University of Tartu
Institute of Computer Science
Course: Introduction to Data Science
Lecturer: Meelis Kull
Name: Janica Wrosch, Bence Klujber
Date: 01.12.2019

Diabetes rate among Pima Indian women

2. Business understanding

Identifying your business goals
Background:

The data we're working with was collected from the women of an Indian stem called "Pima". It is a data set for machine learning projects on classification from Kaggle. It contains several medical attributes - such as blood pressure level - for each woman. One woman is described by one data row. In total we have 768 data points. We have one binary variable which describes whether this woman has diabetes or not. The data was collected because among Indian pima women a lot of them tend to have diabetes. Now it's about to find out why is that.

Business goals:

We are setting two business goals for ourselves which should be answered by our data mining project on the data set. First off, we aim to find interesting relations between two variables at a time. We hope by implementing this two-dimensional analysis to find new unknown correlation between some of the dimensions. These relations can help to understand our data distribution better and maybe detect relationships for having diabetes or not. This analysis is first starting with a graphical comparison to help us visually understand what the data (= the attributes are) is telling us.

The second goal we set ourselves is creating a classification model learned on the data. It should classify the women with critical attribute distribution into diabetes endangered group. We aim to create a classifier which reaches the highest F-value statistic possible. This will help us to ensure we get all diabetes endangered people. The optimal aim we want to reach is that our model could furthermore describe a women's likeliness to get diabetes according to their medical values.

Business success criteria:

We succeed if we can either find some interesting new relations in the data set by our graphical analysis and correlation analysis. Or if we can learn a classification model with F-statistic higher than 75 %.

Our situation:

Inventory of resources:

Our source data is the one data set described above. We will use 80 % of the data to form a training set and use 20 % for testing in the end. The dataset has 9 dimensions, which are the following:

Pregnancies	**Glucose**	**Blood Pressure**	**Skin Thickness**	**Insulin**	**BMI**	**Diabetes Pedigree Function**	**Age**	**Outcome (diabetes = 1)**

Their value will be described further in the next parts.

Requirements, assumptions, and constraints:

We require positive and negative outcomes to train the model. Additionally, we need all measures to be numeric, for classification the outcome variable should better be binary. All these requirements are

already fulfilled. We are assuming very common relations of high BMI, Skin Thickness and blood pressure with the likeliness to be diabetic, as overweight people tend to get diabetes more easily.

We are going to measure the goodness of our model by F-value. So other values that could be used to measure the goodness of a model like accuracy might not be too high for a good model fulfilling our goal. This is our constraint for the model: the good model does not necessarily have a good value for accuracy. It's more important to detect the endangered persons (True Positives) than evaluate all cases correctly.

Risks and contingencies:

Sometimes a pregnancy can also tease a diabetes during pregnancy it could be possible that for these cases our model cannot predict properly. So, we consider this could be a contingency. But we expect that data was taken just from non-pregnant women and outcome variable just measures real diabetes.

Terminology:

In our model we work with a lot of medical measures so that's why we will explain the meaning of the variables in this part.

Pregnancies: how many times this woman was pregnant

Glucose: the level of glucose in the blood 2 hours after taking an oral glucose tolerance test

Blood pressure: this is the diastolic measure of blood pressure, so it is the second part of a blood pressure test where 120 to 80 would be the normal blood pressure – where 80 in this case is the diastolic (lower) measure.

Skin thickness: the thickness of skin is measured next to the arm pit at the triceps, this value is taken in mm. A normal value for women would be in the range of 22 to 25 mm.

Insulin: the amount of insulin in the blood 2 hours after injection of an insulin serum

BMI: classic body mass index (weight in kg / (height in m)^2)

Diabetes Pedigree Function: this function is taking family members into account and measures how likely this person is to get diabetes according to their ancestors.

Age and Outcome (1 = diabetes, 0 = no diabetes)

Costs and benefits:

In our case we think of setting costs to measure the importance of evaluating a person as diabetical. By now we didn't define a value for cost but it should be such as that being evaluated as diabetic has higher costs than not being evaluated like this.

Defining our data-mining goals:

data-mining goals: We slightly stated above that according to our project parts, we define two rules. One will be to detect interesting two-dimensional correlations within the attributes through graphical analysis and correlation analysis.

The other one is to learn a classification model (most likely a decision tree) which classifies given data into either diabetically or non-diabetically person. We want the outcome of our model to give probabilities rather than just 0 or 1. Just like "this woman is 0.8 likely to have diabetes".

Data-mining success criteria:
As both are stated above already:

Getting a model with an F-statistic higher than 0.75.

Finding statistically significant new correlations in the data set.

3. Data understanding

Gathering data
Outline data requirements:

In our project (as said before) we are going to analyse the relation between diabetes and different medical measurements. The required data consist of blood check results, which means that the different attributes are numbers – integers and floats as well. In the first part of our project we are going to find the interesting relations between them. As soon as we have found it, we can start to work with them separately, so we need [require] only a few of the given attributes in one time. In the second part we will train a model, which gives a binary result as prediction, furthermore we are going to try to find a solution with that we can predict a probability between 0 and 1 as a risk for diabetes. So our result will be a non-integer number. All in all the used data is that what Kaggle provides, but as an input data in our project we could use any blood check results. The reason, why we have chosen this dataset, is that the risk of diabetes among Pima Indian women is much higher than among other nations or minorities. We suppose that it should have a reason and we would like to discover it. In the future we are going to be able to discover a diabetes risk regarding to the given data for any minority or nation. Only the required data (blood check result) is needed. Further plans: To fulfil our interests after the project work, we would like to test our code, perhaps visualisations on other data as well for a different group of people to compare. To get some data of people's health still can be difficult and it can cause some problems.

Verify data availability:

Our data is available under kaggle.com. We checked the license: it is public domain, that means we can work legally with it. We already started to prepare the project, our dataframe already works fine and we also did some calculations and some cleaning. Now we can say that our data availability has been verified.

Define selection criteria:

We are going to work only with one data source. The reason for that is the very specific topic of our data mining project, because to be able to work with data of people's health condition is very regulated in every country, it is almost impossible to get other data with similar attributes for testing a software. We believe that more than 760 records will be enough to get our code to work, and if case we can collect or create new or other, perhaps more data, we can suppose that it will work the same way like in the end of the project work. To make a study regarding to diabetes in an American minority seems to be extremely interesting for us. The population of Pima Indians is around 20 000 (+-5000). If we consider that we are now working only with women, we can take around the half of the population. It means the we have 768 records about their condition which is roughly 8% of the whole women population. We suppose that it will be enough to analyse and discover some reasons of diabetes.

Describing data:

We have one table with all of the following information: pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and with an outcome (true or false). To discover the reasons of diabetes we won't need all the values in the same time, but we cannot define the needed attributes in the recent phase of the project. With 8 properties and 1 outcome we have 768 records. The given that suits to our project and to prepare the models. If we would like to develop it further we may find more data, even if it is not connected to Pima Indians. What could be interesting for us: to work with pregnancies as categorical value, so analyse the diabetes with different count of pregnancies separately. We have missing values sometimes, they mean a risk during the project, for example in case of insulin (but it is not the only one).

Exploring data:

In this part we will describe the range of attributes and in relevant cases the most common elements.

Pregnancies [0, 17] – 1
Glucose [0, 100] – 100
Blood Pressure [0, 122] – 70
Skin Thickness [0, 99] – 0
Insulin [0, 846] – 0
BMI [0, 67.1] – 32
Diabetes Pedigree Function [0.078, 2.42]
Age [21, 81] – 22

Now we see that we have to clean the data before we train our model, because 0 means a missing value. We have more times missing values, so we need think about if it can cause a problem.

Verifying data quality:

We have described our data, as well we gave information about the usage of the data. Now here is the time to report the quality of our data. We may have problems in some cases, where insulin or skin thickness are missing. If we clean the data, we still have 537 records. But, we think it is only important for training a model. To make the datamining part, if we e.g. don't use these missing values, we still can get some information from the other attributes of the record. So we will not do the cleaning in the first part, just before training the model.

4. Timetable

- finding data (approx. 2 hr each person)
- creating CRISP-DM (3 hrs each person)
- data cleaning (1 hr each person)
- graphical crossdata plotting research and implementation (2 hrs each person)
    - use of seaborn pairplot for finding a quick overview to work on further on it
- finding multidimensional correlation according to the findings +
- implement visualisation of these (2 hrs each person)
- begin training several classification models with different settings (5 hrs Bence, 11 hrs Janica)
    - we will go for random forests, and as we have a small data set: also, for classification tree, we will measure performance by F-measure
    - we will use grid search to find optimal model setting
- preparing presentation, text and graphics (9 hrs each)
- evaluating which findings are most interesting and working them out understandably for the presentation
- create visualisation for the poster (6 hrs Bence)