Diabetes rate of Pima Indians

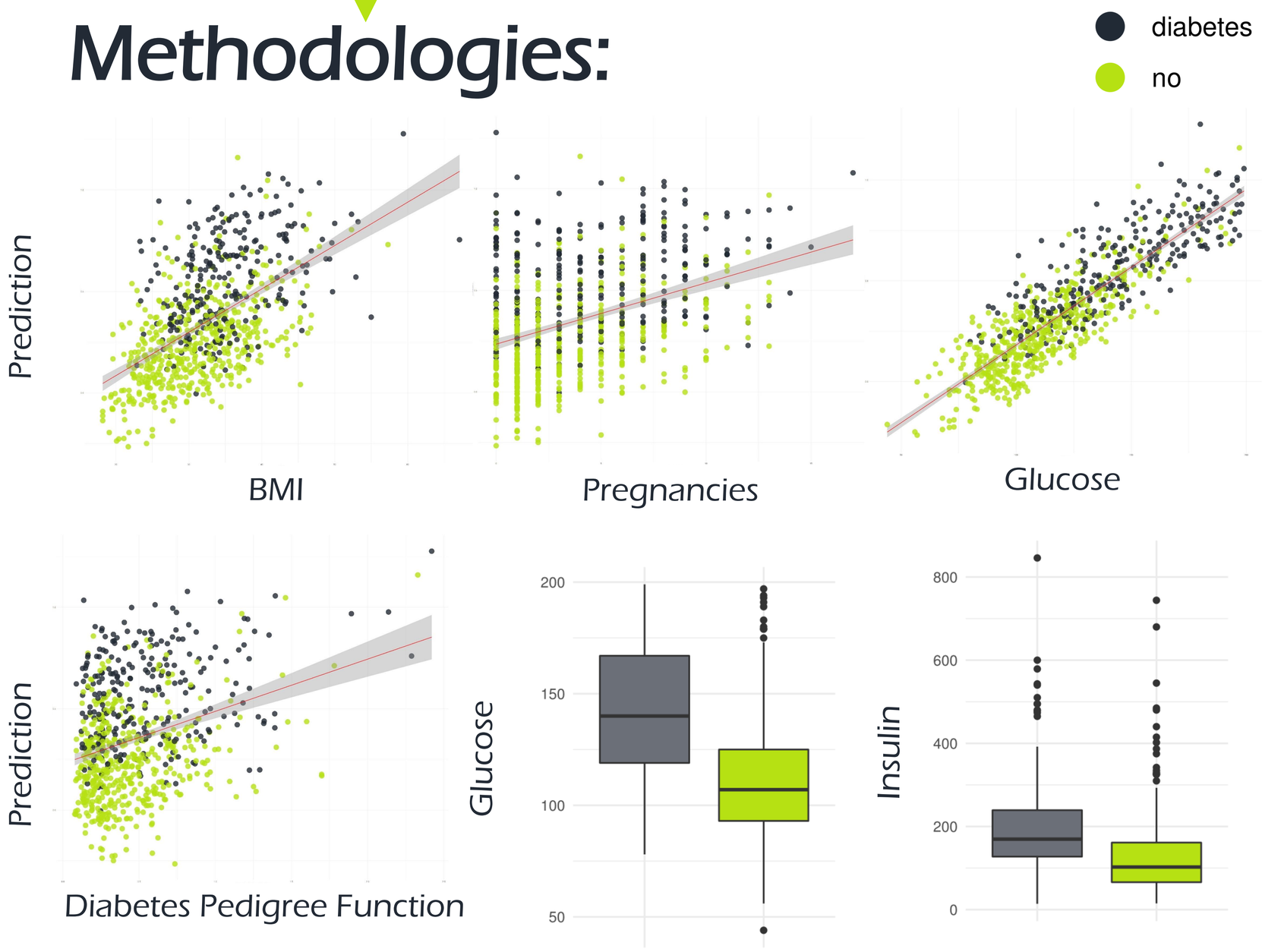
UNIVERSITY OF TARTU
Institute of Computer Science

The idea: The Pima are a group of Native Americans living in an area consisting of what is now central and southern Arizona. The data (from Kaggle) we were working on was collected because among Indian Pima women a lot of them tend to have diabetes. Now it is about to find out why is that.

Process:

I. Hypothesis
II. Gather Data
III. Data Analysis
IV. Reporting

Methodologies:



We generated a linear model that generates a prediction value between 0 and 1. So values between 0 and 0.5 suggest this women does not have diabetes, values above 0.5 suggests she has diabetes. Our scatterplots show for some attributes how the values are predicted according to the value of the attribute. The color shows their actual class. It is to be seen that especially for BMI and Glucose the data points of the groups are quite separated which suggests that in this case our estimator works quite well. Another analysis we made was testing whether there are statistically significant differences in means in between the two groups as well. For example that the average glucose level is higher in the group of women with diabetes, same for Insulin level. Both appeared to be true according to our tests.

Findings:

With analysing the Pima Indian dataset we found out that some attributes show statistical significant differences within the two groups. It means that these attributes help us to identifiy women with diabetes. Our tests using linear model has shown that four out of eight attributes have significant influence on the outcome. These attributes are the following: Amount of pregnancies, BMI, Diabetes Pedigree Function and Glucose.

Predictions:

	Predicted Positives	Predicted Negatives
Actual Positives	100	19
Actual Negatives	1	232

After training different predictive models as a result we have chosen a decision tree model with the upon given confusion matrix according to our F-measure. For training the model we could use only 352 records. To summarize our predictions we can say that our project focuses on analysing the given data, not on predictions.

Janica Wrosch | Bence Klujber

For detailed information about the project follow the given link.