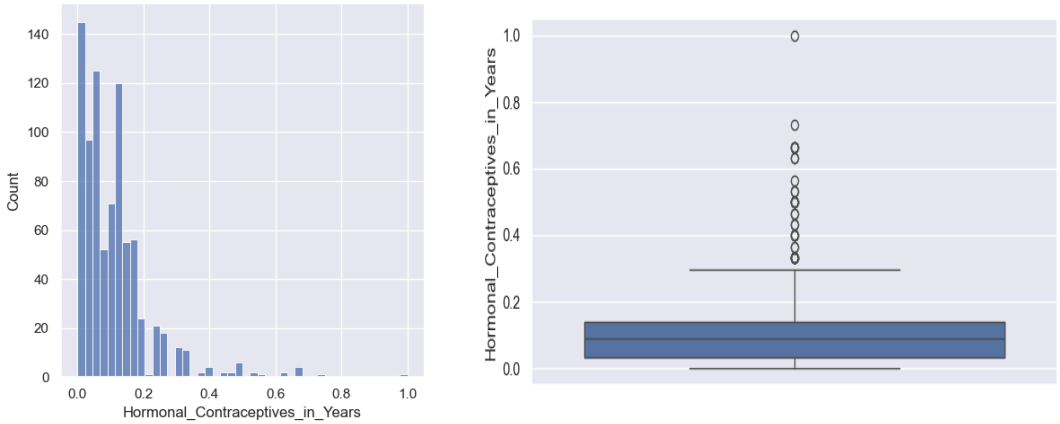
Prediction of cervical cancer

The main purpose of this project was the prediction of cervical cancer using different models such as **OLS** (*Ordinary Least Squares)*, **WLS** (*Weighted Least Squares)* and **RANSAC** (*Random Sample Consensus)*. For result evaluation of each model, **RMSE** and **adjusted R-squared** metrics were used.

The used dataset contained 36 columns (attributes) and 858 rows (examined patients). The predicted column was *Is_Diagnosis_Cancer*. The first step was to calculate the percentage of **missing values** for each attribute. For attributes with more than 85% missing values, the impact on the dependent variable was assessed, **correlation matrix** was calculated. Columns with high correlation coefficients were subjected to **KNN imputation**, on the other hand columns with lower correlation values, lower than 0.4, were removed. Consequently, the missing values problem was handled.

KNN imputation was chosen because it fills in missing data based on similarities in patients' clinical profiles, therefore it is highly useful for medical studies. It was of utmost importance to determine the right number of neighbors for this method, as the smaller values could lead to model overtraining. On the other hand, the larger values could reduce the model's sensitivity to specific changes in the data. In order to implement KNN imputation, it was necessary to normalise the data. For this purpose **MinMaxScaler** was used.

After that, the outliers were detected. Firstly, they were interpreted visually. However, for detection and removal of outliers a **99th percentile method** was used. This method was used because it can give better results in comparison to **Z-score** for not normally distributed data, as the percentile method requires no distributional assumptions. After the implementation of the percentile method, the number of rows was reduced to 748.

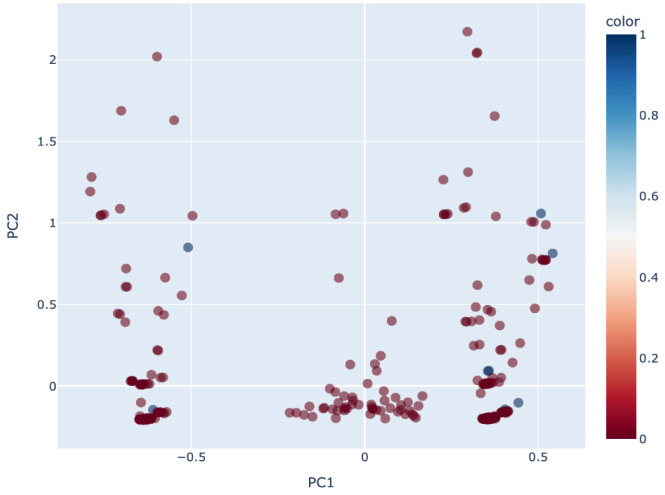


Visualisation of the data distribution for column Hormonal_Contraceptives_in_Years

It was also necessary to split data into **training, validation** and **test set**. It was done in a 60/20/20 ratio. The training set was used for training the model, while the validation and test sets were used to evaluate the performance of the trained model.

After that, an **OLS** model was created, which represented the basic model. It was checked whether it satisfied the **L.I.N.E** assumptions, linearity, independence of errors, normality of errors, perfect multicollinearity, and homoscedasticity of errors. For this purpose **F-Test**, **Durbin-Watson test**, **Anderson Darling test** and **Goldfeld-Quandt test** were used. It was shown that assumptions related to **the normality of errors** and **the absence of perfect multicollinearity** in the data were not satisfied. In this case, the failure to fulfil the assumption of normality of errors was justified because the dataset was large. To eliminate multicollinearity and to reduce dimensionality of the problem, the **PCA (Principal Component Analysis)** method was used. It was crucial to determine the number of wanted components. In this project '**elbow method**' was not used, because the primary goal was to retain **target variance** of data (95.5% was the variance in the project).

Furthermore, each of 11 Principal Components was analysed and higher importance was given to ones that explained the higher variance. Therefore, it was shown that smoking, STDs (Sexually Transmitted Diseases) and contraceptive pills have significant effects on cervical cancer development. These results were expected after visualisation of certain columns.



2D Visualisation of principal components

In order to further examine **heteroskedasticity** in the dataset, **Breusch-Pagan test** and **White test** were used. Both indicated that heteroskedasticity is present, in order to solve this problem the **WLS** model was used. This model is efficient when it comes to heteroskedasticity, as it assigns more weight to data points that have a lower variance, meaning that these points are considered to be more reliable. Therefore, a more accurate regression model will be formed.

After that **RANSAC** model was used, in order to lower the impact of outliers that still are in the dataset. For this model it was of great importance to choose the right **Estimator, Residual-Threshold** and **Max trials**. The **Hauber Regressor** was used, as it implements the loss function (Huber loss) that combines the advantages of mean squared error (MSE) for smaller errors and mean absolute error (MAE) for larger errors. This makes it more robust to outliers which are still in the dataset.

The used metrics were **adjusted R-squared** and **RMSE**. Adjusted R-squared was used in the project, as it is not only good for comparison of the models with different numbers of independent variables, but this metric also does not favor models with the greater number of variables. RMSE was chosen, because it gives information about how much the model errs on average (in actual units).

The best results were achieved using the RANSAC model. For the validation set RMSE was 0.042 and adjusted R-squared was 0.90228, while for the test set RMSE was 0.03861 and adjusted R-squared was 0.938. PCA and WLS models also gave satisfying results. On the one hand, the PCA algorithm reduces the impact of noise and the possibility of minimizing the negative effects of multicollinearity among the attributes. On the other hand, the WLS model reduced the impact of the present heteroskedasticity among the data.

At the end the final results of this project were compared with results from a project that worked with the same dataset. This project used **Gradient Boosting** and **XGBoost** method. Gradient Boosting method has given the perfect results for RMSE and R-Squared metrics, RMSE was 0 and R-Squared was 1. This was explained by the tendency of this method to overfit the model and the use of cross validation in this project. XGBoost has given worse results than my RANSAC model.

My complete project is in this PDF in Serbian and all relevant graphics are available here. On my github https://github.com/Jovana21082003/Cervical-Cancer-Prediciton, full code can be found.