31/10/2019
University of Bradford

Analysis of diabetic data

USING VISUALISATION TECHNIQUES

ISMAIL BHAMJEE/ 16010047

Contents

Introduction 2
Aim 2
Background 2
Dataset 3
Methodology 3
Question 3
Method 3
Visualisation 4
Graph 1 4
Graph 2 4
Graph 3 5
Graph 4 5
Evaluation 6
Advantages and disadvantages 6
Limitations and alternatives 7
Further work 7
Alternative software 7
Conclusion 7
References 8

Introduction

Aim

To produce a piece of analysis along with a visualisation and write a research report for a chosen method of visualisation on a chosen dataset.

Background

Diabetes is a disease in which the body unable to process the blood glucose level. It Is naturally done by a hormone called insulin made in the pancreas, when the insulin levels are low or the body misuses insulin the cells are unable to use the glucose (1).

As of 2014 they are approximately 422 million people with diabetes (2). In the USA alone they are over 29 million as of 2015. Furthermore, it is the 5th most common cause of death in the world (3). Other factors such life style and weigh also have a role in preventing diabetes.

Lastly, they're two types of diabetes the first being the body does not produce enough insulin and the second which is the insulin is misused. The second is what this paper will focus on. As they're a plethora of drugs and procedures which may help (1).

Dataset

The diabetic dataset which was chosen has over 100,000 records which span 50 columns which was from a collection of 9 years from approximately 130 hospitals. This provides a vast amount of data which can be used for mining and data analysis.

Furthermore they're 25 columns per record which contain various medication which could have been possibly used on the patient. This also provides a large data set to analyse. Next, the data contains other factors such as age, race and admission types which may lead to further extrapolation of information. This brings us to the question of which will guide the interpretation of this data into meaningful information. (4)

Methodology

Question

The question which the basis of the research will be is “does the number of medications affect time in hospital”

Method

Analysis will be done via MATLAB 3D and 2D simulations. This is to ensure that the data in question can be compared via a glance. This is due to the dataset being nothing, but a series of numbers and the naked eye cannot understand them in the way they are. But with the addition of visual aids the relationship of the data can be illustrated cleanly and clearly.

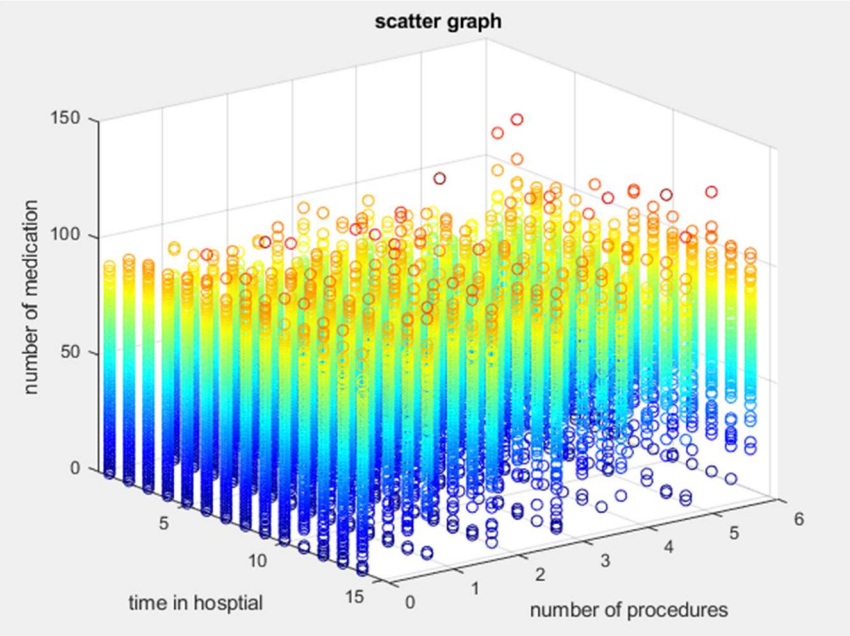
Next trends and correlations can be found in the data which will help with data analysis, this is simple in MATLAB as colour bars can be applied to see the density of an item.

MATLAB is a software which follows a strict set of coding conventions to allow the data to be displayed. It is highly accurate in terms of acknowledging all the data along with any number precision associated with the task.

From the data base only 3 columns will be used for visualisation this is due to the nature of the question. The question is simple which means that the full extent of the data base will not be used.
Lastly during testing, the 3 data types will be placed against each other to extract meaningful information. As opposed to only one method which refers to each column as being a part of an XYZ 3D graph.

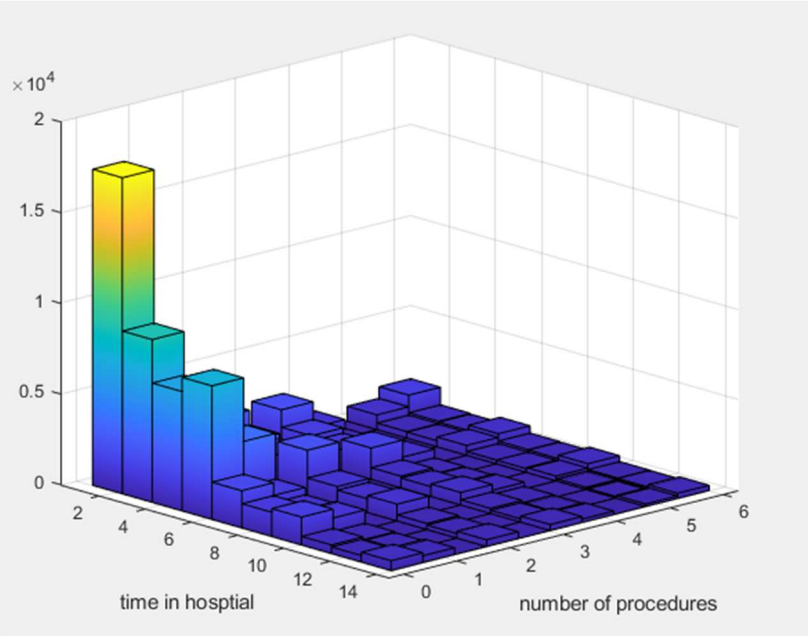
Visualisation

Graph 1



This scatter graph shows the relation between 3 columns in the database. The X axis column is time is hospital, the y axis Is the number of medication and the z axis is the number of procedures. From this graph we can see that the number of medications does not exceed 90 in the whole dataset. Furthermore, we can see that the greater time spend in the hospital means that less procedures were used. On the contrary it shows that less time in the hospital would cause more procedures.

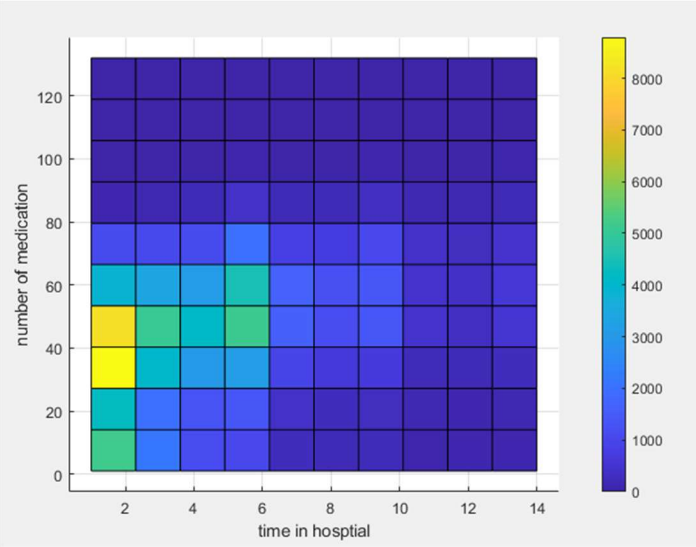
Graph 2



The second graph shows the time in hospital relative to the number of procedures.in this we can see that they're more people who spend less days in hospital and the bulk of those

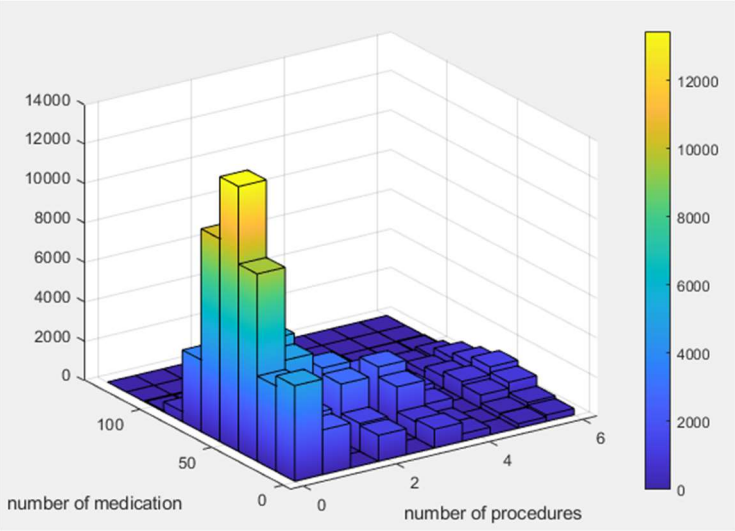
people had less than 2 procedures. The is shown by the 3D histogram. however, this model does not allow for a corrects 2D conversion as the nodes behind the larger pillars are not visible

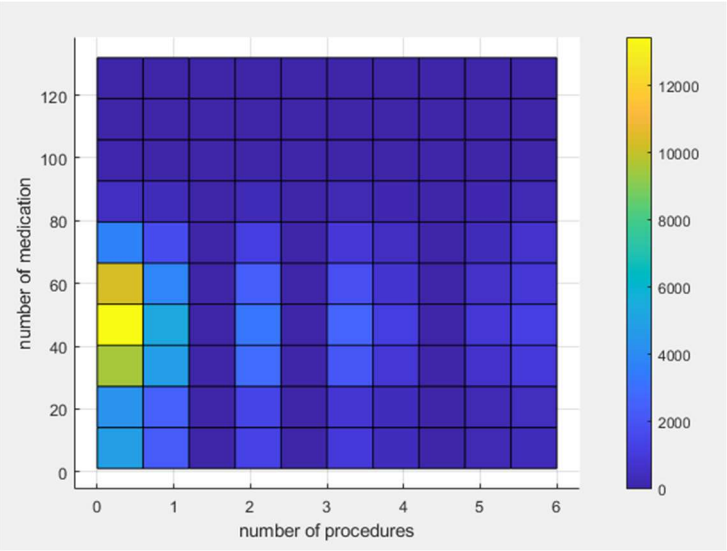
Graph 3



The above 2D diagram shows the relationship between time in hospital and number of medications. From this we can see people who received the most medication were people who stayed in hospital between 1-6 days. The catchment area was between 30 and 70. From this the people who stayed 2 or less days had the majority of medications between 30 and 50. As marked by the colour bar, which is a welcomed visual aid

Graph 4





The following plots show the relationship between number of procedures and number of medications. This was done using a 3D graph and a 2D graph. This was to demonstrate the features used in data modelling that the user can move the model around to see what it represents. But this 3d image cannot be projected in a document so a 2d alternative had to be done.

The Graphs show that the number of procedures between 0 and 1 had the highest number patients use medications between 30 and 80 times. Moreover, this graph also shows that every 2 procedures the number of people who take medication increasers this is shown by the changes in colour.

Evaluation

The use of graphs and models proved useful for interpreting the data for visual analysis. This is shown by the conclusions and discoveries made.

Advantages and disadvantages

Advantages to the scatter plot was to see where each of the individual records line up in relation to each other. This gave a visual image at first glance to see what was going on with the data which the raw excel file could not provide. However, disadvantages to this were that the graph was too densely populated and had to rotated to see clearly. This made it nigh impossible to get the full graph on a 2D plain.

Furthermore, an In-depth break down of the columns could have helped expand the information which was acquired from this project. Due to smaller result numbers providing more precise readings as well as more information to determine the outcome of the question proposed.

Limitations and alternatives

The work was hindered by some Limitations in the type of graphs to use, initially a 3D mesh was considered to map the data but this was later scrapped due to errors in generating the visual. But after considering the density of the data a histogram would be the most appropriate visual as it can represent all the data correctly. With distinct margins between numerical changers being visible.

Furthermore, MATLAB allowed the data to be transformed internally unlike WEKA which was an additional bonus if damaged data was to be removed.

Further work

An omission to the research was to break down the columns using other rows to determine what other factors affected them. For example, the weight of a patient. This would have been done by clustering weight categories available in the data set already.

This was never done as the task would be too large at hand and would pose and answer other question which weren't relevant.

An alternative approach would be to use classifiers to train the data and to visualise the output. Weka allows for data training, but its visualisation methods aren't as rich as MATLAB's. However, the use of classifiers in MATLAB would have been more difficult as they would have to be written manually.

Alternative software

MATLAB wasn’t the only visualisation tool that was considered. Microsoft excel and WEKA were also considered. But after a small amount of testing, a lack of control and preciseness of the graphs were shown to be apparent. So that software was quickly abandoned, this is further demonstrated on the WEKA website (5).

Conclusion

To conclude the use of medication can affect the time a patent has to stay in hospital, as shown by the above research. Time in the hospital Is bad due to many factors such as financial and phycological. This papers aim was to find out if medication and procedures had a correlation to time in hospital as well as to find logical associations so that future patients wouldn’t need hospitalisation, or have their times reduced
Next with the given data analysis the question proposed was answered and the discovery was made that medication and procedures which are done in the initial parts of hospitalisation can prevent further need to hospitalise. This is proven by most people who didn’t return to hospital as shown by the density of blocks in the afore mentioned graphs.

References

(1)

National Institute of Diabetes and Digestive and Kidney Diseases. (2019). *What is Diabetes? | NIDDK*. [online] Available at: https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes [Accessed 19 Oct. 2019].

(2)

Who.int. (2019). *Diabetes*. [online] Available at: https://www.who.int/news-room/fact-sheets/detail/diabetes [Accessed 20 Oct. 2019].

(3)

Diabetes. (2019). *Since 1996, the number of people with diabetes in the UK has risen from 1.4 million to 3.5 million. Diabetes prevalence is estimated to rise to 5 million by 2025.*. [online] Available at: https://www.diabetes.co.uk/diabetes-prevalence.html [Accessed 14 Oct. 2019].

(4)

Archive.ics.uci.edu. (2019). *UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set*. [online] Available at: https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008 [Accessed 12 Oct. 2019].

(extension)

Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.

(5)

Brownlee, J. (2019). *How to Better Understand Your Machine Learning Data in Weka*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/better-understand-machine-learning-data-weka/ [Accessed 15 Oct. 2019].