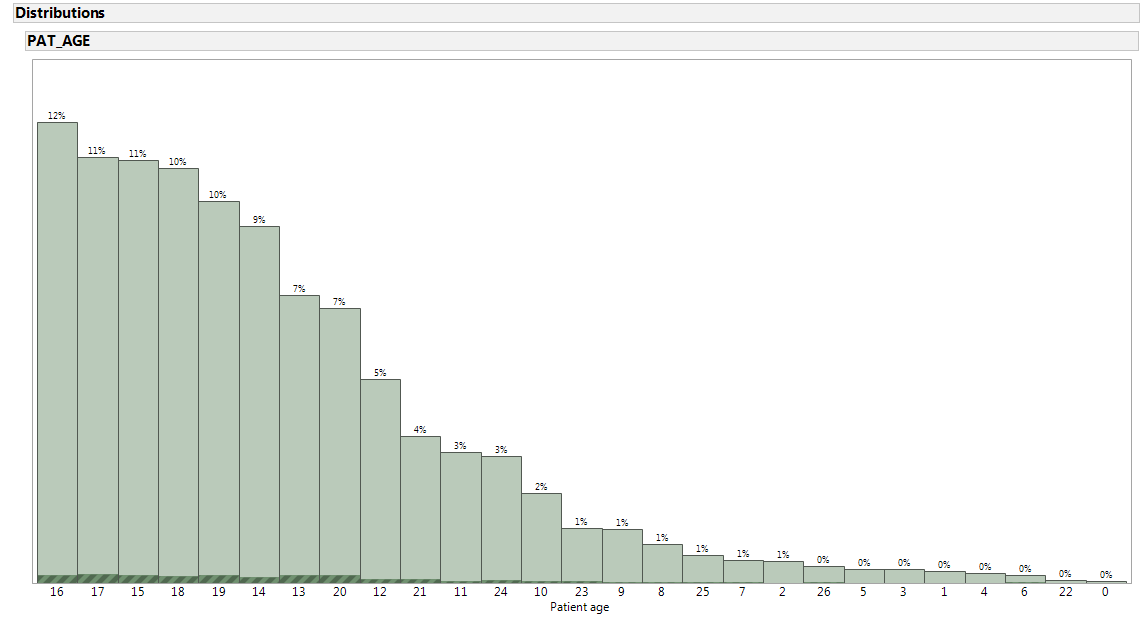
### **Age (PAT\_AGE)**

**Reason:**

Based on our research, patient age is one of the data elements that is used to determine the risk of mortality [4]. For example, the risk mortality of a patient who is 70 years old and suffers from an ischemic heart disease is higher than a patient whose age is 40 years and suffers from the same heart condition.In this particular figure below, the patient age group is taken instead of the actual patient’s age and the classification of age is as follows -



**Data Exploration:**

*Patient age group* 

*Figure 4. Distribution (percentage) of total records based on patient age*

*Inference:* From figure 4, we can understand that majority of the patients fall into the age group categories 14-19 which signify age groups 55-84. This is because we are considering the case of circulatory diseases which are more prevalent in people of this age group.

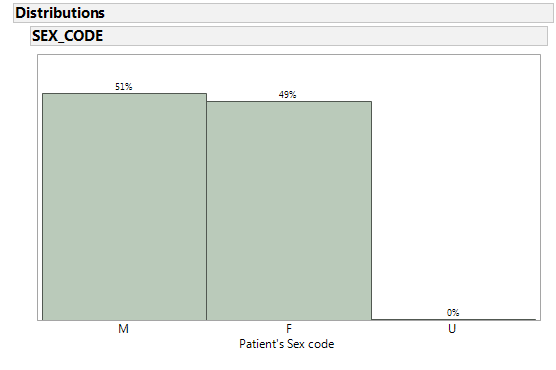
**Data cleansing approach:**

The field PAT\_AGE doesn’t contain any invalid or erroneous data values.

### **Sex (SEX\_CODE)**

**Reason:**

Based on our research, patient age is one of the data elements that is used to determine the risk of mortality [4]. For example, Heart attack symptoms in women may be different from those experienced by men. Many women who have a heart attack do not know it. Women tend to feel a burning sensation in their upper abdomen and may experience lightheadedness, an upset stomach, and sweating. Because they may not feel the typical pain in the left half of their chest, many women may ignore symptoms that indicate they are having a heart attack [10]. Accordingly the diagnosis for men and women may be different.

**Data exploration:**

*Figure 5. Distribution (percentage of total records) based on patient sex code*

*Inference:* As seen in figure 5, the distribution of data is pretty even among male and female genders. This will help in classifying the risk mortality for both men and women effectively.

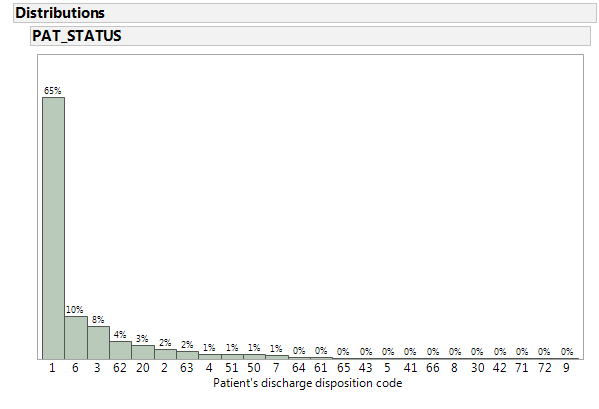
**Data cleansing approach:**

Upon investigation, we found that only a negligible percentage (4%) of the total number of records had empty values. These records can be removed while preprocessing.

### **Discharge Disposition (PAT\_STATUS)**

**Reason:**

A patient discharge status code is a two-digit code that identifies where the patient is at the conclusion of a healthcare facility encounter. Based on our research, patient age is one of the data elements that is used to determine the risk of mortality [4]. This parameter can provide us the details on how a patient discharge status can affect the risk of mortality. For example, if the risk mortality for acute myocardial infarction with a discharge code *07* (Left against Medical Advice or Discontinued Care) is high, the hospital can allocate adequate resource to take care of the patient once he is discharged.

**Data exploration:**

*Figure 6. Distribution (percentage of total records) based on patient’s discharge disposition code*

*Inference:*  As seen in figure 6, almost 65% of the patients had a *routine discharge*. This was expected.

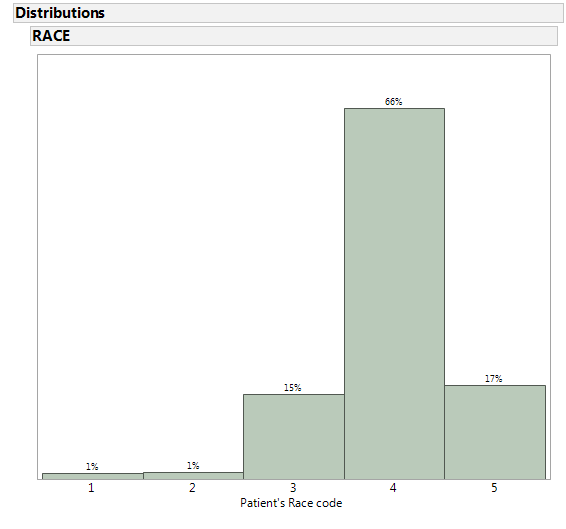
**Data cleansing approach:**

Upon investigation, we found that only a negligible number of records (247) of the total number of records (96000) had empty values. These records can be *removed* while preprocessing.

### **Race (RACE)**

**Reason:**

Based on our research, patient age is one of the data elements that is used to determine the risk of mortality [4]. For example, most of the studies found hypertension to be significantly higher in Blacks than Whites [15].

**Data exploration:**

*Figure 7: Distribution (percentage of total records) based on patient’s race code*

*Inference:* From figure 7, we can clearly see that most of the patients are ‘white’ (category 4). Also, 17% of the total number of patients are listed in the category 5 which belongs to ‘other’.

**Data cleansing approach:**

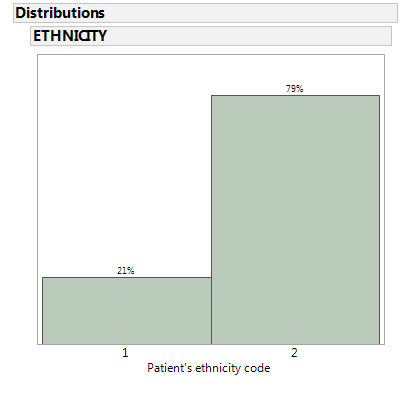
Upon investigation, we found that a negligible number of records (411) of the total number of records (around 96000) had empty values. These records can be *removed* while preprocessing.

### **Ethnicity (ETHNICITY)**

**Reason:**

Changes in circulatory diseases differ by race as well as ethnicity. To better differentiate various trends between African-American, Hispanic, and White ethnic groups in circulatory disease we have included the ethnicity as one of the data elements that is used to determine the risk of mortality.

**Data exploration:**



*Figure 7. Distribution (percentage of total records) based on patient’s ethnicity code*

*Inference:* From figure 7, we can observe that most of the patients are not of Hispanic origin (category 2). Also, 21% of the total number of patients are listed in the category 1 which belongs to ‘Hispanic origin’ as expected since we are dealing with data from Texas which has a significant Hispanic population.

**Data cleansing approach:**

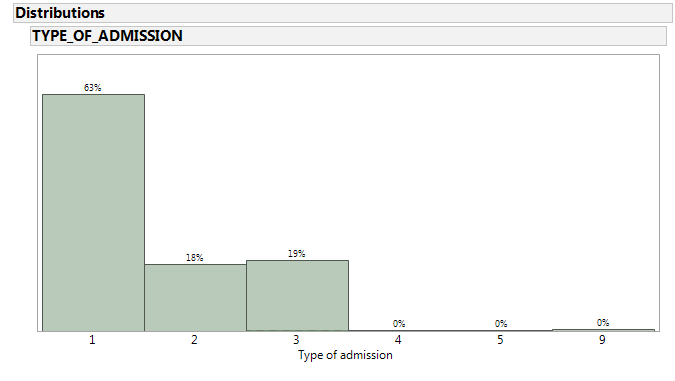
Upon investigation, we found that a negligible number of records (187) of the total number of records (around 96000) had empty values. These records can be *removed* while preprocessing.

### **Type of admission (TYPE\_OF\_ADMISSION)**

**Reason:**

This code indicates the manner in which the patient was admitted to the health care facility. The type of admission (emergency, trauma, urgent etc.) can determine the severity of the patient’s condition, which can in turn have an impact on the patient’s risk of mortality.

**Data exploration:**



*Figure 9: Distribution (percentage of total records) based on type of admission*

*Inference:* From figure 9, we can clearly see that most of the admissions are emergency admissions (category 1). Also, 18% of the total number of admissions were listed in the category 2 which belongs to ‘urgent admissions’.

**Data cleansing approach:**

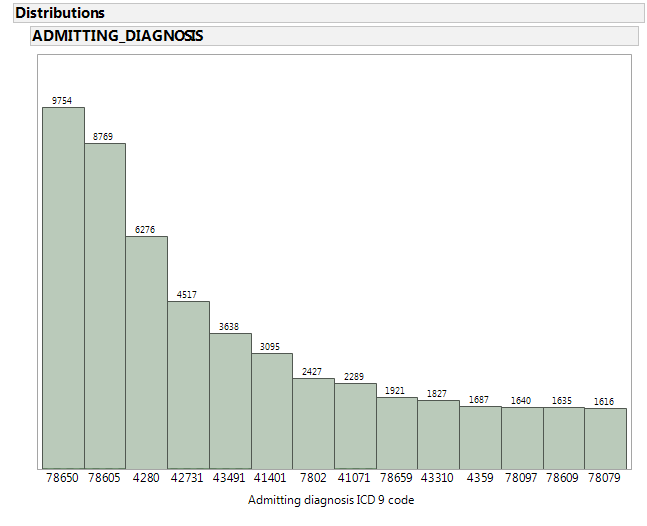
Upon investigation, we found that though there are no empty values in the field but a negligible number of records (247) of the total number of records (around 96000) belong to category 9 (no information available). These records can be *removed* while preprocessing.

### **Admitting diagnosis (ADMITTING\_DIAGNOSIS)**

**Reason:**

This code indicates the admitting diagnosis of the patient who was admitted to the health care facility. The type of admission (ICD-9-CM Codes) can accurately determine the type of condition that patient suffered from when he was initially admitted. Certain admission codes can indicate potential seriousness of the patient condition, which in turn can impact the risk of mortality.

**Data exploration:**



*Figure 10. Distribution (percentage of total records) based on admitting diagnosis ICD9 code*

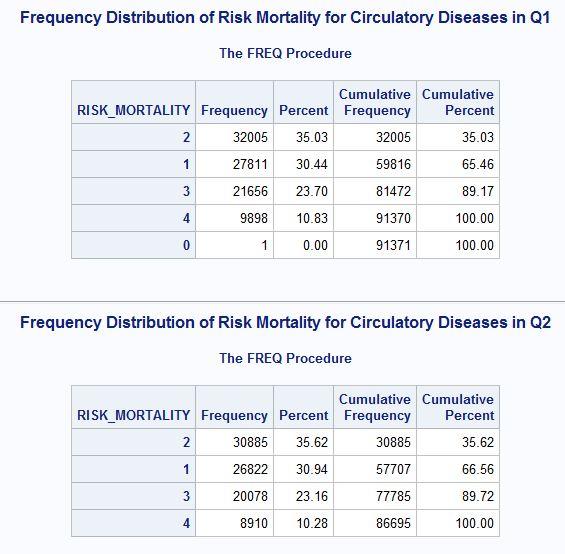
*Inference:* From figure 10, we can clearly see that majority of the patients are admitted with an ICD 9 code *786.50 (unspecified chest pain) followed by 786.05 (shortness of breath)* which are symptoms of circulatory diseases.

**Data cleansing approach:**

Upon investigation, we found that a negligible number of records (94) of the total number of records (around 96000) had empty values. These records can be *removed* while preprocessing.

### **Risk Mortality**

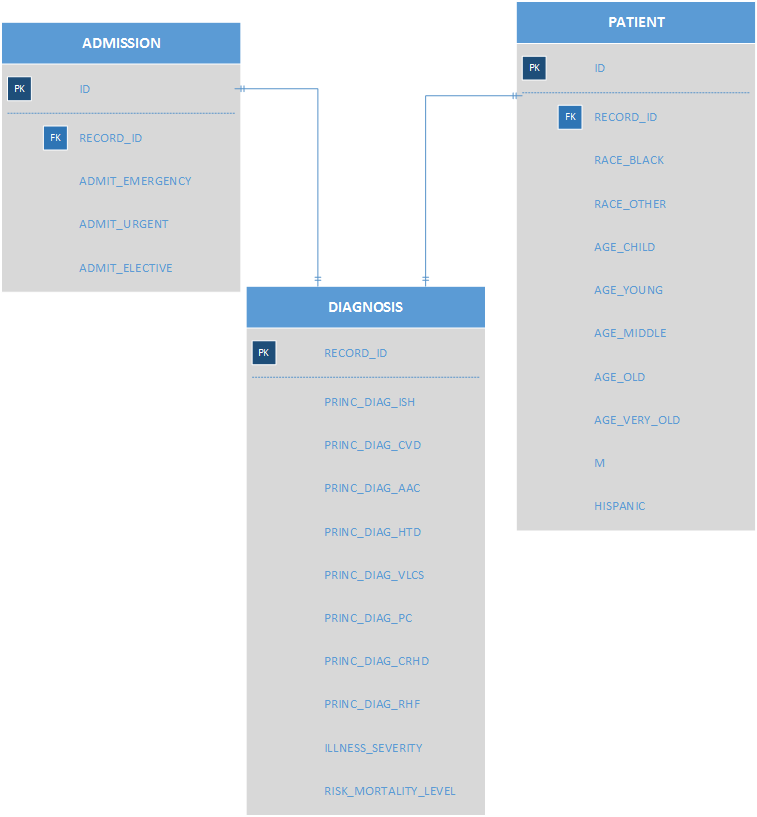
This variable is the output variable which we are going to classify in our data mining project.



*Figure 11: Distribution (number of records) for every class of risk mortality*

In figure 11, we have indicated the number of records under each class of risk mortality for both the quarters Q1 and Q2 separately. In our classification problem, we consider output risk mortality classes 3 (major) and 4 (extreme) important since these demand special medical attention. Thus, from figure 11 we can understand that there might not be a need for oversampling as there around 33% of the records belong to risk mortality classes 3 or 4 in both the quarters.

## **Case Table and Nested Case Tables**



*Figure 12: Case table and nested case tables*

# **6. IMPLEMENTATION**

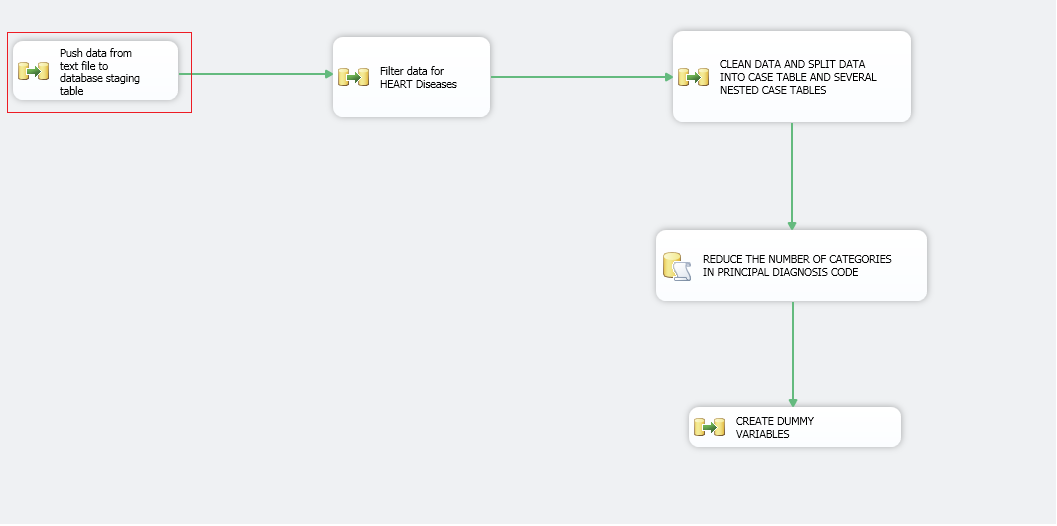
## **a.Data cleaning using SSIS.**

Table 1 below compiles the list of tasks that we are doing to cleanse the data. For each variable of significance, we are checking if it is NULL or empty, special character, whether a filter is required and whether they need to be grouped together.

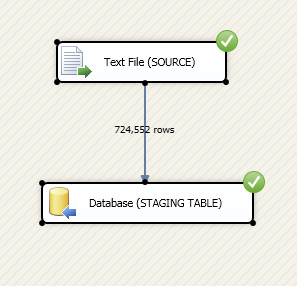
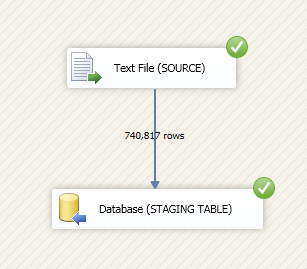
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Cleaning Criteria | | | | |
| Column | **NULL Removal** | **Special Character** | **Filter data/Subset** | **Grouping to categories** | **Comments** |
| Record\_ID | Yes | Yes | No | No |  |
| Discharge | Yes | Yes | No | Yes |  |
| THCIC\_ID | Yes | Yes | No | Yes |  |
| Provider\_name | Yes | Yes | No | Yes |  |
| Type\_of\_Admission | Yes | Yes | No | Yes |  |
| Source\_of\_Admission | Yes | Yes | No | Yes |  |
| Pat\_State | Yes | Yes | No | Yes |  |
| Pat\_ZIP | Yes | Yes | No | No |  |
| PAT\_Country | Yes | Yes | No | No |  |
| Pat\_status | Yes | Yes | No | No |  |
| Sex\_code | Yes | Yes | No | No |  |
| Race | Yes | Yes | No | No |  |
| Ethnicity | Yes | Yes | No | No |  |
| Length of Stay | Yes | Yes | No | No |  |
| Pat\_age | Yes | Yes | No | No |  |
| Admitting diagnosis | Yes | No | No | Yes |  |
| Princ\_Diag\_Code | Yes | No | Yes | Yes |  |
| OTH\_DIAG\_CODE\_1 | Yes | No | No | No |  |
| OTH\_DIAG\_CODE\_2 | Yes | No | No | No |  |
| OTH\_DIAG\_CODE\_3 | Yes | No | No | No |  |
| PRINC\_SURG\_PROC\_CODE | Yes | No | No | No |  |
| PRINC\_ICD9\_CODE | Yes | No | No | No |  |
| Risk\_mortality | Yes | No | No | No |  |
| Illness\_severity | Yes | No | No | No |  |
| apr\_drg | Yes | No | No | No |  |

*Table 1: Cleaning criteria for different columns*

## **Importing data from flat files**



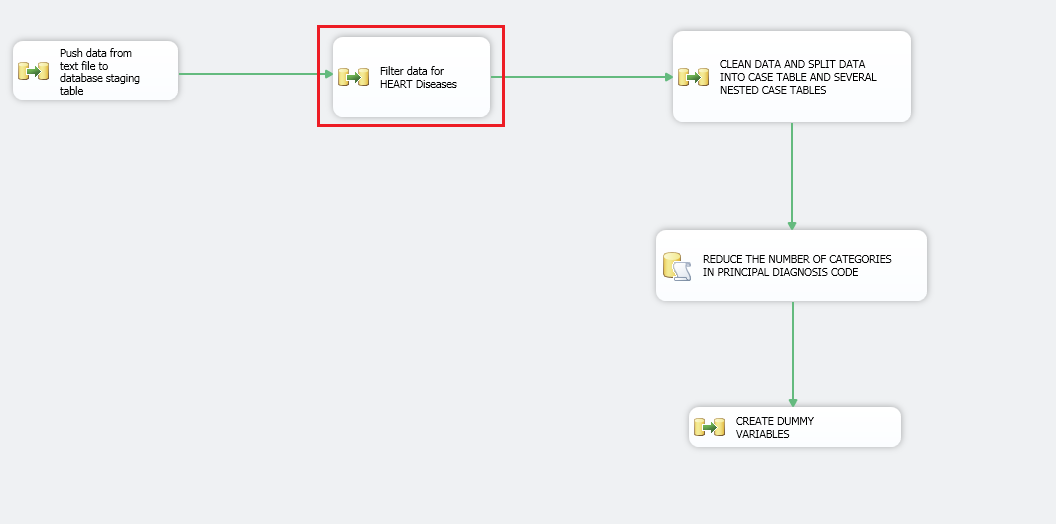
*Figure 13: Overall control flow*



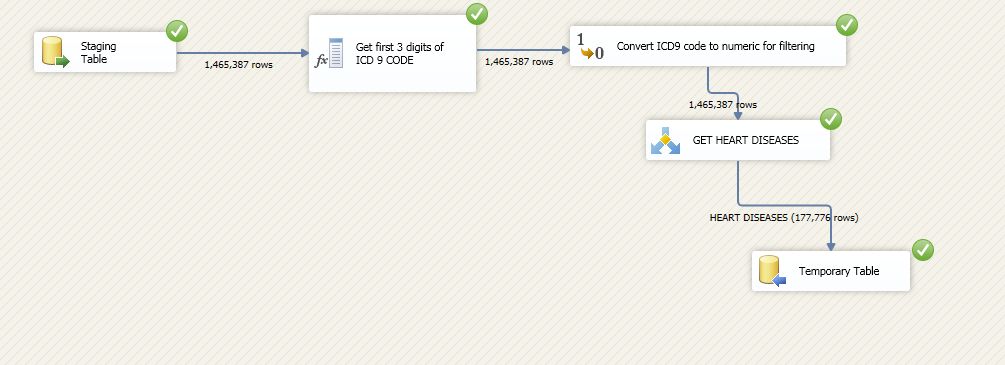
*Figure 14: Importing data from flat file to temporary staging table*

As shown in figure 13, we imported data from the base1 files of quarter 1 and quarter 2. The data flow for the first control flow block (highlighted) is shown in figure 14. A total of 1,465,369 records were imported from the flat files.

## **Filter data for heart diseases**

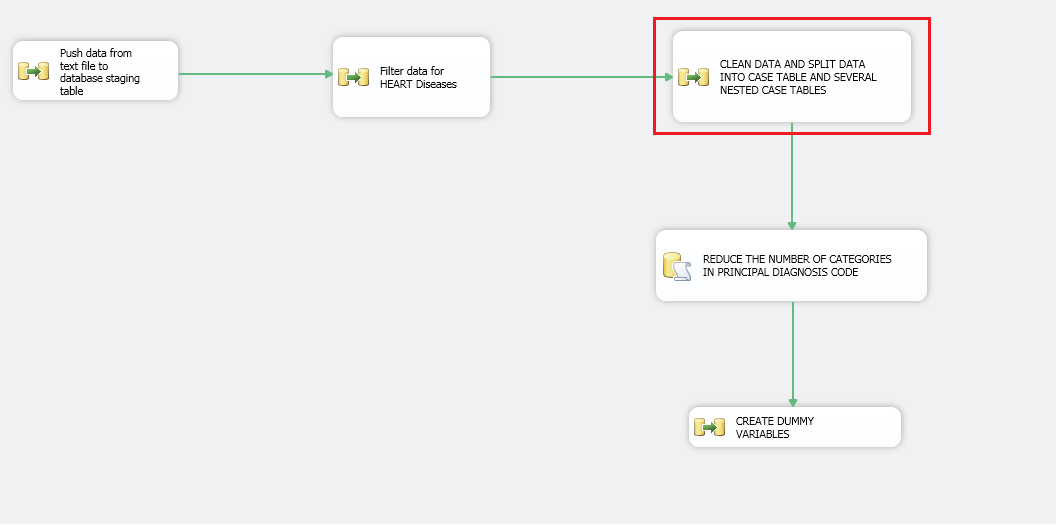


*Figure 15: Overall control flow*

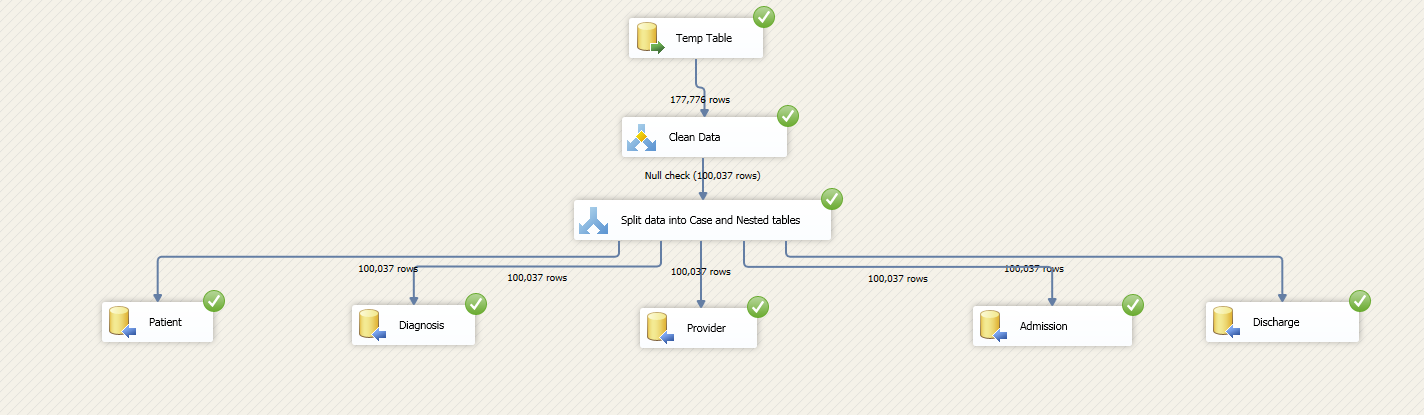
*Figure 16: Data flow for the control flow “Filter data for HEART diseases”*

As shown in figure 16, the first three characters were extracted from the Principal diagnosis code (ICD 9) and then it was converted to a numeric field and records with values in the range 390 to 459 (inclusive) which represents heart diseases. As shown in figure 16, after filtering we ended up with 177,776 records which were pushed into a temporary table.

## **Cleaning data and separating data into case and nested case tables**



*Figure 17: Overall control flow*

*Figure 18: Data flow for the control flow highlighted in figure 17*

As shown in figure 18, the data was cleaned by removing any nulls or empty values. Also, any special characters were removed from the columns according to the criteria mentioned in table 1. Then, specific columns were selected from the temporary table and data was moved to the case table (Diagnosis) and other nested case tables. **As we can see after cleaning we have approximately 100,000 observations.**

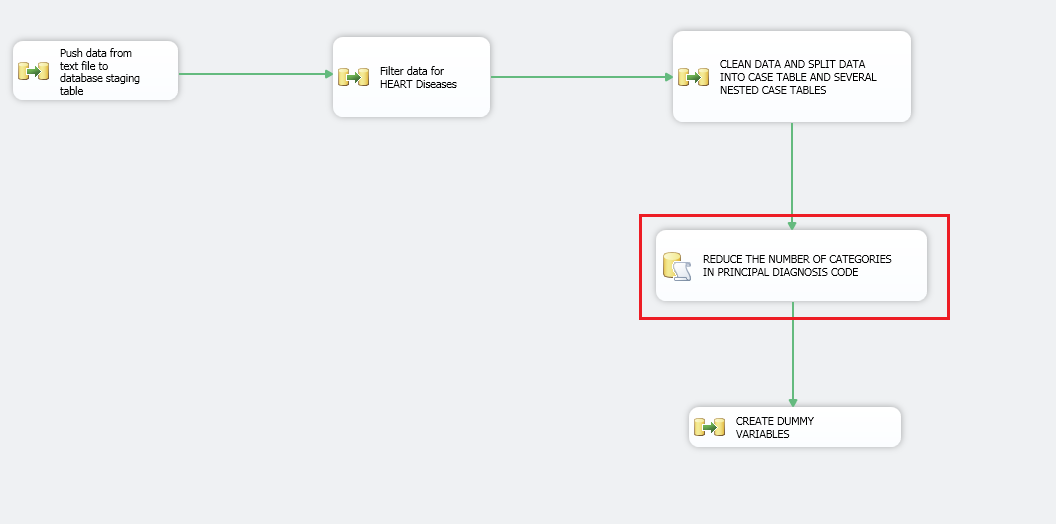
## 

## 

## 

## 

## **Reducing the number of categories in principal diagnosis code**



*Figure 18: Overall control flow*



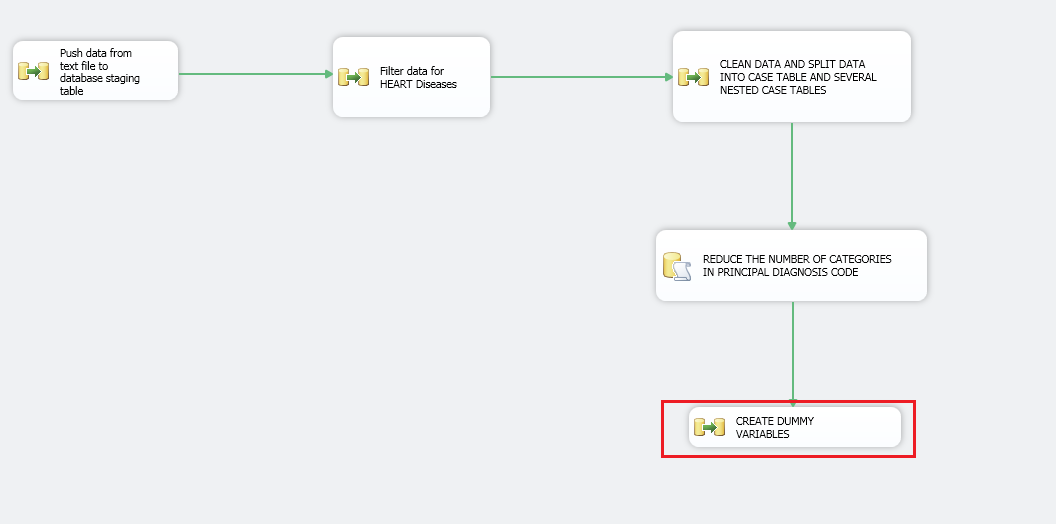
*Figure 19: SQL script used in the execute SQL task highlighted in figure 18*

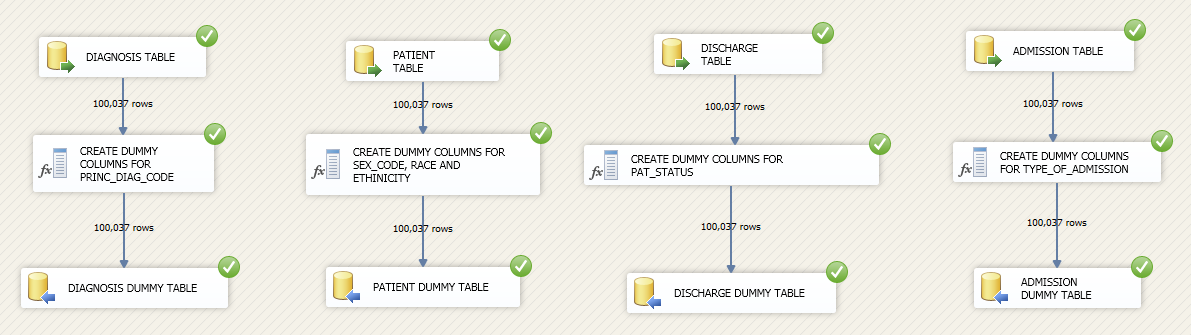
As shown in figure 19, we used a SQL script to reduce the number of categories in Principal diagnosis code**.** Table 2 explains the different abbreviations used in the categorization of principal diagnosis codes.

|  |  |  |
| --- | --- | --- |
| Abbreviation | Actual Name | Range (based on first 3 digits of ICD9 code) |
| RHF | Rheumatic fever | 390-392 |
| CRHD | Chronic rheumatic heart disease | 393-398 |
| HTD | Hypertensive disease | 401-405 |
| ISH | Ischemic heart disease | 410-414 |
| PC | Diseases of pulmonary circulation | 415-417 |
| OTH | Other forms of heart disease | 420-429 |
| CVD | Cerebrovascular disease | 430-438 |
| AAC | Diseases of arteries, arterioles, and capillaries | 440-448 |
| VLCS | Diseases of veins and lymphatics, and other diseases of circulatory system | 451-459 |

*Table 2: Categorization of principal diagnosis codes*

## **Creating dummy variables**

*Figure 20: Overall control flow*

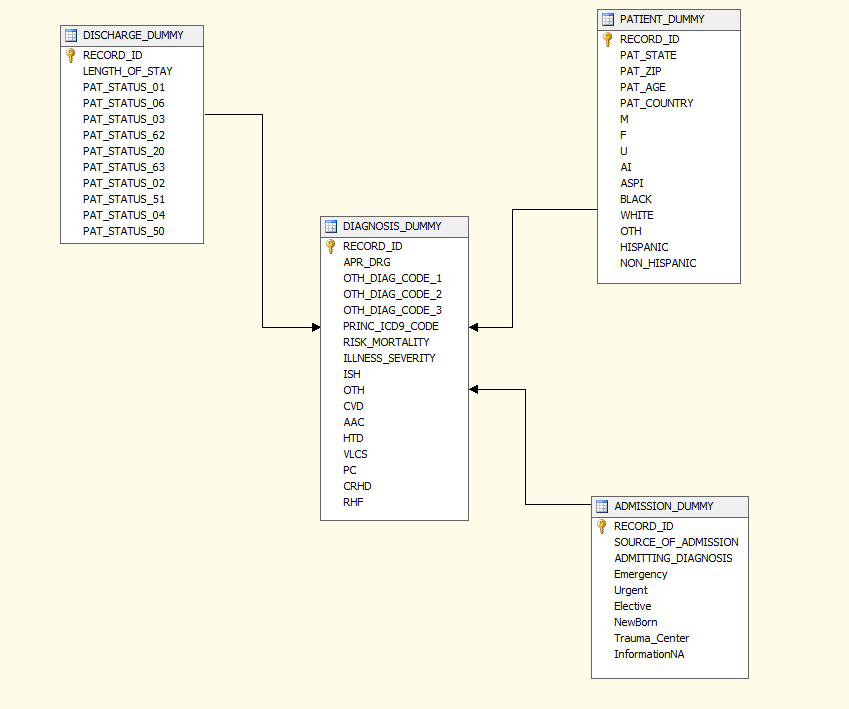


*Figure 21: Data flow for the control flow “Create Dummy variables”*

As shown in figure 21, certain columns in case table and nested case tables were converted into dummy variables using SSIS. The columns which were converted into dummy variables are as follows:

* Principal diagnosis code
* Sex\_Code
* Race
* Ethnicity
* Patient discharge status
* Type of admission

## **Creating Data Source View (DSV)**



*Figure 22: Data source view with case and nested case tables*

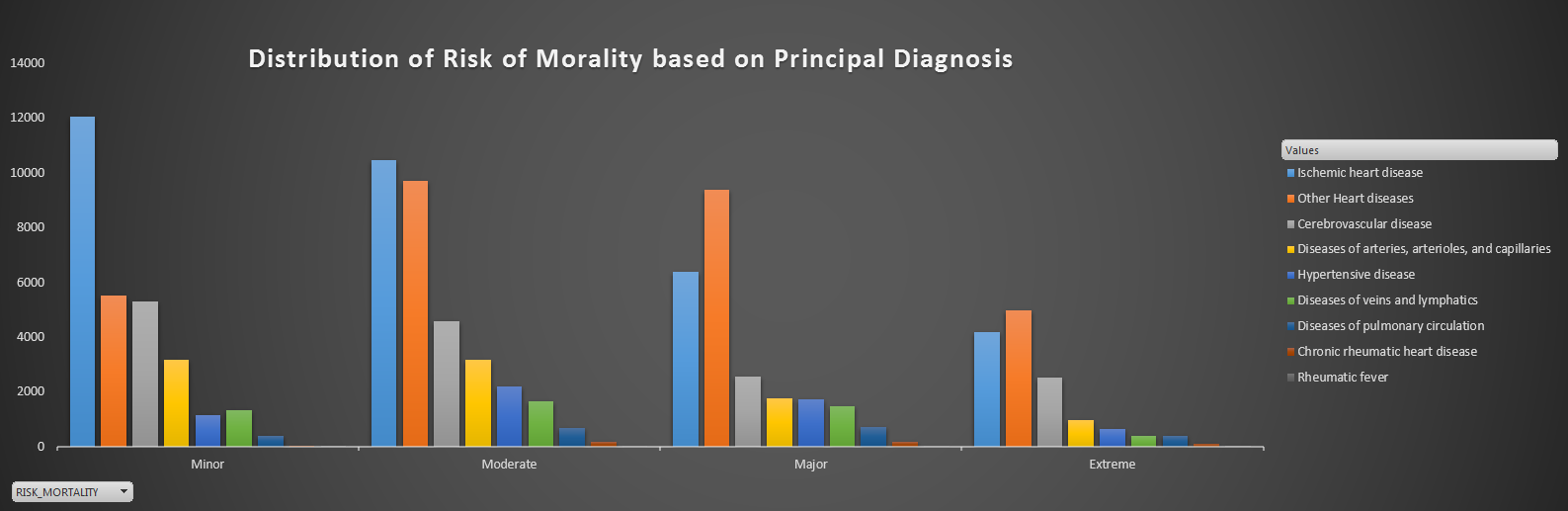
As seen in figure 22, the data source view consists of a case table (DIAGNOSIS\_DUMMY) and nested case tables (DISCHARGE\_DUMMY, PATIENT\_DUMMY, ADMISSION\_DUMMY). There are a few dummy variables which were derived from the original data as seen.

Data explorations in Excel using DSV

After creating a DSV, we explored the data further to look for any further relations between the various variables.

## **Data Exploration**

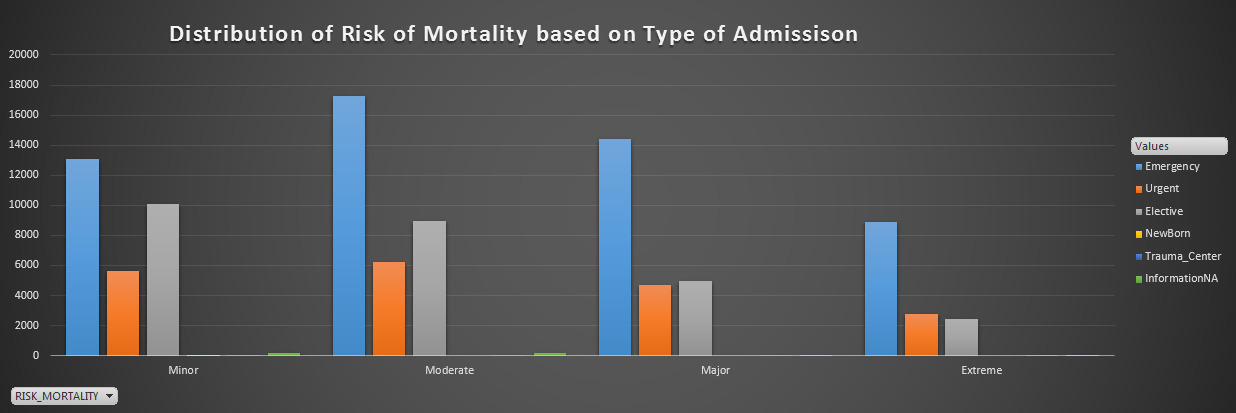
### **Distribution of Risk Mortality on the basis of principal diagnosis**



*Figure 23: Risk Mortality with respect to Principal Diagnosis*

*Inference*: Figure 23 describes the Risk Mortality with respect to Principal Diagnosis. The top three contributors of all risk mortality is Ischemic Heart Diseases, Cerebrovascular Diseases and Other Heart Diseases (Cardiamyopathy, Cardiac Dysrhythmias etc.). With the help of this graph we can ascertain the effect on mortality for each of the categories of Principal Diagnosis.

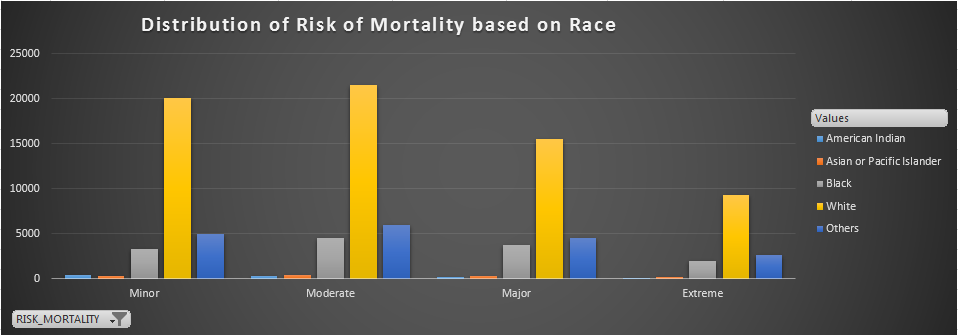
### **Distribution of Risk Mortality based on Type of Admissions**



*Figure 24: Risk Mortality with respect to Type of admissions*

*Inference*: Figure 24 describes the Risk Mortality with respect to Type of Admissions. The graph shows that Emergency admission type is the top contributor to mortality at all levels. This graph could be used to bolster the emergency care services to keep the mortality cases as low as possible.

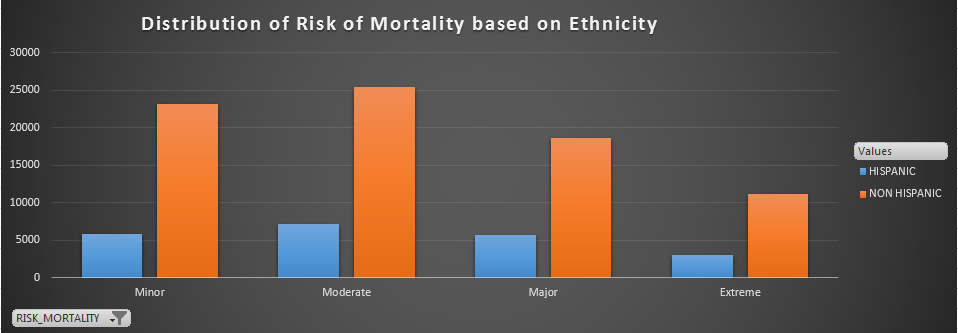
### **Distribution of Risk Mortality based on Race**



*Figure 25: Risk Mortality with respect to Race*

*Inference*: Figure 25 describes the Risk Mortality with respect to Race. The graph shows that Whites are the top contributor to mortality at all levels. This could be because the proportion of whites among the population is large as compared to other races whose data was collected.

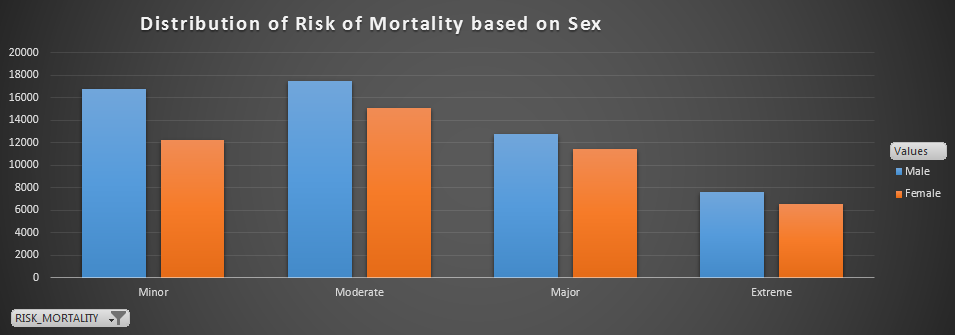
### **Distribution of Risk Mortality based on Ethnicity**



*Figure 26: Risk Mortality with respect to Ethnicity*

*Inference*: Figure 26 describes the Risk Mortality with respect to Ethnicity. The graph shows that Non-Hispanics are the top contributor to mortality at all levels. This could be because the proportion of Non-Hispanics among the population is large as compared to people of other ethnic backgrounds whose data was collected.

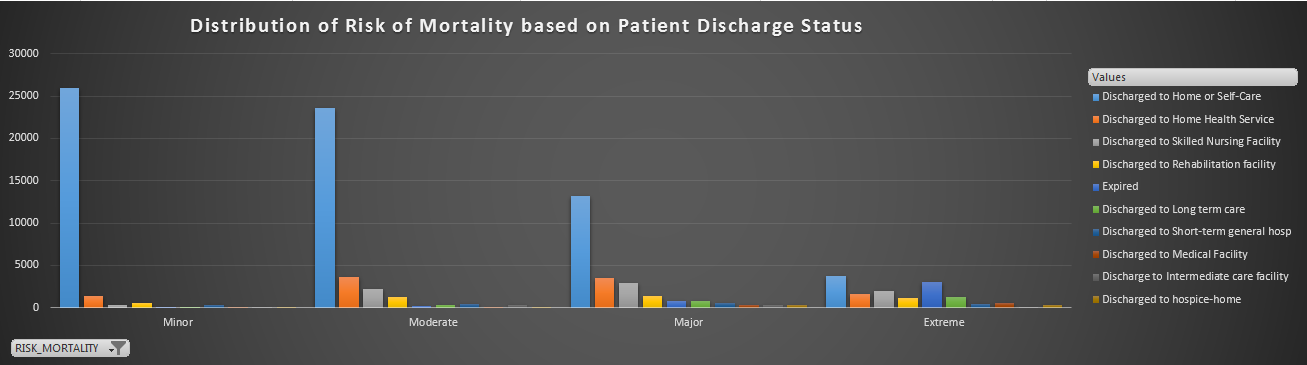
### **Distribution of Risk Mortality based on Sex**



*Figure 27: Risk Mortality with respect to Sex*

*Inference*: Figure 27 describes the Risk Mortality with respect to Sex. The graph shows that Men are the top contributor to mortality at all levels. Both men and women are equally affected by coronary heart diseases (CHD). Though mortality of men due to CHD’s are higher as compared to women, women have unique symptoms that may not be manifested in men [16].

### **Distribution of Risk Mortality based on Patient Discharge Status**

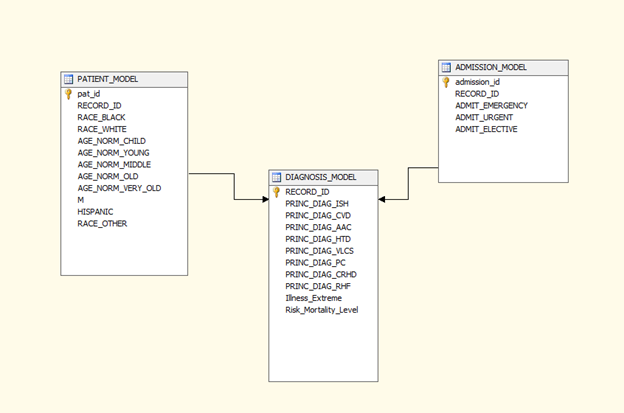


*Figure 28: Risk Mortality with respect to Patient Discharge*

*Inference*: Figure 28 describes the Risk Mortality with respect to Patient Discharge. The graph shows that Discharge to home or self-care category is the top contributor to minor, moderate and major mortality. While for extreme mortality cases the contribution is equal among the various types of discharge status of the patient. This graph also tells us how the hospitals should be more involved with the patients after the treatment rather than allowing them to go home or opt for self-care.

## **b. SQL Server based mining**

### **Data Source View (DSV)**



*Figure 28: Data Source View (DSV)*



As shown in figure 28, the data source view was created with a case table that focuses on the diagnosis related data and two nested tables - Patient and Admission. Since, the aim is to classify the risk mortality the case table (*Diagnosis\_Model*) stores diagnosis related information like DIAGNOSIS codes and Illness extremity that are critical factors of risk mortality and the nested tables (*Patient\_Model* and *Admission\_Model*) store patient’s demographic and admission information respectively. As seen in all the three tables, dummy variables were created for all categorical variables.

### **Algorithms**

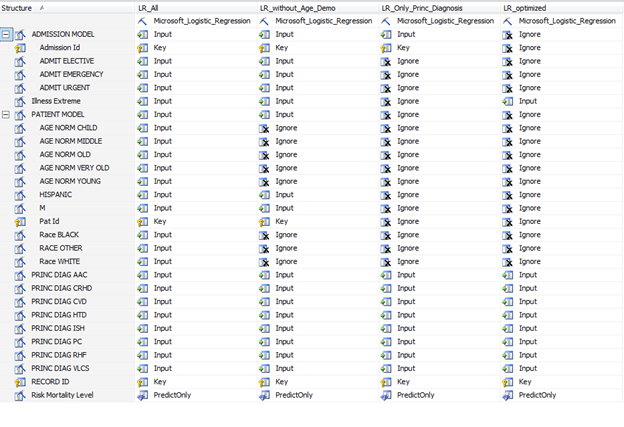
Since our predictor variables are categorical and outcome variable is categorical we had to choose among the following algorithms -

* Logistic Regression
* Naive Bayes

We tried to run the same model using different algorithms and found that logistic regression seemed to produce the best results. Thus, we decided to adopt logistic regression for our problem.

**Note:** Another important factor that we consider an important factor for our classification model is the reduction of false negatives. This scenario occurs when a particular patient is classified as “*Not Critical*” when the patient is actually “*Critical*”. This increases the risk for the patient due to lack of immediate or critical care required for his/her ailment. Therefore, we have made sure that it is one of the important factor while considering the best model for our risk mortality classification.

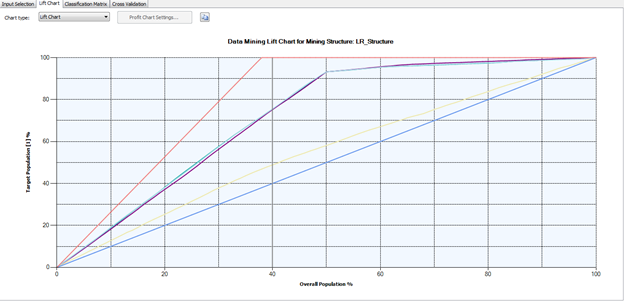
### **Results obtained by running different models using Logistic Regression**

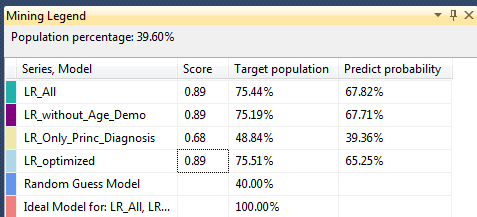


*Figure 29: Different models run using logistic regression*

As shown in figure 29, logistic regression was applied to a different subset of predictors from the mining structure. Also, after each model was run the predictor variables which did not contribute significantly to the classification accuracy were removed. In other words, if the model produced the same classification accuracy even after one of the variables were to be removed, these variables were not considered for the next iteration. This is to obtain a parsimonious result with maximum efficiency.

1. **Comparison of different models**

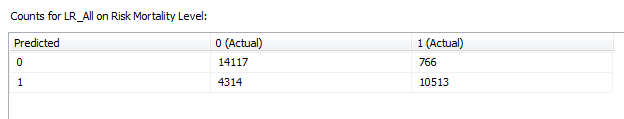
*Figure 30: Lift charts for different models run using logistic regression*



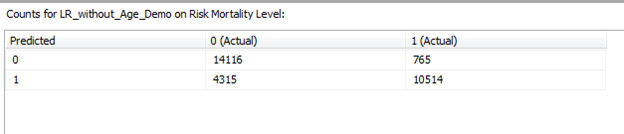


*Figure 31: Legend*

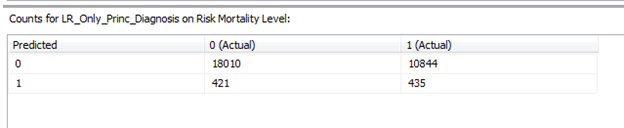
From figures 30 and 31, we can see that the score remains the same for models LR\_All, LR\_without\_age\_demo and LR\_optimized whereas LR\_only\_princ\_diagnosis has a significantly lesser score compared to the other three models. This is because in LR\_only\_princ\_diagnosis, illness\_severity is not included which led to a drop in the lift as reflected by the yellow line in figure 30.



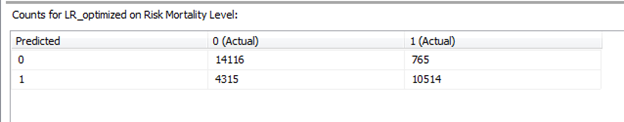
*Figure 31: Classification matrix for the model “LR\_All***”**



*Figure 32: Classification matrix for the model “LR\_without\_age\_demo***”**



*Figure 33: Classification matrix for the model “LR\_only\_Princ\_Diagnosis***”**



*Figure 34: Classification matrix for the model “LR\_Optimized”*

As seen fromfigure 31-34, the misclassification error is the same for models LR\_without\_age\_demo, LR\_all and LR\_Optimized. The number of misclassifications become drastically high for the model LR\_only\_Princ\_Diagnosis because it does not include illness severity as a predictor variable.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Name** | **Description** | **No. of variables** | **Misclassification error rate** |
| *LR\_All* | includes all variables | 25 |  |
| *LR\_without\_age\_demo* | excludes patient age and demographic information | **15** | **17.1** |
| *LR\_only\_princ\_diagnosis* | includes only principal diagnosis codes | **8** | **37.9** |
| *LR\_Optimized* | includes principal diagnosis codes and illness severity | **9** | **17.1** |

*Table 3: Comparison of models run using logistic regression*

1. **Recommendation of a best model and evidence**

From running different models consisting of different subsets of predictors using logistic regression we find **the model with the least misclassification error and which makes use of minimal number of variables is LR\_Optimized.** It uses the following predictor variables :

1. PRINC\_DIAG\_RHF - Rheumatic fever
2. PRINC\_DIAG\_CRHD - Chronic rheumatic heart disease
3. PRINC\_DIAG\_HTD - Hypertensive disease
4. PRINC\_DIAG\_ISH - Ischemic heart disease
5. PRINC\_DIAG\_PC - Diseases of pulmonary circulation
6. PRINC\_DIAG\_OTH - Other forms of heart disease
7. PRINC\_DIAG\_CVD - Cerebrovascular disease
8. PRINC\_DIAG\_AAC - Diseases of arteries, arterioles, and capillaries
9. PRINC\_DIAG\_VLCS - Diseases of veins and lymphatics, and other diseases of circulatory system
10. Illness Severity

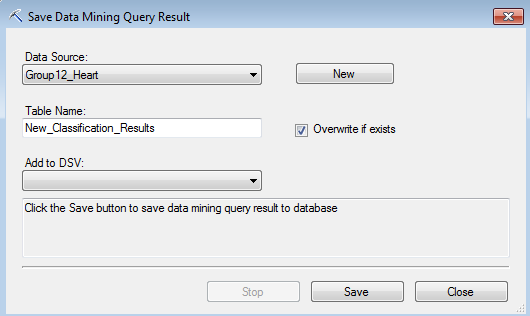
As parsimony is an important feature in selection of the best model in data mining, we recommend LR\_Optimized (which consists of the 10 predictor variables as shown above) as the best model.

## **c. Classification performed on new data**

### As seen in the previous section we ran our model against a test dataset to determine the misclassification error rate**.** In order to test the best model we ran the model (LR\_Optimized) against approximately 993 records which are new to the model

### 

*Figure 35: Selecting the table to predict new data*



*Figure 36: Saving the classification results in a table in the database*

As seen in figure 36, the classification was done on the new data and all the predictor information along with the actual and predicted class values were stored in a database table. These values were used in creating the classification matrix as shown in figure 37.

|  |  |  |
| --- | --- | --- |
| Predicted | 0 (Actual) | 1(Actual) |
| 0 | 457 | **39** |
| 1 | 151 | 346 |

*Figure 37: Classification matrix after running the best model on new data*

From the classification matrix shown in figure 37, we can calculate the misclassification error as 19.13 % which is not drastically higher than that of the test data set. Thus, we can infer that this is a good classifier.

# **7. CONCLUSIONS**

We have successfully developed a classifier for classifying patients who suffer from circulatory diseases into either high risk or low risk category using logistic regression. This will enable hospitals in providing expedited services and facilities to such patients thereby preventing the number of deaths that occur due to such circulatory problems. From our analysis, we have found that the two major parameters which help in predicting the risk mortality level are the Principal diagnosis code and illness severity.

# 

# **REFERENCES**

1. Jesus J Caban and David Gotz.Visual analytics in healthcare – opportunities and research challenges. *Journal of the American Medical Informatics Association.*
2. David Gotz, Harry Stavropoulos and Jimeng Sun. ICDA: A Platform for Intelligent Care Delivery Analytics. *US National Library of Medicine National Institutes of Health*
3. Gordon H. Robinson, Louis E. Davis, and Richard P. Leifer. Prediction of Hospital Length of Stay. *Health Services Research*
4. 3M Health Information Systems. APRTM-DRG Classification Software -Overview
5. P. Fontaine.Using Severity Adjustment Classification for Hospital Internal and External Benchmarking. *2004 IFHRO Congress & AHIMA Convention Proceedings, October 2004*
6. Length of Stay (LoS). Wikipedia. [*https://en.wikipedia.org/wiki/Length\_of\_stay*](https://en.wikipedia.org/wiki/Length_of_stay)
7. Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. Health Information Science and Systems.
8. Laursen TM, Wahlbeck K, Hällgren J, Westman J, Ösby U, Alinaghizadeh H, et al. (2013) Life Expectancy and Death by Diseases of the Circulatory System in Patients with Bipolar Disorder or Schizophrenia in the Nordic Countries. PLoS ONE 8(6): e67133. doi:10.1371/journal.pone.0067133
9. Statistics about data elements. [*https://www.health.ny.gov/statistics/sparcs/sysdoc/elements\_837/table\_x12.htm*](https://www.health.ny.gov/statistics/sparcs/sysdoc/elements_837/table_x12.htm)
10. Texas Heart Institute. [*http://www.texasheart.org/HIC/Topics/HSmart/women.cfm*](http://www.texasheart.org/HIC/Topics/HSmart/women.cfm)
11. Margaret Jean Hall, Ph.D.; Shaleah Levant, M.P.H.; and Carol J. DeFrances, Ph.D.(2013) “Trends in Inpatient Hospital Deaths: National Hospital Discharge Survey, 2000–2010” (Issue No. 118 March 2013)
12. Drye, Elizabeth et al. “Comparison of Hospital Risk-Standardized Mortality Rates Using In-Hospital and 30-Day Models: Implications for Hospital Profiling.”*Annals of Internal Medicine* 156.1 Pt 1 (2012): 19–26. *PMC*. Web. 26 Oct. 2015.
13. Krumholz HM, Wang Y, Mattera JA, Wang Y, Han LF, Ingber MJ, et al. An Administrative Claims Model Suitable for Profiling Hospital Performance Based on 30-Day Mortality Rates Among Patients With Heart Failure. Circulation. 2006; 113(13):1693–701. [PubMed: 16549636]
14. <http://www.cdc.gov/heartdisease/>
15. Kurian AK, Cardarelli KM. Racial and ethnic differences in cardiovascular disease risk factors: a systematic review.
16. A.H.E.M. Maasand Y.E.A. Appelman,Gender differences in coronary heart disease