

University of Windsor

Master of Applied Computing

[Advanced Database Topics](https://blackboard.uwindsor.ca/webapps/blackboard/execute/courseMain?course_id=_147620_1) ([COMP 8157](https://blackboard.uwindsor.ca/webapps/blackboard/execute/courseMain?course_id=_147620_1))

**Milestone 2**

**(Till Data models)**

**Instructor:**

Prof. Dr Kalyani Selvarajah

**Group Members:**

Disha Amrish Gajjar (110022319)

Shailja Gupta (110019120)

Sowmiyaa Chakravarthy Vasan (110011686)

Project title: **Data Mining Framework for Healthcare Analytics**

**Sources**: We have chosen Data mining model for our project. We search the journals and found two topics that caught our interest. The links to those articles are: <https://dl.acm.org/doi/abs/10.1145/3176653.3176740>

<https://dl.acm.org/doi/abs/10.1145/2023582.2023587>

**Data Source:** We are using publically available data “ Texas Hospital Inpatient Discharge Public Use Data File” : <https://dshs.texas.gov/thcic/hospitals/Inpatientpudf.shtm>

**Milestone 2**

In this project, we are considering group of patients suffering from HIV or diagnosed with drug/substance abuse and how analysis of healthcare data can help us point to services that can be improved. In our data mining problem, we focus on patient profiling by grouping patients with HIV and drug abuse condition and they would require special needs and care while being treated for the illness they have been admitted for. Due to their underlying condition, these groups could develop certain side-effects or illness physically or mentally that can be addressed if aware of. This way, it improves clinical efficiency as analysis and study would help in providing right treatment based on their condition and side-effects. This in turn cuts down the patient count for each doctor that increases their efficiency too. Further, admin of healthcare organizations can plan their resources based on the study in order to accommodate such patients as quickly as possible and at an affordable price.

From an admin’s view, this project aims to improve healthcare experience for these targeted group (HIV and drug abuse patients) by increasing the special rooms as per demand to provide better care. Also, help in providing pain free services based on their type of illness with pre-existing HIV or drug abuse disorders. Further, as admin it would be a good idea to come up with programs and department that provide counselling for drug abuse patients in order to prepare them towards a healthy living.

Having more information about the patient suffering from HIV and drug abuse order will help the administration of the hospital to improve clinical efficiency. There are different units in hospitals such as Coronary Care Unit, Pediatric Unit, Rehabilitation Unit etc which are recommended to patients depending on their illness. The utmost goal of healthcare industry is to provide care and comfort to the patient while on treatment. On similar lines, as quoted in our data mining problem, patients with HIV or drug abuse condition would require extra care when they come in for other illnesses. Hence, the issue in the healthcare industry to provide better services to specially categorized patients such as HIV or drug abuse patients can be improved by mining into data to predict better services to them. This couldalso help clinics and admins to understand the resource utilization and predict future allocation based on current utilization trend. It allows clinics to be prepared and to provide better care based on the illness of such patients considering their underlying medical conditions.

Hence, among the different pathways that are defined under healthcare analytics, our data mining problem falls under “right care” pathway to beprovided to the patients. Data mining problem also address client’s risk for retention in care failure before client falls out of care.Healthcare industry has always faced this issue where patients under special category or a condition fall out of care due to their underlying condition adding to their illness or disease. Hence, clinical efficiency is lost in such failures of retaining care for HIV or drug abuse patients.

**Data mining process for classification of HIV/drug abuse patients**

The data mining problem is worked upon by following the data mining steps. We selected the data for HIV/drug abuse patients at an earlier stage to have a focused dataset to solve our business problem. The data mining problem is to classify a new patient into one of the specialized units in the hospital that forms our target variable. As we classify patient based on certain parameters, they behave as input variables such as illness, ethnicity, age, type of admission etc. On identifying the required variable for our data mining problem, we selected appropriate data and performed dimension reduction and other visualization techniques to clean, transform and validate the dataset required for classification. Further, we partitioned the data to build a model using training set and to test the accuracy of the model using validation set.

**DATA EXPLORATION**

The THCIC data is cleaned and examined for the following variables that are required for our data mining problem. Based on the identified target variable, required input variables or predicted variables are selected in this part of data mining.

**Predictor Variable**: Predictor variables are mapped to the target variable through an empirical relationship. They can be categorical, continuous or integer. Predictions can be of three types: decisions, rankings and estimates.

1. TYPE\_OF\_ADMISSION (Categorical Variable):

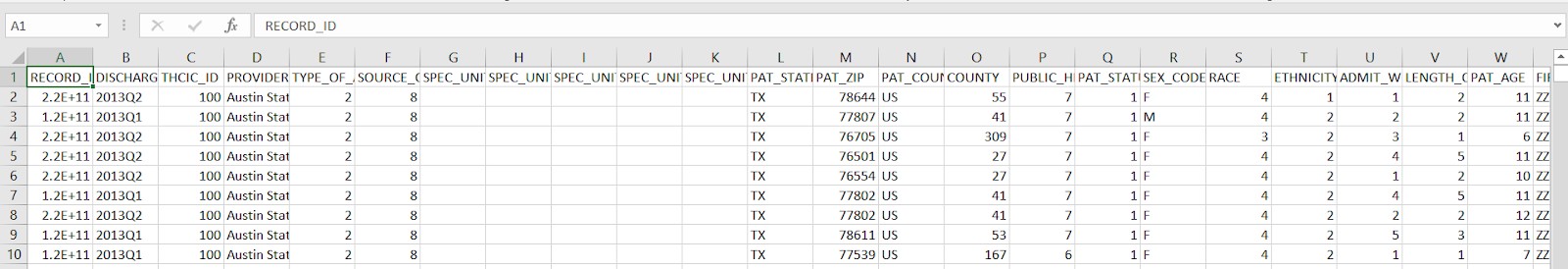
This field is used to indicate the type of admission of the patient. It consists of values such as “Emergency”, “Urgent”, “Elective”, “Newborn”, “Trauma Center”, “Information not available” and “Invalid”. The patient requiring immediate assistance will fall under the emergency or urgent category depending on the situation. This predictor will help us in getting the type of admission information about the patient. Since, we are classifying the patients suffering from HIV and drug abuse it will facilitate in giving us the clarity of the patient’s condition when he/she was admitted.

1. SOURCE\_OF\_ADMISSION (Categorical Variable):  
   There can be various reasons for which the patient can be admitted. Sometimes, they are referred by a clinic, recommended by lawyers in the court or transferred from other hospitals. This field will source number which will specify their source of admission. This predictor will help our analysis of patients suffering from HIV and drug abuse by providing the details of their source. For example, if there are more patients recommended by court this might imply that people suffering from HIV and drug abuse are either in worst helpless condition or the ones which are referred from other hospitals may signify that there is still lack of medications available to cure them. After performing the thorough analysis some conclusion can be made.
2. PAT\_STATUS (Categorical Variable):  
   This field indicates the status of the patient when he/she is about to leave the hospital. For example, this can contain, discharged, expired etc. depending on the situation. It will store numbers which will depict the status of the patient. This predictor will help us know that whether the patient was cured fully when he/she left or there was just little improvement. This can help us track the patients which were not fully cured.
3. RACE (Categorical Variable):  
   This field will store the information about the patient's race. It will include a code referring to a certain race. “1” refers to American Indian/Eskimo/Aleut, “2” refers to Asian or Pacific Islander, “3” refers to Black, “4” refers to White, “5” refers to Other and “`” refers to Invalid. There are certain races which are prone to specific type of diseases. If we come to know about that information, then patients arriving later of same race will be highly prone to that specific disease. Having this information can facilitate in providing cure to people of that race too.
4. ETHNICITY (Categorical Variable):  
   This field includes the ethnicity of the patient. It indicates that whether a patient is Hispanic or not. It includes code signifying that value. “1’ implies Hispanic Origin, “2” implies “Not of Hispanic Origin” and “`” implies Invalid. This predictor will be helpful for knowing the ethnicity of the patient. If there are more people from a certain ethnicity suffering from HIV and drug abuse. Then in future, precautions can be taken, or awareness can be spread to those specific ethnicity people about the HIV and drug abuse in order to avoid it. As hospital/clinic administration, programs can be organized for people around in the nearby areas.
5. LENGTH\_OF\_STAY (Continuous Variable):  
   This field indicates the duration for which the patient was admitted in the hospital. It is calculated by subtracting the day patient entered the hospital from when he/she leaves the hospital. The minimum length is 1 and maximum length is 9999. This predictor will help us know the duration for which the HIV and drug abuse patients stays in the hospital. Having an average number for this can help the hospital/clinic administration to plan accordingly. For example, like it can facilitate in the bed management that tentatively how long the patient will stay in the hospital or clinic.
6. PAT\_AGE (Categorical Variable):  
   This field specifies the age range of the patient. We are using the range of 22 to 26. Mostly, the AIDS and drug abuse is common in the youth. Choosing this predictor and fixing it to a value of 22 to 26 will help our classification.
7. FIRST\_PAYMENT\_SRC (Categorical Variable):  
   This field indicates the source of the payment done by the patient. There are various options a patient can choose to pay like he can opt for Insurance, Medicaid etc. This field will include codes indicating the option which patient has opted for. This predictor will help us know the mode of payment chosen by the patient. This can facilitate in knowing the financial status of the most patients.
8. TYPE\_OF\_BILL (Categorical Variable):  
   This field includes the information about the claim data. It includes of three digits. The first indicates the type of facility, second shows the type of care and the third shows the sequence of the claim. This predictor will help the hospital/clinic administration know the details about the kind of facility used by the patients so that they can plan in the future to accordingly manage for that facility. This also gives information about the type of care undertaken by patients suffering from HIV and drug abuse. this can help the administration to plan for that care unit. Maybe they can plan inventory for that unit.
9. TOTAL\_CHARGES (Continuous Variable):  
   This includes the total sum the patient must pay for the services he/she availed at the hospital. This predictor will help the hospital/clinic administration to know how much amount of money is charged by the patients suffering from HIV and drug abuse. This will help the administration for financial statements.
10. PRINC\_DIAG\_CODE (Continuous Variable):  
    This code indicates the patient’s principal diagnosis when he/she arrived at the hospital. This predictor will facilitate us in having information about the patient’s principal disease.
11. RISK\_MORTALITY (Categorical Variable):  
    This field indicates that what are the chances of a patient to die. This value has been assigned considering All Patient Refined (APR) Diagnosis Related Group (DRG) from the 3M APR-DRG Grouper. This predictor will help us knowing that what would be the dying chance of the patient suffering from HIV and drug abuse.
12. ILLNESS\_SEVERITY (Categorical Variable):  
    This field indicates that what is the level of severity of the illness of the patient. It includes four code:”1” indicating Minor “2” indicating Moderate “3” indicating Major and “4” indicating Extreme. This value has been assigned considering All Patient Refined (APR) Diagnosis Related Group (DRG) from the 3M APR-DRG Grouper. This predictor will help us know the suffering level of the patients having HIV and drug abuse. Classifying patients on this basis can help planning the medications.

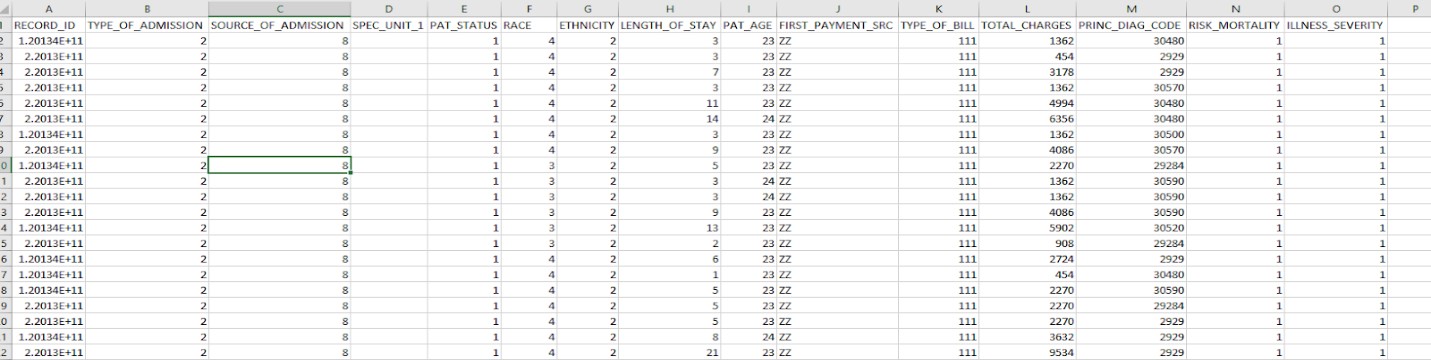
**Target Variable**: The target variable is the one for which we need the output or classification hence in this case our target variable is SPEC\_UNIT\_1.

1. SPEC\_UNIT\_1 (Categorical Variable):  
   This field indicates the specialty units in which most days during stay occurred based on number of days by Type of Bill or Revenue Code. In order by number of days in the unit. It includes codes such as “C” which means Coronary Care Unit, “P” Pediatric Unit etc. With the analysis of this variable will help us identify the speciality unit which the patient was in.

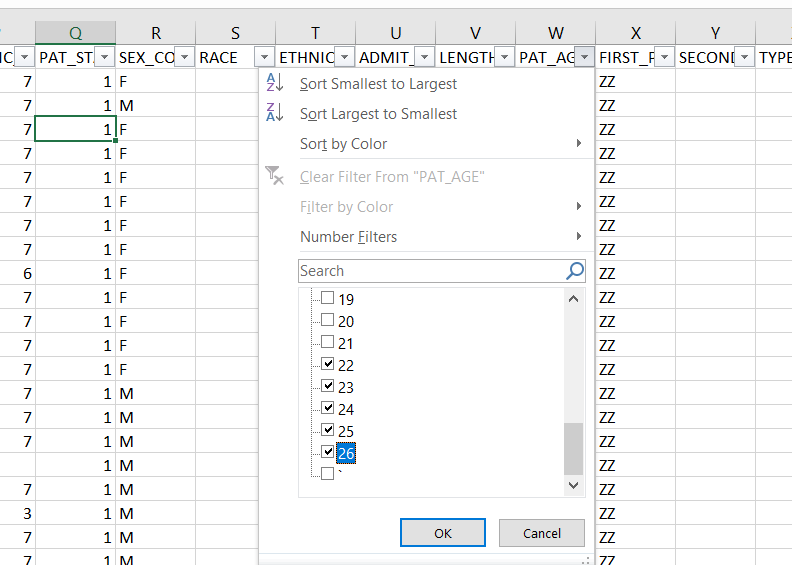
**DATA IDENTIFICATION**

We have identified 14 variables that can be used as predictors for our target variable. These predictors can be used to create models that will help us in our predictions. Once we analyse the given data, we need to first identify the data that is useful for us. Hence, we need to reduce our data from 194 columns to 15 columns including the target variable. This reduction in the number of columns is done as the other columns are of no value while classifying the Specialty units, they just increase the volume of data for our business case.

Thus, we select those 15 variables and create another sheet of data that does not have any unnecessary columns.

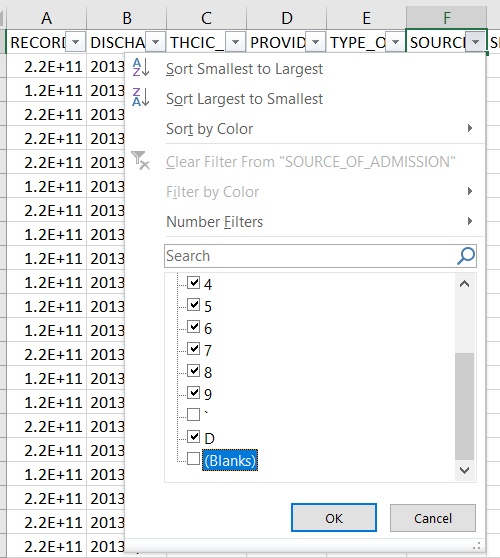


Once we are done with columns, we need to work on reducing the number of rows by identifying the data that we need. Here, we need the data for patients who are ‘HIV and drug/alcohol use patients. As we analyse the column PAT\_AGE we see that the values 22-26 for that variable specifies these patients. Hence, using this criterion we filter out the data and utilize it for further tasks. Hence after this step our data reduces from 719,371 rows to 52,146 rows. This is done as our classification is based on a subset of patients having a specific condition associated with them. If we include other data, it can lead to wrong predictions as well.



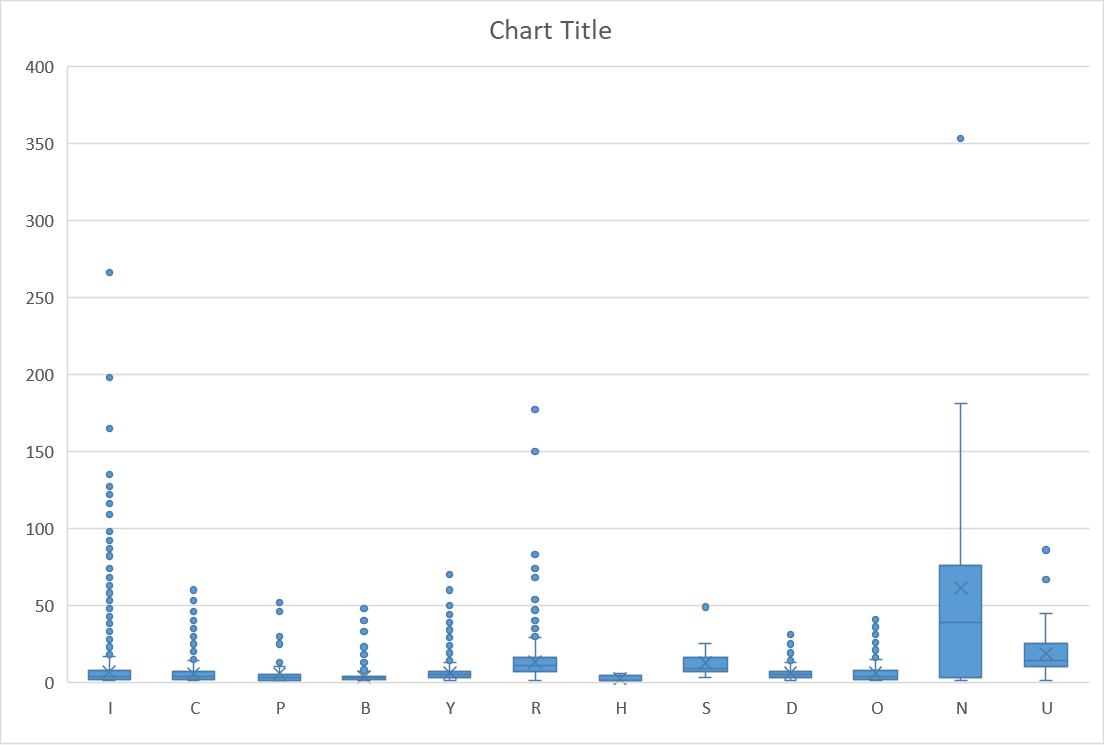
**DATA CLEANING**

In this step we must identify the anomalies present in our data. For this we need to check each variable that is involved in our analysis. Hence, we observe that there are not many trash values in each column, but there are blanks and the symbol [‘] which need to be filtered out and removed.

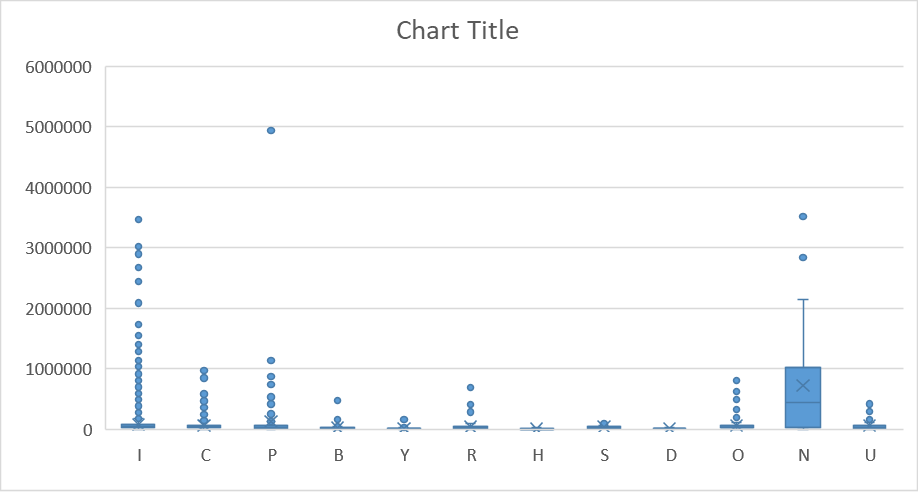


This step is needed because having blanks and symbols will cause hindrance in our classification as they add errors to our analysis. Also, null values specify nothing which means assuming them as zero also is not correct. After we are done with this step for all variables, our number of rows reduce from 52,146 rows to 32529 rows.

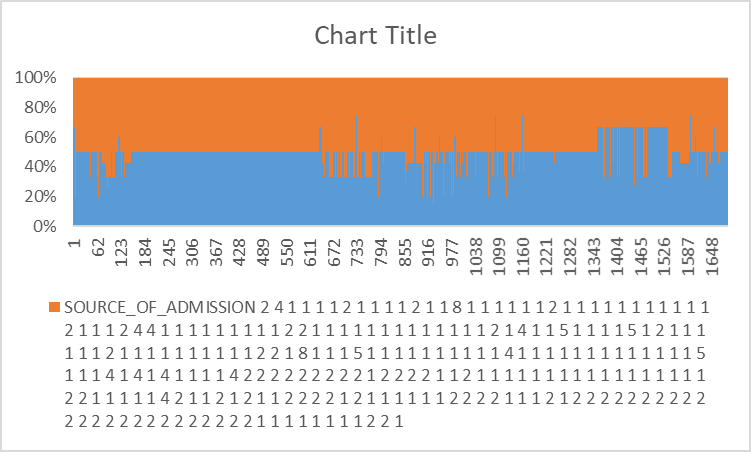
Now we try to understand the relationship between predictor and target variables.



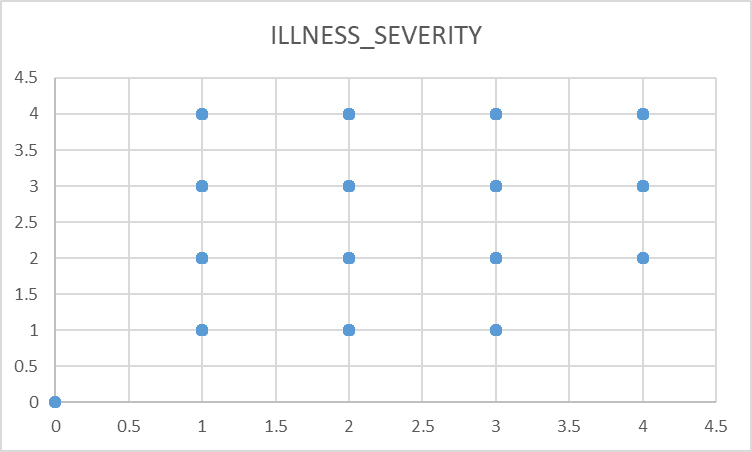
In the above graph we check the relation between predictor LENGTH\_OF\_STAY and SPEC\_UNIT. As we see the box plot, we can identify the relation between the type of specialty unit and the length of stay. The values are distributed around the median quite uniformly for such a large data set. Even the number of outliers is very less. This shows that using length of stay as a predictor will enhance our analysis. Usage of such variables help us reduce average error in case of our predictions.



In this graph, we create a box plot to understand the relation between SPEC\_UNIT and TOTAL\_CHARGES. As we know, that in general cases our charge or cost of treatment depends on the disease. Also, the specialty unit is specific for specific type of disease. Hence, the cost and specialty unit are directly linked to each other. This assumption of ours is proven true by the graph above where the distribution box for the cost is similar for similar specialty units. Even in this case the number of outliers is less.

Apart from this let us understand the relation between multiple predictors as well.

This is a stacked bar plot for variables SOURCE\_OF\_ADMISSION and TYPE\_OF\_ADMISSION. Similar columns tend to capture similar information which can lead to multi collinearity. This must be avoided for which we can use just one column instead of both.



It is the similar case with variables ILLNESS\_SEVERITY and RISK\_MORTALITY. They capture similar information hence we can eliminate either of these.

Apart from this we have a lot of categorical variables that need to be transformed to perform further analysis for upcoming reports.

**CREATING A MODEL SET**

In the case of supervised algorithm, it means that the algorithm learns through the data and then apply the pattern it understood to the data for which we need to classify our outcomes. The data that we have for this contains 32,529 rows in the table. This data can be split into three parts where one part can be used by the algorithm to learn the pattern of the data, the second part can be used to validate this pattern that the algorithm learned and check if it gives us appropriate results that is with error rate which is tolerable for our use and the third part of the data can be used to test the algorithm. This third dataset is helpful in the case where there are multiple algorithms run and we need to select the best possible model and then run it on this test data set. When an algorithm learns from a training data set, and it applies this model to validation data set it might seem accurate, but this might not necessarily mean that it is correct, hence to learn the characteristic of a data set wholly we need a separate test data. Hence, the three parts of our data will be training data, validation data and test data.

Here in this case, we have divided our data into three parts where 40% of the data is training data, 30% of data is validation data and the last part of 30% of data is test data set on which the model is applied to verify if the algorithm is working accurately. Here what happens is that the algorithm will use the training data set and understand how the target variable varies in accordance with changes in predictor variables. Then it will apply this pattern on our validation data set and check how correctly it is able to classify the records, we can check error percentage for this. After this it again applies this pattern to a separate test data to confirm if the model is working properly.

**Work Continous**….. We are trying to access a prediction tool or either python/R to analyze the data and create the data models and for regression.