Domaca naloga 4

Neža Kržan, Tom Rupnik Medjedovič

1 Short descripiton of Data

We chose the Health Data of Women from the Pima Indian Tribe, a dataset primarily used for diabetes research. It contains eight standard medical variables necessary for diagnosing diabetes for 768 women aged 21 to 81. The goal of the dataset is to explore factors contributing to diabetes development, with a binary outcome indicating whether or not a woman has diabetes.

2 Missing data

Missing data was observed only in the variables Glucose, BloodPressure, BMI, SkinThickness, and Insulin.

We observed several positive correlations between missing values, indicating that missing data in one variable was more likely associated with missing values in another. The strongest correlation was found between the variables Insulin and SkinThickness. We also determined that missing data in the Insulin variable is not random but rather more likely to occur in older women, those with higher BloodPressure, and lower values of the DiabetesPedigree variable. Additionally, missing values in SkinThickness are associated with higher BloodPressure and are further linked to the Pregnancies variable. We also found that missing data in SkinThickness is related to older age and lower values of the DiabetesPedigree variable.

For BloodPressure and Glucose, we concluded that missing data appears to be missing MCAR (Missing Completely At Random).

We found that the probability of missing values for any given variable (with the exception of BMI) tends to be higher in women with diabetes compared to those without.

3 Missing value imputation

Based on the results of our analysis, we can divide the methods into two groups. The first group includes the use of the glm() function on the original data and the listwise or pairwise deletion methods. For these cases, we expected slightly worse results, as logistic models rely on available data, potentially leading to different measures of central tendency and higher variability, which affects coefficient estimates. This was confirmed in our case, as coefficients and 95% confidence intervals within this group were identical, largely due to the glm() function removing rows with missing values, making these methods effectively the same.

The second group includes all other methods we analyzed: Random Forest, Imputation using Mean Values, and Multiple Imputation via Chained Equations (MICE). All three methods yielded similar coefficient estimates and their 95% confidence intervals (particularly Random Forest and MICE). Although the method of imputing mean values produced narrower confidence intervals, this is not ideal, as it introduces many identical values for variables with many missing entries, reducing variability and potentially underestimating it.

Based on the above analysis, we would recommend using either Random Forest or Multiple Imputation via Chained Equations (MICE) for handling missing values, as both methods appear suitable for our data. Although both methods had almost the same results, theory says different. With Random Forest we

underestimate variability and overestimate connectivity, so that is why Multiple Imputation via Chained Equations (MICE) is the most suitable method.