Data Mining Presentation for a Hospital

# Introduction \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

To fully grasp the concept of Data Mining, first you need to be aware of how much data we actually create in today’s world.

“Every two days now we create as much information as we did from the dawn of civilization up until 2003” (Siegler, 2010)

That sounds like an overwhelming amount of data. After finding ingenious ways of storing this much data, we needed to discover ways of analysing vast amounts of data. Up until recently, only 5-10% of generated data ended up being analysed (Mark Elshaw,2016). The concept of data mining deals with the data that never got any attention before.

“Data mining is the extraction of implicit, previously unknown, and potentially useful information from data” ( Witten, I.H. and Frank, E, 2005). We create computer models that sift through vast amounts of data organized in data sets, looking for regularities and patterns.

Data mining can play an enormous role in predictions, economical patterns, business development and decision making.

# How it applies in Health Care\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

According to [this NCPA article](http://www.ncpa.org/sub/dpd/index.php?Article_ID=23148), An estimated 10 % - 20% of cases are misdiagnosed, which exceeds drug errors and surgery on the wrong patient or body part, both of which receive considerably more attention. Further on, 28% of 583 diagnosis mistakes reported anonymously by doctors have been classified as life threatening, or had resulted in death or permanent disability.

Diagnosis is a vital part of patient care. The process of diagnosis also follows a pattern. You take the symptoms of the patient, its characteristics and you make a prediction based on that, then you run tests to see if the prediction is correct. Tests usually take a long time, time which can be vital in life threatening cases. If the tests fail, that means that whatever you predicted was wrong, and you have to try another prediction, and repeat the cycle until you get it right.

Even something as ugly as a disease follows a pattern. When you think about a disease, you think about its symptoms and you think about the characteristics of people that make them vulnerable to it. These classifications are made using patient records who have been successfully diagnosed and proved to have a condition.

Data mining can automate the process of diagnosis. Vast amounts of past patient records can be fed into data set. The data mining process will create a model based on that information, which will automatically analyse the data whenever you ask the data set a question. Not only it will decrease the time of diagnosis significantly, it will do it with precision. Data mining can play a vital role in increasing the percentage of successful diagnosis and decreasing the time it takes, implicitly offering your patients a quicker reaction to their symptoms.

# Data mining lifecycle\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The first thing that would happen in this process must be the problem definition. We have to know what we are trying to do, that will determine our plan and what data we need. If what we want to do is predict if a patient might have diabetes, then we will be interested to acquire a lot of data related to past patients who have been tested for diabetes. This is the first step and defines the plan for the process.

The next thing that would happen in this process would be acquiring a data set. We cannot do anything without data! Data may come in various different forms, that is why the next step is vital to the successful creation of a data mining model.

The data has to be pre-processed. There are several factors we have to take into account. First, we have to consider if the dataset is big enough to cover all outcomes of our problem. If my dataset contains 90% records with people that tested positive and only 10% of people that tested negative, then our model will not be able to clearly distinguish between the two. Also, the more records we have, the more precise the model will be. We have to make sure there is a balance between the two, to increase the prediction efficiency.

We have to determine our primary inputs, that is, what data is relevant to our model. For example, if we try to decide if a person is likely to have diabetes, or not, we will definitely be interested if close family members of that patient has diabetes (*Genetics and diabetes,* 2016). Then we have to eliminate redundant data, that is data that is not relevant to our question. In the case of diabetes, the exact address of the patient might not be of interest, as we are simply trying to predict whether s/he has diabetes, which does not relate to the person’s geological home address.

After the process of eliminating redundant data, we have to normalize our relevant data. Normalisation means scanning for errors in the data, or for values that are off the charts and might complicate our mathematical model. For each numeric set of values, there is a mean and a standard deviation. Numeric values that are not within these parameters have to be normalized to improve the consistency of the model.

We have to check the data for missing values or other inconsistencies. This is important, otherwise the model will not take in consideration the data that has columns missing, or that conflict with other data (same ID).

At this stage, we should be ready to take our data mining process a step further. We have solved the input pre-processing, and we are now concerned with the output. First, we have to choose a tool to perform Data Mining. Many tools are available in the market, most popular being RapidMiner and WEKA. In this case, we are interested in comparing a patient’s record to past patients and predict if the patient has diabetes or not, therefore, we are creating a predictive model that performs a classification. We create an initial model, and then we evaluate it. During the evaluation, we split the data set in a training set and a validation set. We run the model on the training set and then we test it on the validation set. If it fits the outcome 100% of the time, then we created the perfect model. That is very unlikely. Usually model creation and model evaluation imply an iterative process, meaning that you will find yourself repeating the process multiple times to increase the model efficiency. At the end of the process, we extract the model with the highest efficiency and that would be our output.

# Data Pre-Processing and Attribute Selection\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

First thing to notice in our database is that all the records are of Pima Indians women aged 21 and above. According to [this website](http://www.diabetes.co.uk/diabetes-and-ethnicity.html) dedicated to diabetes study, ethnicity has been determined as a risk factor, therefore the model we extract from this database can only be efficiently run on women of the same ethnicity.

Our attributes are:

1. Number of times pregnant

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test

3. Diastolic blood pressure (mm Hg)

4. Triceps skin fold thickness (mm)

5. 2-Hour serum insulin (mu U/ml)

6. Body mass index (weight in kg/(height in m)^2)

7. Diabetes pedigree function

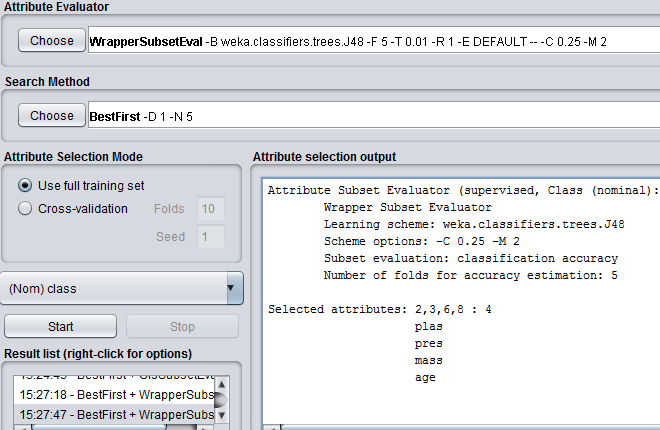
8. Age (years)

Plasma glucose concentration and insulin levels have been directly associated with increased diabetes risk, so they are important factors in our database. (Abdul-Ghani and DeFronzo 2009).

Age, BMI and blood pressure are very common questions in diabetes tests and have been proven to be linked to the increased risk of diabetes (Council 2016).

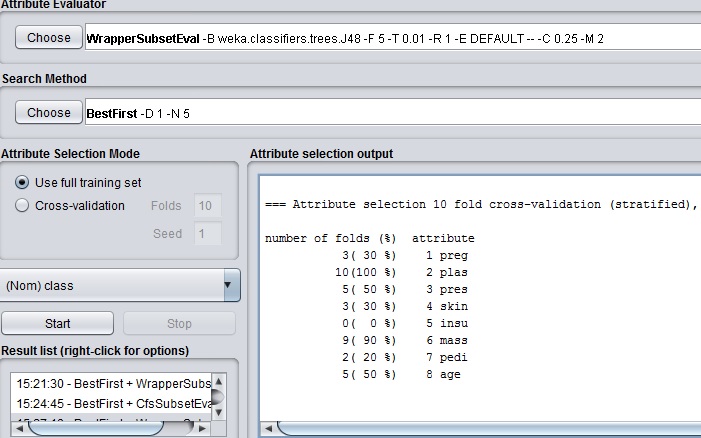
So from research, we have deducted that some of these attributes may not be relevant, namely “number of times pregnant”, ”Triceps skin fold thickness” and “Diabetes pedigree function”. We will have to run tests using WEKA to see the opinion of the machine learning algorithms as well.

Based on our research, we will choose subset: **Plas, pres, insu, mass, age**

We will run some attribute selection methods in WEKA on the database and see what we get.WEKA – Select Attributes – WrapperSubsetEval(with J48) -method BestFirst – run on training set and WEKA will give us the ideal subset(based on the training set.

So, based on the Wrapper on training set, we will use subset: **plas,press,mass,age**

Running the same wrapper with cross-validation will give us:



WEKA sees plas and mass as two very important factors, while age and pres have been considered only half the time. We can see that insu, pedi and preg have under 30% rank, meaning they are not seen as important by WEKA. Since we already have a subset of **plas,pres,mass,age,** let’s see if preg and skin affect our results in any way. So we choose subset **plas,pres,mass,age,preg,skin.**

After running our attribute seletction methods, we end up with 4 subsets that we will use to evaluate our models.

Subset 1: Initial – All attributes

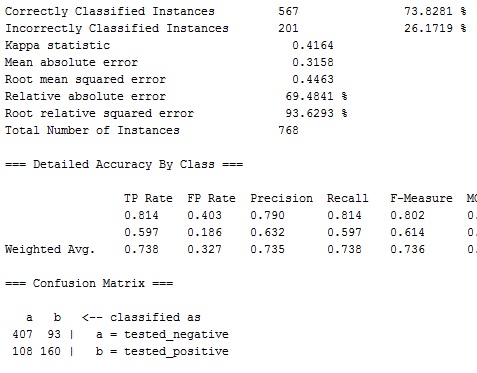
Subset 2: Research – All but preg,skin,pedi

Subset 3: Wrapper on training set: plas,mass,age,press

Subset 4: Wrapper on cross-validation: plas,pres,mass,age,preg,skin

# Model Evaluation\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

First we need to know our baseline. Running the dataset with a J48 classifier with no modifications will give us 73.8% efficiency



The confusion matrix states that 93 people were said to be negative, while they were positive, and 108 people were said to be negative, while they were positive. These are pretty bad results, especially considering that if our model would always guess negative, it would get 65% accuracy (500/768).

So we need to improve it somehow. Let’s run the same for all our subsets:

Subset 2(Research – All but preg,skin,pedi) – 74.7396%

Subset 3(plas,mass,age,pres) - 75.78%

Subset 4(plas,pres,mass,age,preg,skin- 75.65%

Our best results are for subset 3, namely 75.78%, which gives us a confusion matrix of:

a b <-- classified as

421 79 | a = tested\_negative

107 161 | b = tested\_positive

It is only a 2% better accuracy, so it is not yet impressive. We have to look through our data and see if there is anything there that throws off our ML models. We can see there are several instances where the BMI and blood pressure values are extremely low(<5). These are not normal values in humans and might have been entered by mistake.

We will run a filter and then test the subsets again.

WEKA – Filters – Unsupervised – Instance – RemoveWithValues

Remove all instances with BMI < 3.5

Remove all instances with pres < 7.5

We end up with 729 instances from 768, but the values are a bit more standardized now. Let’s run our subsets again:

Subset 1: 74.6 %

Subset 2: 74.8 %

Subset 3: 74.3 %

Subset 4: 74.4 %

This is not impressive, so we will try a second type of model, namely the MultiLayerPerceptron, to see if we get better results.

First run, before eliminating any instances:

Subset 1: 75.3 %

Subset 2: 76.56 %

Subset 3: 77 %

Subset 4: 75.65 %

From these, results, we can see that this model fits our problem better, since it gave us a very good 77% for Subset 3. We will run the same model with same subsets after eliminating the instances with incorrect values:

Subset 1: 75.3 %

Subset 2: 74.8 %

Subset 3: 75.7 %

Subset 4: 75.44 %

The best results we got were for MultiLayeredPerceptron, using Subset 3 (plas,press,mass,age), which is 77 % with the confusion matrix:

a b <-- classified as

431 69 | a = tested\_negative

107 161 | b = tested\_positive

At the beginning, we specified that the average rate of error in the real-world diabetes prediction is between 10 – 20 %. Our model gives us an error of 23 %, which is worse than that. Therefore, I don’t believe this model could be used in real-world. I think there are a couple of important missing attributes, notably if close relatives of the person have diabetes (diabetes is hereditary). If we can collect more information about the patients, we can construct a much better model, but unfortunately based on the number of instances and attributes we have available, the model is not precise and in the best case scenario it still has an error rate of 23 %.

# References:

Witten, I.H. and Frank, E. (2005) Data mining: Practical machine learning tools and techniques, Second edition. 2nd edn. San Francisco, CA: Morgan Kaufmann Publishers In. pp.13(Preface)

Center, J.D. (2016) Genetics and diabetes: Are you at risk? Available at: http://www.joslin.org/info/genetics\_and\_diabetes.html (Accessed: 1 December 2016).

Physicians Misdiagnose at an alarming rate (2013) Available at: http://www.ncpa.org/sub/dpd/index.php?Article\_ID=23148 (Accessed: 1 December 2016).

Siegler, M. (2010) Eric Schmidt: Every 2 days we create as much information as we did up to 2003. Available at: https://techcrunch.com/2010/08/04/schmidt-data/ (Accessed: 1 December 2016).

Noureddin Sadawi (2014) Model evaluation: Introduction to the cross validation and hold-out methods. Available at: https://www.youtube.com/watch?v=JxweGr03W9g (Accessed: 1 December 2016).

*Council, A. (2016) Take the test [online] available from <https://doihaveprediabetes.org/prediabetes-risk-test.html> [3 December 2016]*

*Abdul-Ghani, M. A. and DeFronzo, R. A. (2009) Plasma glucose concentration and prediction of future risk of type 2 diabetes. [online] 32 (Suppl 2). available from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2811468/> [3 December 2016*

*<http://www.diabetes.co.uk/diabetes-and-ethnicity.html> [3 December 2016]*