HW10 Team A5

Herman Rull, Joonas Halapuu, Aron Sisask

Task 1. Setting up

https://github.com/RatsemaatMacrosomia/

Task 2.

Identifying your business goals

Background

Currently, there is no clear predictor whether a newborn baby will be macrosomic. People at Women's Clinic have some sense regarding the factors which affect the outcome. For example, excessive weight gain and having gestational diabetes are two attributes that raise the risk of fetal macrosomia. However, there are still some unanswered questions like, in which phase does weight gain have an effect on the weight of the infant. Or how much does mother's initial weight matter? Which has a bigger effect, relative or absolute weight gain?

Business goals

Women's Clinic wants to provide a clear guideline how women should behave during their pregnancy so that the baby is born healthy and childbirth goes well. This task is often difficult since humans differ from person to person and therefore what applies for one person does not necessarily apply for another. By looking at thousands of subjects, Women's Clinic hopes to develop some golden rules that might be easy and memorable enough for women to follow.

Business success criteria

Our project is considered successful, when doctors gain some insight regarding the risk of fetal macrosomia. As this is a subjective criteria, then our contact person Kristiina Rull can judge if any meaningful insight is gained. Furthermore, if the models we predict are accurate,

they can be used when consulting pregnant women about optimal weight gain during pregnancy.

Assessing your situation

Inventory of resources

- https://comserv.cs.ut.ee/home/files/Plhu_ITMI_2020.pdf?study=ATILoputoo&reference=A1F71E775A7E7A1709665D2A1EDD2AFD3E3E3C43 - Similar project done 2 years ago.
- Introductory information - Small overview of the project given to us by our contact person (in Estonian).
- Kristiina Rull - Doctor at Women's Clinic. She is our contact person from the University of Tartu. When we have domain specific questions or have questions about data, then she can help.
- Data - We have data of about 2000 women throughout their pregnancy. This contains their weights at week 20, 24, 28, 34 and childbirth, whether they have been diagnosed with GDM and whether they have been notified of their results of GDM test. More detailed information about the data can be found below.

Requirements, assumptions, and constraints

Project deadline is 17th December. By 1st December we will have to submit our homework task which contains the first analysis of the business and data. Also by then we must have a plan on how we will go forward with the project.

Risks and contingencies

There is a risk that we do not have enough time to achieve all the goals. Should that happen, we can drop some of the goals of the project.

Terminology

- **Fetal Macrosomia** is used to describe when a newborn is heavier than 95% of other children with the same gestation period.

- **Gestational diabetes(GDM)** is high blood sugar (glucose) that develops during pregnancy and usually disappears after giving birth.

Costs and benefits

Because this is a student project, then probably the most expensive asset in this project is the data. Since we did not collect the data, then we do not know how much this might have cost.
One of the costs in this project is also the time of the team members.
Regarding benefits, the price of the health of a newborn and the mother is invaluable. Expressing the benefits in terms of money is difficult. That is not to justify the project, but rather to make a point that this question is not very relevant at our project.

Defining your data-mining goals

Data-mining goals

Our goal is to train different types of models on the dataset which predict a probability whether a newborn is macrosomic. Since the data we have is already labeled, then we can use supervised learning methods.
Based on the created models we want to provide information which features in the data correlate with the risk of fetal macrosomia the most.

Data-mining success criteria

Our main goal is to describe which factors can be used to predict the risk of fetal macrosomia the best. We are successful if we can show that gaining a large amount of weight in a certain phase of the pregnancy is correlated with a higher chance of macrosomia compared to gaining the same amount of weight in some other phase. Lack of correlation is also valuable information, since then we show the absence of this critical phase.

Task 3

Gathering data

Outlined data requirements.

To achieve this particular task of identifying the factors that can be in correlation with the development of fetal macrosomia, in terms of required data, we need measurements of the pregnant women's weights (and BMI) over the course of the pregnancy and after giving birth. The weight of the newborn is also needed. Furthermore, since some studies have shown that the way the woman's body deals with glucose also affects fetal macrosomia, then data about the pregnant woman's blood sugar levels and diabetes is needed. Any additional information about the fetuses or pregnant women might be useful but is not essential.

Verify data availability.

The required data for this task was given to us by Tartu University Hospital Women's Clinic. The data is not publicly available and it was modified to protect the confidential information of the pregnant women.

Define selection criteria.

As mentioned above the data was customized for us by the people in Tartu University Hospital Women's Clinic. All of the data given was pre manipulated and so most of the data is relevant. The irrelevant parts of the data are the date fields, which give the exact date of giving birth or exact date of when the weight measurements were taken since we already have the weights of women at given weeks of pregnancy. Aside from the date fields as saied everything else in the data is relevant.

Describing data

Source of the data is from the Women's Clinic database gathered over the course of two years (2018 and 2019). There were in total 2027 cases of which 1771 are chosen because some of the cases were not measured over the same time intervals. Each case/patient is described by 57 fields. The data contains information about the women and their fetuses. Information of the women is mostly about their weight changes, height and ability to obtain glucose (blood sugar levels and diabetes) while most of the information of the fetus is about

the fetuses weight and age. Some of the data is obtained through calculations (like body mass index or age of the pregnant woman) and rest of the data is raw examination results (like weight of the fetus and weight of the newborn).
The data is suitable for the analysis and although more cases would be better the amount of data at the moment should also be enough.

Exploring data

While exploring the data I found that the bmi of the pregnant women at the beginning of their pregnancy was denoted sometimes with a dot and other times with a comma as the decimal separator. Therefore it would need some manipulating before any further handling.
The women's ages, newborns' weights, women's heights and weight gains during the whole pregnancy were very similar to normal distribution.
Sex between the newborns is distributed equally in the examinations. Only a handful of women (39) had given birth to a big baby before while 16% of the labours resulted in a baby with macrosomia.

Verifying data quality

The dataset fulfills all of the requirements in the "outline data requirements" sub-task and therefore is suitable for the project. There are some minor problems with the data regarding decimal separators and the data is not fully complete (a few values are missing).
Overall the quality problems should be small enough that using alternative data resources is not necessary.

Task 4. Planning your project

Task	Tools	Herman	Joonas	Aron
Homework 10 - First steps of project	Google Docs, pandas, Excel	4h	4h	3h
Research about previous work done in the area	Google, research papers	2h	2h	2h
Data exploration	Excel, pandas	2h	3h	2h
Data Preparation: selecting data, cleaning data, constructing data, integrating data, formatting data	Pandas, Excel, Jupyter notebook	1h	8h	0
Building models using supervised machine learning methods (repeatable step)	Pandas, sklearn, numpy, Jupyter notebook	6h	0	9h
Evaluation of models (repeatable step)	Pandas, sklearn, Jupyter notebook	3h	0	3h

Poster	Google Slides	12h	12h	12h