**BUSINESS INTELLIGENCE AND APPLICATIONS PROJECT**

**ANALYSIS OF HOSPITAL DATA ON HANDLING DIABETES AND IMPROVING PATIENT SATISFACTION**

**Details:**

**Name: Mansi Goyal Sachi Angle**

**Section: C C**

**Registration number: 140905232 140905542**

Introduction

Diabetes can strike anyone, from any walk of life.  
And it does – in numbers that are dramatically increasing. In the last decade, the cases of people living with diabetes jumped almost 50 percent – to more than 29 million Americans.   
Worldwide, it afflicts more than 380 million people. And the World Health Organization estimates that by 2030, that number of people living with diabetes will more than double.   
Living with diabetes places an enormous emotional, physical and financial burden on the entire family. Annually, diabetes costs the American public more than $245 billion.

Considering how prevalent Diabetes is, hospitals over the world have a large number of patients admitted yearly for the treatment of it. Successful treatment of the disease leads to a larger and happier client base. Apart from medical factors, successful treatment depends on many administrative factors too, like availability of resources and the number of hours during which the hospital staff are available for check ups and treatment. Analyzing data also gives insights on which type of diabetes, out of four, should the hospital be more prepared for. If one of the four types is more common than the others, the hospital can control its expenditures by stocking an excess of medicines for that type only. Thus, analysing data can lead to better management of diabetes by the hospital and an improved overall experience for the patient.

Significance of Project:

One of the major issues faced by the Healthcare industry today is the lack of tools which can be used to measure the productivity of the employees and the extent to which patients are benefitting from the services provided by them.

The project goal is to use Business intelligence (BI), which is the use of computing technologies for the identification, discovery and analysis of business data - like sales revenue, products, costs and incomes, to overcome the above mentioned obstacles. As a result, it will enhance the overall performance which in turn can provide high quality services at lower costs.

Hospitals want to improve on two base factors - Patient Satisfaction Index, and The extent to which the disease was handled and contained by them. They can also stock medicines in accordance with data which can tell them which disease is more likely to occur across various countries.

There are various possible factors which can contribute to Patient Satisfaction Index and the Percentage Management of Diseases by the hospital - like Consulting Hours, Knowledge of the Hospital Staff, etc. But to improve on these indices, a hospital should know exactly which features should be changed and improved. Some features don’t contribute as much as others. Improving on the wrong features would not have as big an effect on Patient Satisfaction and Management of Diseases as opposed to working on the important features, as that would be time, money and resource consuming. Our project analyses the data to find out which features are important to improve the two indices.

Hospitals can control their expenditure by not stocking an excess of medicines which are less likely of being used. Analysis of the data also enable us to to see which disease is more likely to occur in which country.

Improving on the Patient Satisfaction Index, and the Percentage Management of Diseases, and knowing which disease is more likely to occur lets the hospital be prepared, with medicine, staff and other facilities which leads to them successfully handling and treating the patients. Subsequently, this leads to an increase in the size of their client base.

Software Used

The data was extracted, integrated, and cleaned using python libraries - Numpy and Pandas - on Anaconda, an open Data Science platform.

Machine Learning techniques were used to learn about the data using sklearn, a Python library, on Anaconda.

Power BI was used to plot the graphs.

Methodology

F:\flowchart.png

DATA INTEGRATION

* For Percentage Management of Data, one database was collected for every year, from 2012 till 2015. Any changes in the format of stored data that varied between the years had to be made uniform while combining the datasets.
* Each year had a separate database for the Patient Satisfaction Index, this was integrated into the main database, using the common key ID.

DATA CLEANING

* While cleaning the data, errors in inputting the data were checked for. Different people use different symbols to indicate NULL values and this needs to be made uniform.
* Redundant and Incomplete entries were removed.
* Irregularity in the data was checked for and corrected. The data entered may have been entered in more than one form - A countries name may have been written with two different spelling across the database, or a percentage may have been mentioned as a multiple of 1 in a few rows, and 100 in the remaining. We need regularity and uniformity in our database, all the irregularities had to be corrected.
* The numerical data was present in string format, with additional symbols like ‘%’. The data had to be converted into float numbers, by writing python scripts to remove the character symbols, and to convert the data to a usable form.

MACHINE LEARNING

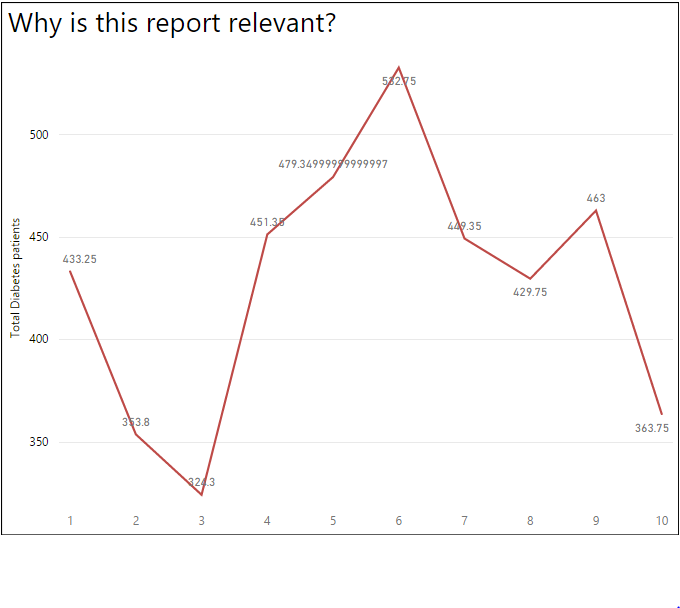
* Machine Learning was performed on the dataset to obtain insights on the features.
* Decision Tree Classifiers were used. Decision tree learning uses a decision tree as a predictive model which maps observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modelling approaches used in statistics, data mining and machine learning.
* “Feature\_importances\_” is an attribute of the decision tree classifier. It is an array of values, each corresponding to an input feature. The higher the value, the more important the feature for that Index. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature.
* For every output index like “Overall Satisfaction”, “Percentage Management of Diabetes”, “Percentage Medical Errors”, the input data was fed into the Machine Learning Model.
* The input comprised of attributes like “Staff Knowledge”, “Foot Disease Assessment done within 24 hours”, “Number of Hospital Beds”, and so on.
* The Machine Learning Model learnt from this data, and a “Feature\_importances\_” array was obtained for each of the output indices. These arrays indicate how important each input feature is for the corresponding output feature.
* Using this, hospitals will know how to improve each output feature - they will know which input feature needs to be improved to obtain the corresponding increase in the output feature.

VISUALIZATION OF DATA

* A number of graphs were plotted to capture and analyze the different trends in the data using Power BI.
* The different type of graphs used were : Histograms, Bar Graphs, Pie Charts, Line Graph etc.
* The results obtained from the Machine Learning section were used to plot these graphs. They were extracted and stored in separate Excel Sheets.
* These Excel sheets were then imported into Power BI and used for showing the different patterns.
* The next section examines each graph in detail.

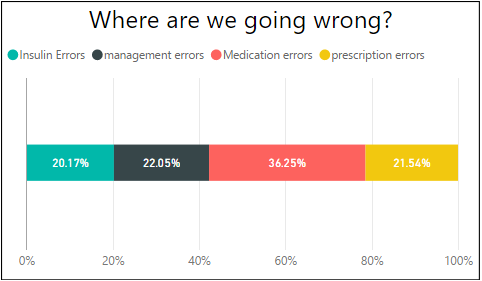
**Results and Graphs**

A number of graphs were plotted to capture and analyze the different trends in the data. Relevant and actionable insights were provided to the hospital administration by the use of following graphs.

1. The number of diabetic patients per 20 hospitals was used to plot the graph. It helps in estimating the number of expected patients suffering from diabetes. 

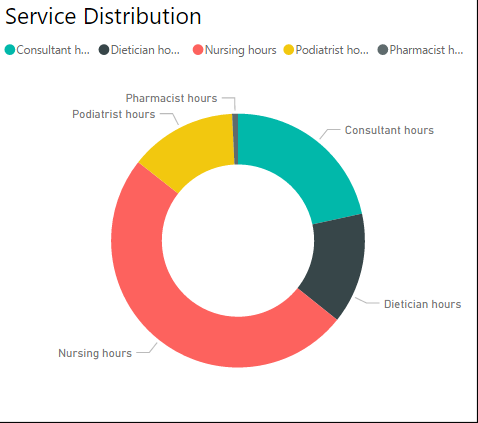
Graph 1

2. Average percentage of different types of errors committed by the hospitals was used to plot the graph. It describes the distribution of errors across the different categories.

. 

Graph 2

3 The average number of hours per service was used to plot the graph. It helps in determining the services with maximum and minimum demand by patients/

. 

Graph 3

4. Graphs drawn to analyze the effectiveness of treatment of diabetic patients:

The main factors which decide the quality of diabetes treatment include:

a) Percentage management of diabetes

b) Able to take control of diabetes care

c) Percentage renal replacement therapy

d) Medication errors

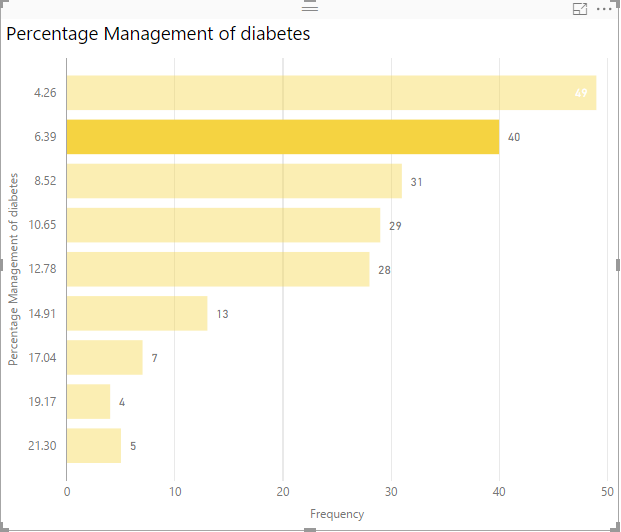
e) Prescription errors

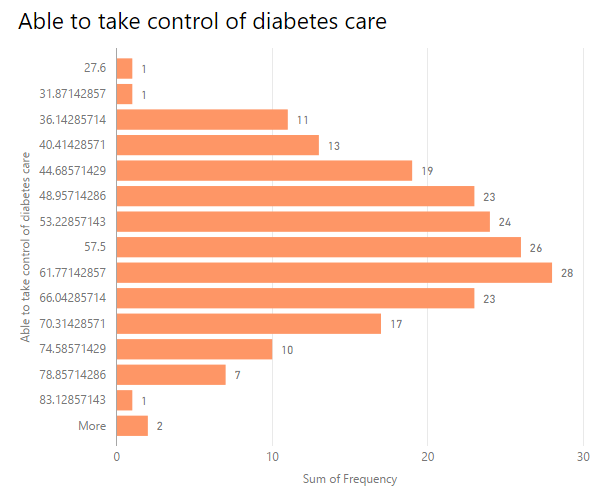
f) Management errors

g) Insulin errors

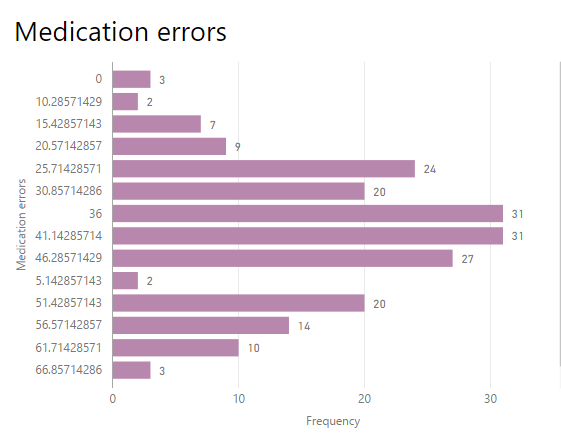
Graphs 4.1 to 4.4 are histograms which tell about the distribution of data in various bins.

Similar histograms were drawn for the other factors as well.

Graph 4.1 

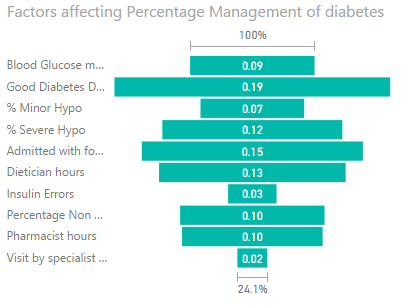


Graph 4.2

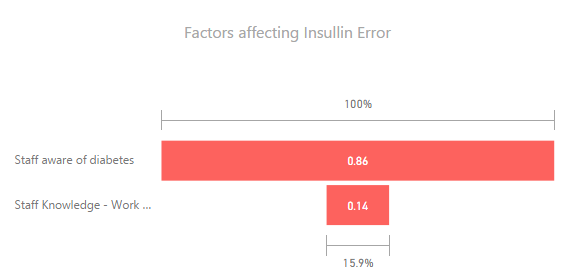


Graph 4.3

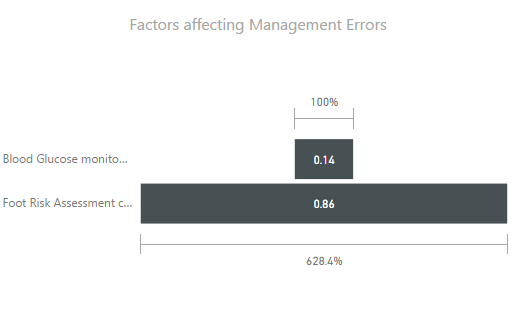
Each factor is further supported by several subfactors. The graphs given below analyze the contributions of different subfactors in strengthening the factor:



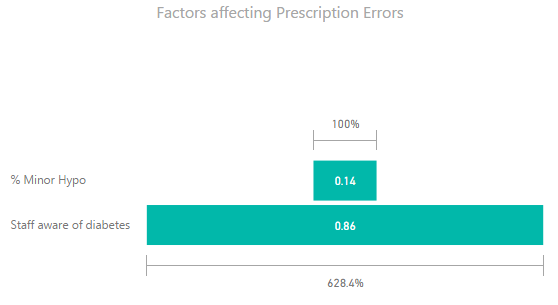
Graph 4.4



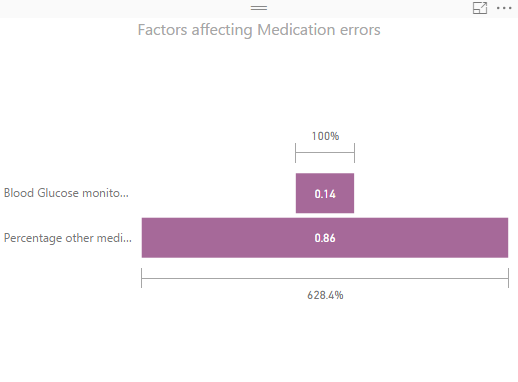
Graph 4.5



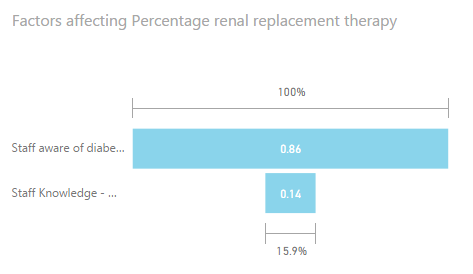
Graph 4.6



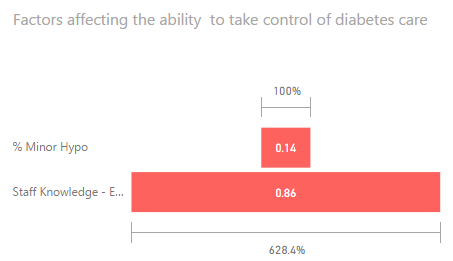
Graph 4.7



Graph 4.8



Graph 4.9



5. Graphs drawn to determine the patient satisfaction:

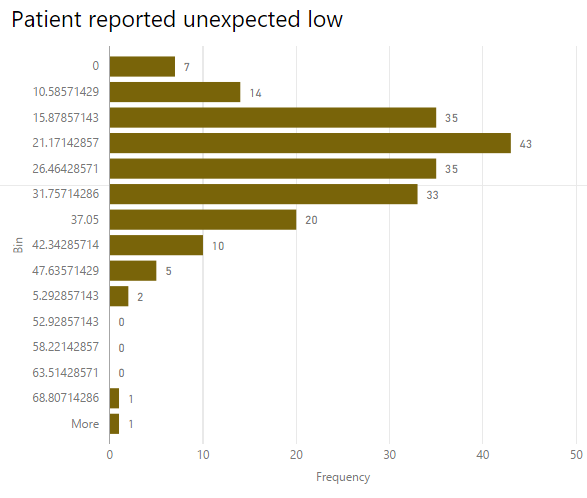
The main factors which decide the level of satisfaction experienced by the patients include:

a) Overall Satisfaction

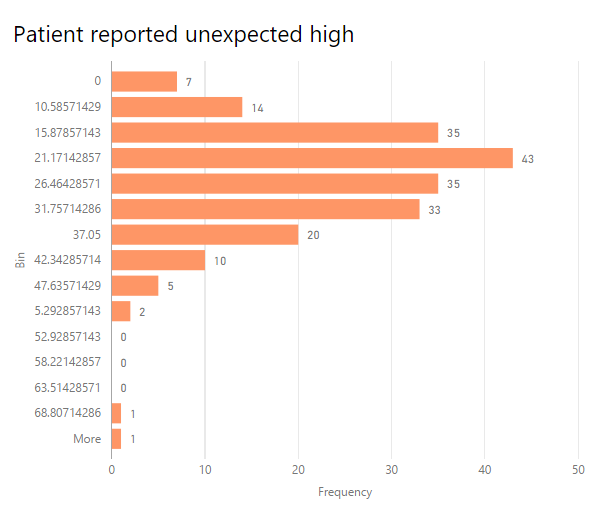
b) Patient reported unexpected high

c) Patient reported unexpected low

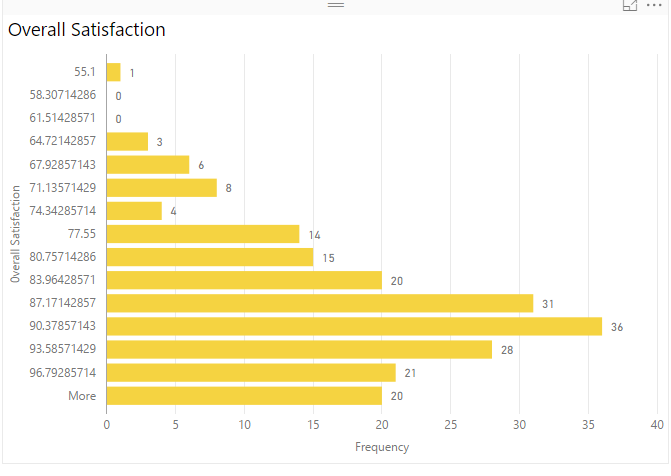
Graphs 5.1 to 5.3 are histograms which tell about the distribution of data in various bins.



Graph 5.1



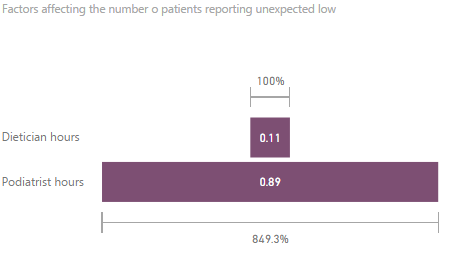
Graph 5.2



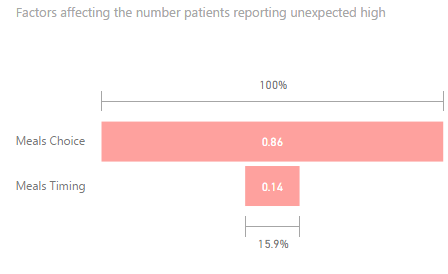
Graph 5.3

Each factor is further supported by several subfactors. The graphs given below analyze the contributions of different subfactors in strengthening the factor:

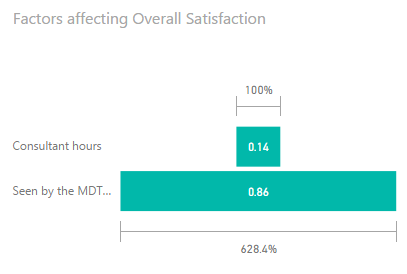
Graph 5.4



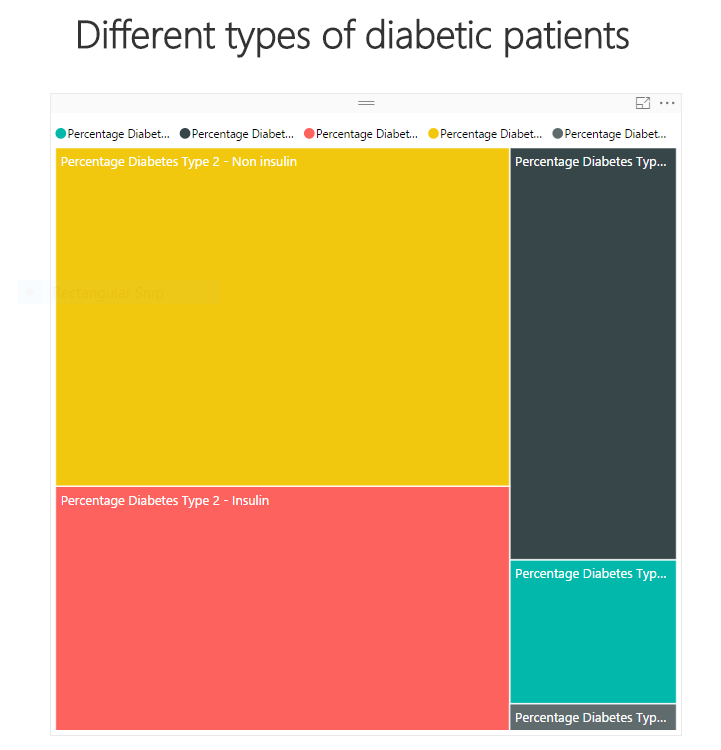
Graph 5.5



Graph 5.6



6. Graph 6 describes the distribution of patients across different categories of diabetes:



---------------------------------------------------------------------------------------------------------------------

Conclusion

1. The high numbers in Graph 1 indicate that diabetes is a fairly common disease therefore hospitals should be better prepared to deal with it.Hence reinforcing the relevance of the report.
2. Graph 2 helps the hospital in determining where the largest number of errors are being committed. In the graph it can be seen maximum number of errors occur in the medication sector while insulin errors occur the least. Hence the hospital needs to focus more on the medication sector to improve its service.
3. Graph 3 indicates the demand for different services provided by the hospital. The distribution helps the hospital in determining which service is used the most, hence requires allocation of more manpower and budget and vice-versa. For instance in the graph time spent on nursing is maximum and least on pharmacist.
4. The graphs 4.1 to 4.3 are histograms which were used to determine the threshold value for a hospital to be in the top 10% or bottom 90% in that particular category.
5. Graphs 4.4 to 4.11 help in determining how a factor is affected by different subfactors. Thus the hospital can work in those particular domains to get better results and enhance it's services. For instance the following subfactors were deemed most important for a particular factor for improving the quality of treatment provided to diabetic patients:

**a) Percentage management of diabetes :** Good diabetes days

**b**) **Able to take control of diabetes care:** Staff Knowledge – Emotional Support

**c) Percentage renal replacement therapy:** Staff aware of diabetes

**d) Medication errors:**  Percentage other medical

**e) Prescription errors:** Staff aware of diabetes

**f) Management errors:** Foot risk assessment completed during the hospital stay

**g) Insulin errors:** Staff aware of diabetes

In the above data the subfactor “Staff aware of diabetes” is repeated thrice. Hence by educating the staff more about diabetes the quality of the treatment can be enhanced exponentially.

1. The graphs 5.1 to 5.3 are histograms which were used to determine the threshold value for a hospital to be in the top 10% or bottom 90% in that particular category.
2. Graphs 5.4 to 5.6 help in determining how a factor is affected by different subfactors. Thus the hospital can work in those particular domains to get better results and enhance it's services. For instance the following subfactors were deemed most important for a particular factor for improving the overall satisfaction level of diabetic patients:

**a) Overall Satisfaction : Seen** by the MDT within 24 hours

**b**) **Patient reported unexpected high:** Meal of choice

**c) Patient reported unexpected low:** Podiatrist hours

The above results align with the general conceptions. For instance a person is in one of his worst moods when he is hungry and doesn’t get food of his choice. Also the deciding factor for the patient is whether he is seen by the doctor within 24 hours of getting admitted to the hospital. In severe cases of diabetes the nerves in the feet get damaged leading to a lot of discomfort. Hence if a patient does not get enough consultation time with the podiatrist he is left highly dissatisfied.

1. Graph 6 helps the hospital administration in identifying the different type of diabetic patients coming to the hospital for treatment. Hence to maximize resource utilization and profits the hospital should focus more on the category of diabetes with the maximum incoming patients which is Diabetes Type-2 Non Insulin.

References

1. Data obtained from : <http://www.digital.nhs.uk/>
2. <http://scikit-learn.org/>
3. [www.wikipedia.com/businessintelligence](http://www.wikipedia.com/businessintelligence)
4. Fundamentals of Business Analytics – Seema Acharya