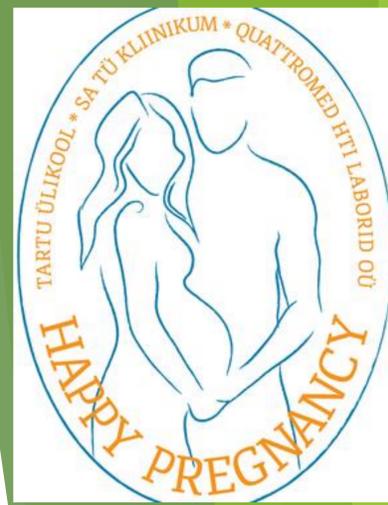# UNIVERSITY OF TARTU

Predicting newborn's weight and macrosomy based on risk factors like gestational diabetes and others

Silvia Pihu, UT CSI, silvia.pihu@ut.ee

Introduction

Gestational diabetes (GDM) is a disorder of carbohydrate metabolism which occurs during pregnancy. Timely diagnosis of the disease reduces the incidence of perinatal problems (birth traumas, need for caesarian section, etc.) (Kirss et al. 2015). Preliminary study was carried out by Kirss and others in 2015 with smaller data. Often too large (macrosomic) baby causes perinatal problems. Thus the **weight of newborn and macrosomy** (the weight of baby, that exceeds the standard weights' 90th percentile)were chosen **as target variables**. The attributes finally chosen for analysis were: the weight of woman, duration of pregnancy (gestation), gender of newborn, GDM tested- not-found-found-not tested, having macrosomic child before , having GDM before, having gestational hypertension, diabetes in family and weight-gain during pregnancy.

The **goals** were:

1. Try to predict the weight of newborn using these attributes by machine learning (regression) or weight clusters (classification)
2. Try to predict macrosomy using these attributes by machine learning (classification)

Results

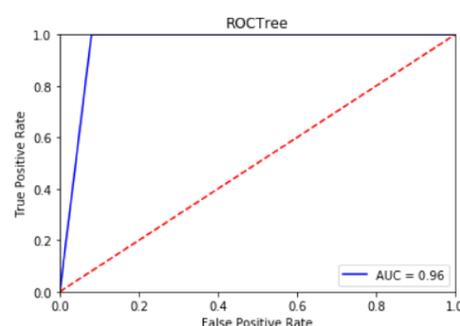Regression resulted in relatively good results:

method	MSE
linear	0.0091
Ridge	0.0091
Lasso	0.0141

But, from the coefficients by Lasso method, it can be concluded, that only the duration of pregnancy determine the weight of newborn (all other coefficients equal zero).

The classification resulted in different results by different models:

```
   model    accuracy
0    1NN     0.938144
1    3NN     0.820619
2   11NN     0.707216
3   21NN     0.705155
4  Tree0     0.915464
5 EntTree0   0.905155
6     RF     0.731959
7 EntTree0   0.690722
8 SVM_RBF    0.841237
9 SVM_poly   0.690722
```

As can be seen the best results were obtained by KNN with one neighbour and Decision tree model. The ROC-curve by 1NN is provided below. The best confusion matrix also resulted from 1NN:

320	30
0	135

Data

Data originate from the study "**Development of novel non-invasive biomarkers for fertility and healthy pregnancy**" carried out in 2012-2015 (http://www.happypregnancy.ut.ee), funded by European Regional Development Fund.

Methods

Python 3 (including its libraries numpy, pandas, sklearn) and JupiterNotebook were used to conduct the analysis. As the data was heavily unbalanced (macrosomic baby being rare), in classification **overweighting** for macrosomic babies (tripled) was used. Some of the float variables were also normalised. The determining attributes chosen were not significantly correlated to each other. **Regression** was conducted by **linear, Ridge and Lasso** methods. Ten newborn weight (together with macrosomy and gender) classes were built by cluster analysis(K-means), but using these clusters as target value did not give good results and results are not provided. In **classification**, many different models were used (**KNN with diferent number of neighbours, Tree decision, Random Forest, SVM polynomial and radial kernel**).

Conclusions

- The predictions are possible, but in further work, the choice of predictive features should be focused on. The delivery normally or by medical intervention should also be considered as target variable.
- From choice in this work, duration of pregnancy mainly determines the weight of newborn
- In predicting macosomy, 1NN and Decision Tree models gave best results

Acknowledgements

https://github.com/silviapihu/happypregnancy