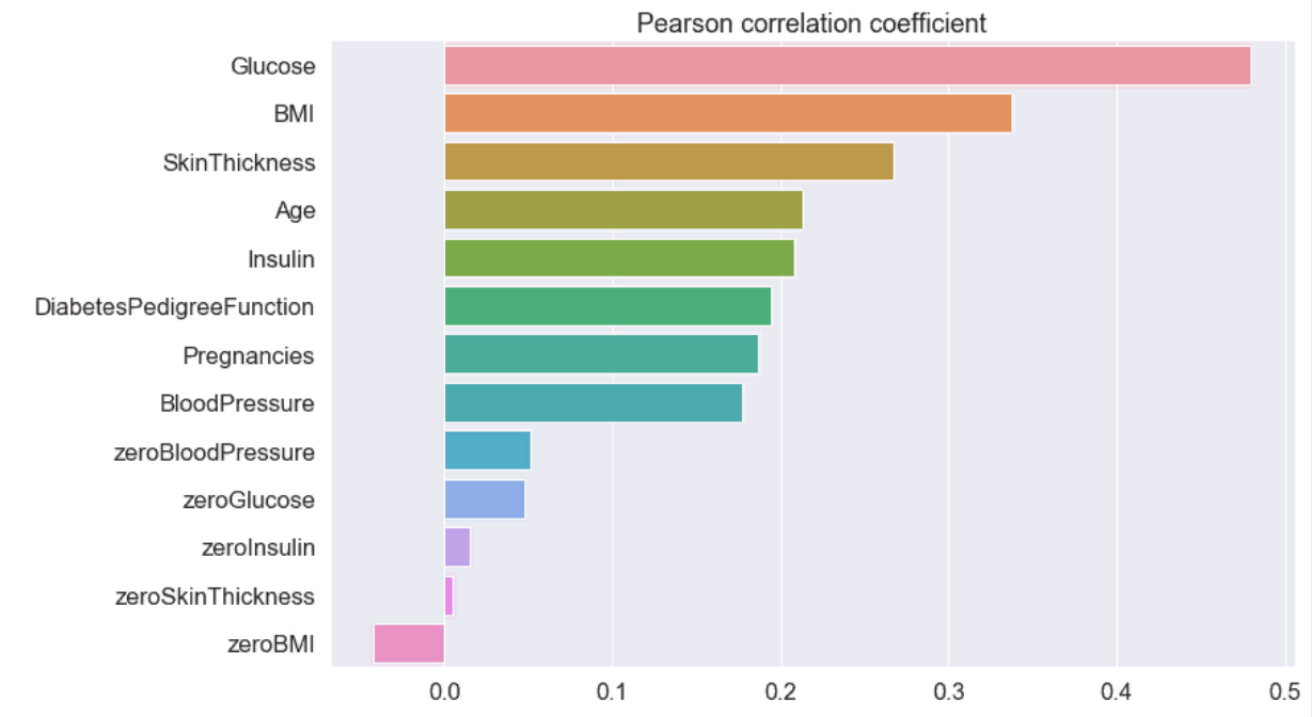
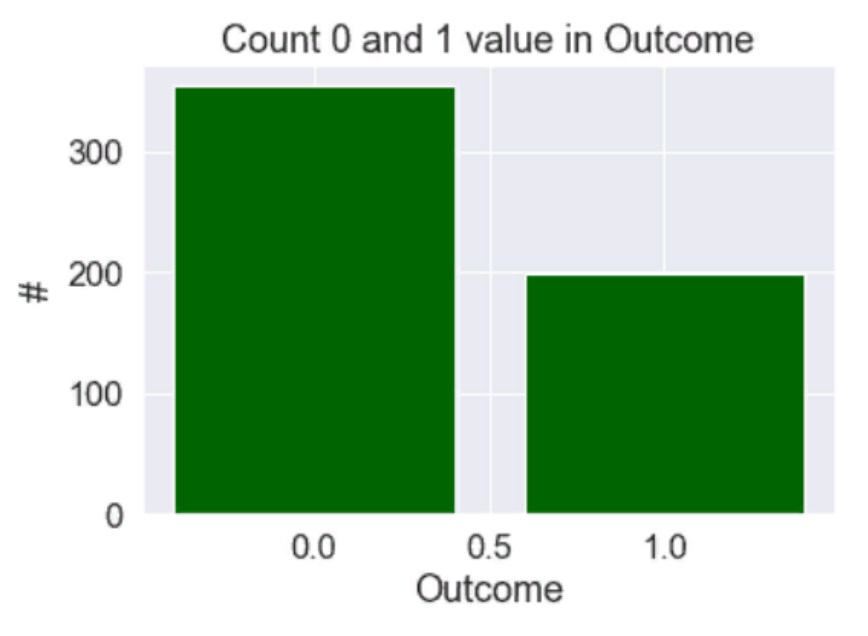
DATASCIENCE CASE

Aleksandra Lezhankina

alex_lezh@mail.ru

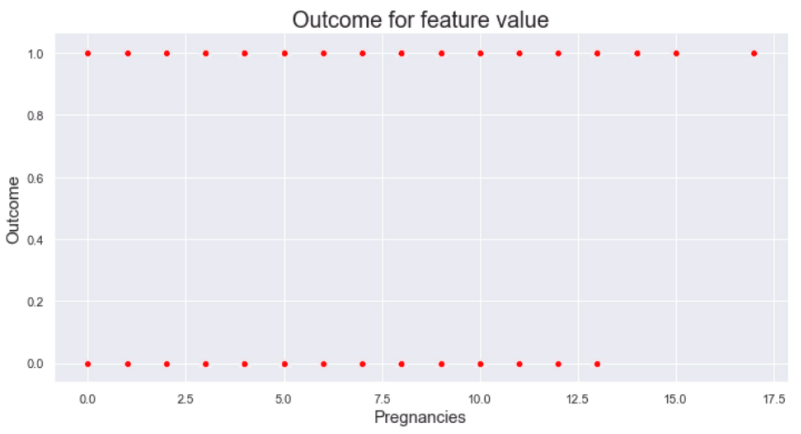


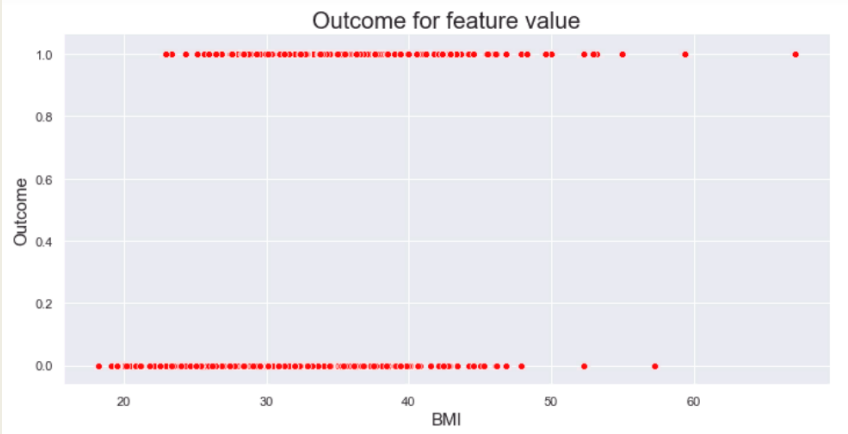


Data preprocessing

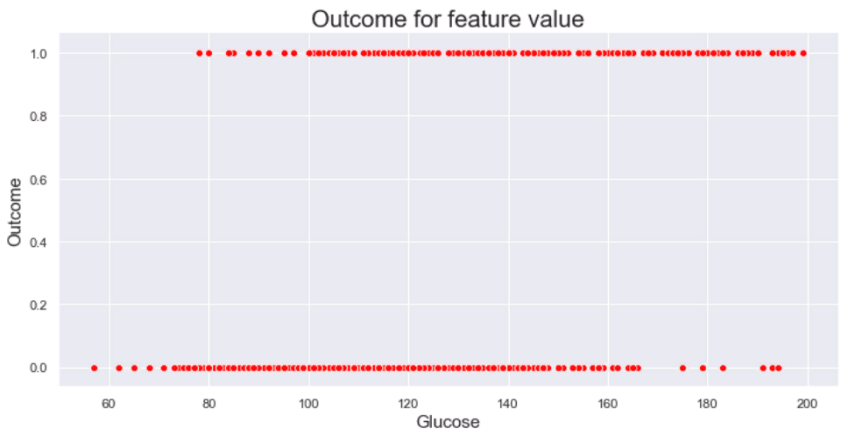
- Used in the analysis: Python (numpy, pandas, sklearn, matplotlib, seaborn);
- Imbalanced target;
- Zero values in features 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI' -> Transform into the mean value -> Create new binary features illustrating existence of a zero value in appropriate original feature;
- Normalize variables.

Hypotheses

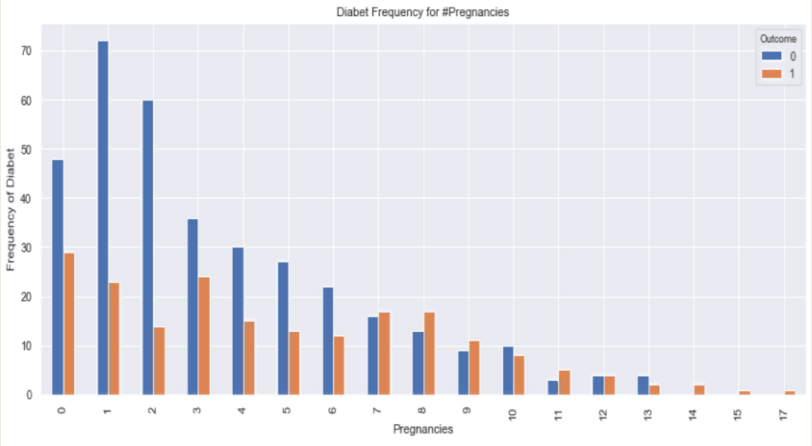




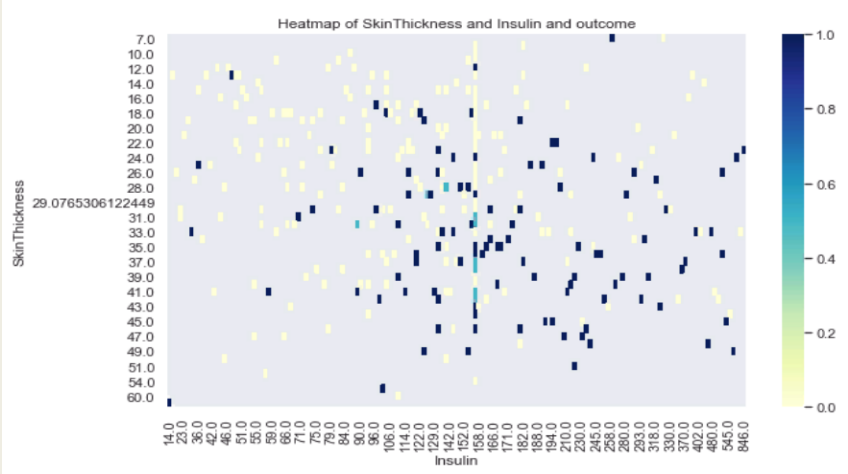
High BMI => 1 & Low BMI => 0



High Glucose => 1 & Low Glucose => 0



High #Pregnancies => 1



Low SkinThickness & Insulin => 0

High Age => 0

Results

- Model: Logistic Regression
- Accuracy metric: precision
- Train accuracy: 0.794
- CV accuracy: 0.823
- Prediction

	correct	Incorrect
0	350	4
1	183	15