Report on learning practice # 2

Analysis of multivariate random variables

Performed by
Alexander Razin
Mikhail Lovtskiy
Mark Evgrafov
Julia Pimkina
Ac. group J4132c

Saint-Petersburg
2022

Table of contents:

Dataset preparation:

```
# Colums renaming and data preparation
source_df = source_df[['gameDuraton', # this is the value we will predict (target)
                       'blueWins', # this is our category sorter
                       'blueWardPlaced', # all other values are the predictors
                       'blueWardkills',
                       'blueKills',
                       'blueDeath',
                       'blueChampionDamageDealt',
                       'blueTotalGold',
                       'blueTotalMinionKills',
                       'blueJungleMinionKills',
                       'blueTotalHeal',
                       'blueObjectDamageDealt']]

# show new dataset
source_df.head(7)
```

✓ 0.2s Python

	gameDuraton	blueWins	blueWardPlaced	blueWardkills	blueKills	blueDeath	blueChampionDamageDealt	blueTotalGold	blueTotalMinionKills	blueJungleMi
0	22.050000	0	38	13	15	31	56.039	37.001	440	
1	21.950000	1	57	18	19	8	60.243	41.072	531	
2	15.533333	0	28	7	5	20	24.014	22.929	306	
3	34.966667	0	129	39	26	36	101.607	63.447	774	
4	39.066667	1	114	35	27	40	134.826	74.955	831	
5	26.116667	1	65	23	26	18	59.839	52.221	576	
6	28.100000	0	72	26	16	31	70.270	47.107	601	

Fig.1. Dataset preparation.

In this lab, 14 dimensions will be considered, 1 target for predictive analysis, one dimension for categorization, and 12 predictor values.

From Lab 1: *"Our dataset is composed of a curated collection of over 200 publicly available COVID-19 related datasets from sources like Johns Hopkins, the WHO, the World Bank, the New York Times, and many others. It includes data on a wide variety of potentially powerful statistics and indicators, like local and national infection rates, global social distancing policies, geospatial data on movement of people, and more."*

1. Plotting a non-parametric estimation of PDF in form of a histogram and kernel density function for MRV (or probability law in case of discrete MRV).

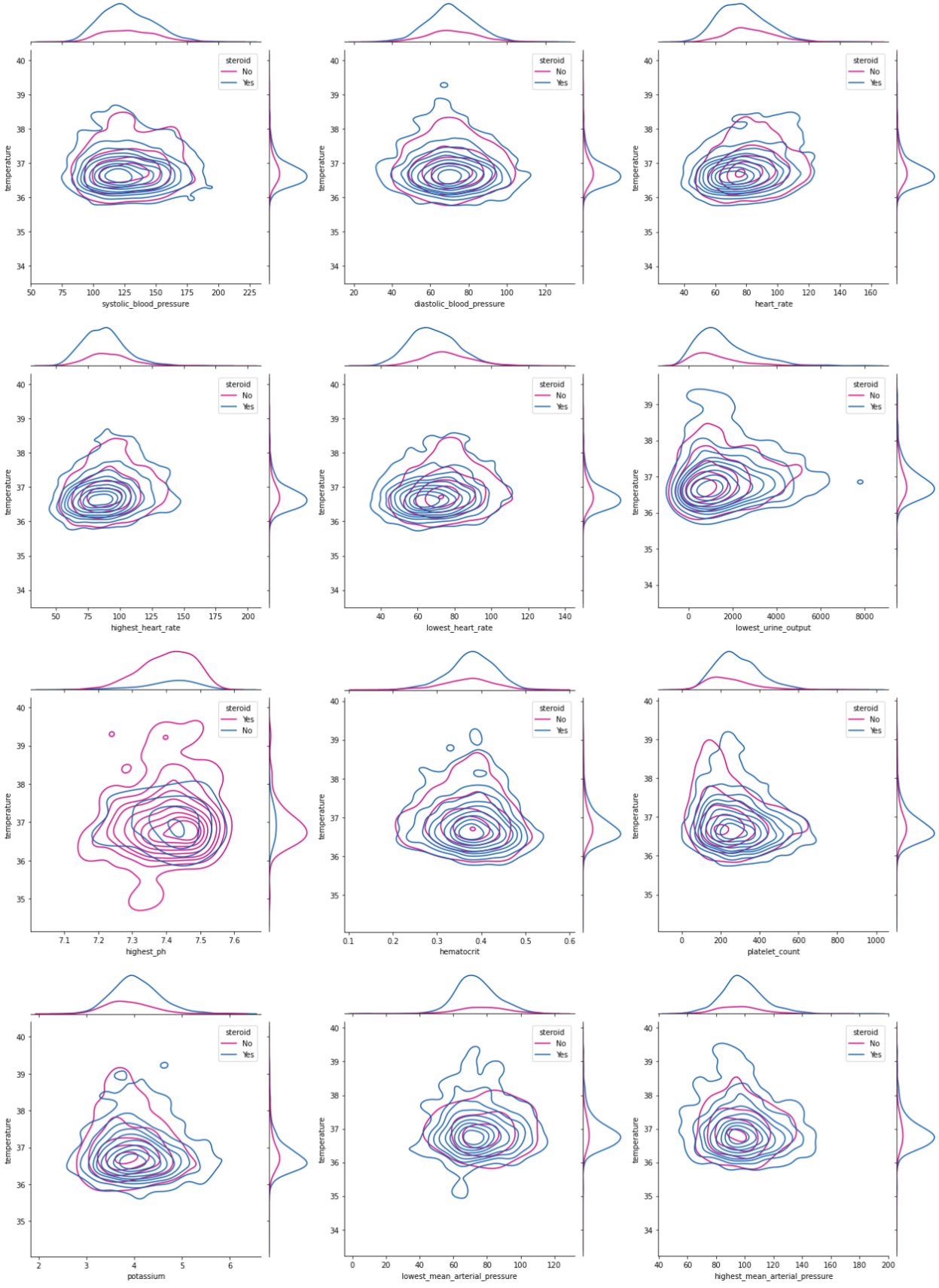

Fig.2. KDF plots.

2. Estimation of multivariate mathematical expectation and variance.

Column name	mathematical expectation	variance
systolic_blood_pressure	127.378	426.906
diastolic_blood_pressure	71.074	161.892
heart_rate	78.755	243.792
highest_heart_rate	88.972	294.274
lowest_heart_rate	70.692	183.380
lowest_urine_output	1513.093	1512313.740
highest_ph	7.404	0.006
hematocrit	0.376	0.003
platelet_count	279.154	14103.923
potassium	4.013	0.256
lowest_mean_arterial_pressure	74.460	169.481
highest_mean_arterial_pressure	98.091	230.062
temperature	36.800	0.264

Fig.3. Multivariate mathematical expectation and var

3. Non-parametric estimation of conditional distributions, mathematical expectations and variances.

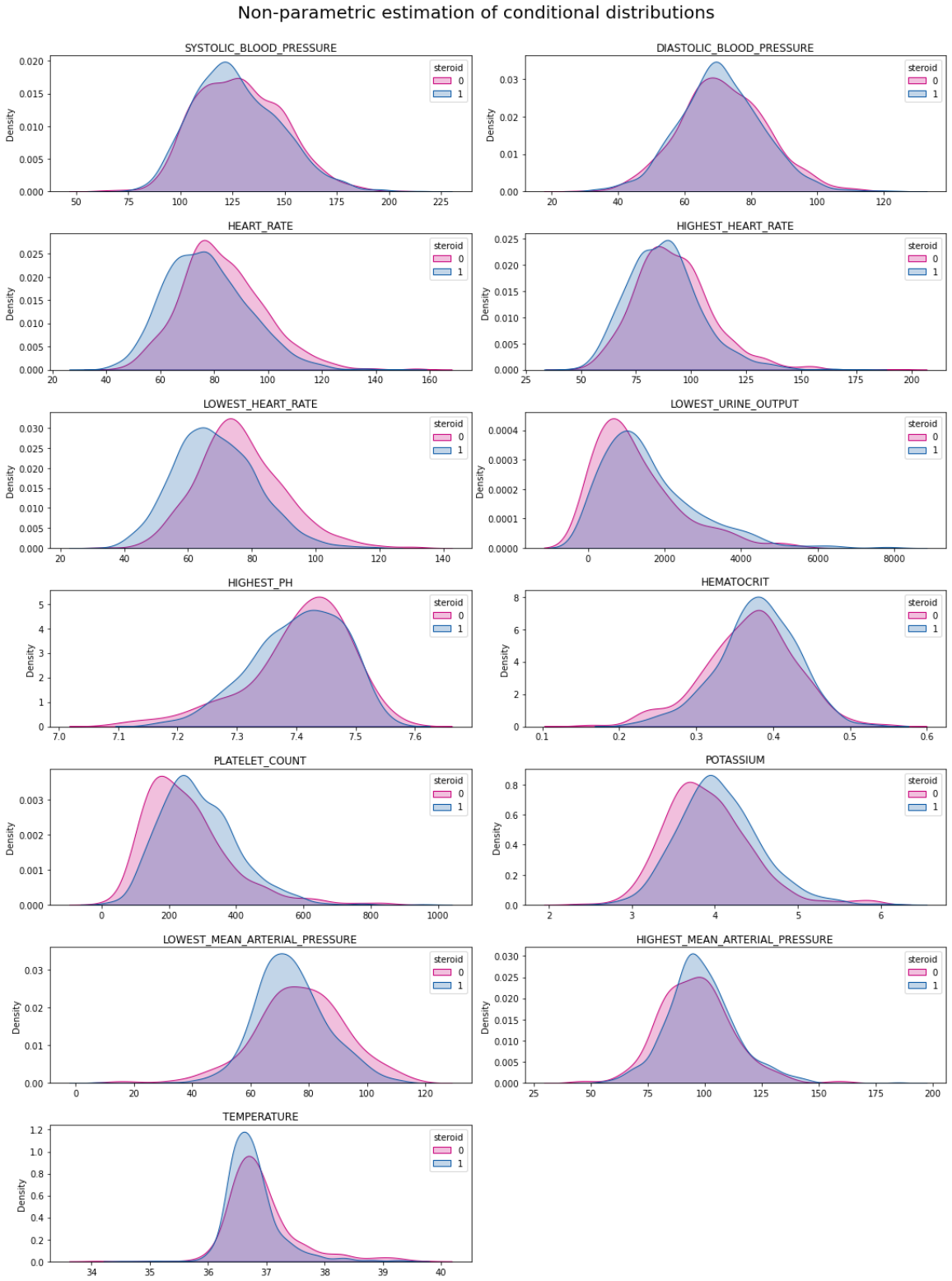

Fig.4. NPE visualization.

	steroid	No	Yes
systolic_blood_pressure	mean	128.534766	126.979983
	std	20.597861	20.673142
diastolic_blood_pressure	mean	71.683944	70.864230
	std	13.094254	12.589663
heart_rate	mean	83.210593	77.227410
	std	15.500808	15.360945
highest_heart_rate	mean	92.288777	87.834847
	std	17.736884	16.803939
lowest_heart_rate	mean	76.480454	68.707307
	std	13.588559	12.945702
lowest_urine_output	mean	1267.104265	1590.910045
	std	1112.684338	1255.318104
highest_ph	mean	7.405789	7.403398
	std	0.084260	0.076234
hematocrit	mean	0.368917	0.378634
	std	0.059671	0.052433
platelet_count	mean	252.843318	287.910276
	std	127.373787	114.467865
potassium	mean	3.912045	4.048889
	std	0.519694	0.496300
lowest_mean_arterial_pressure	mean	77.014085	73.880000
	std	15.415745	12.350257
highest_mean_arterial_pressure	mean	96.098592	98.544000
	std	15.700797	15.020128
temperature	mean	36.915696	36.760164
	std	0.616056	0.467212

Fig.5. NPE results.

4. Estimation of pair correlation coefficients, confidence intervals for them and significance levels.

	Feature	Corr coefficient	Significance level	Confidence interval
0	systolic_blood_pressure	0.04692	0	[-0.0919866665684353 ... 0.18588965965830329]
1	diastolic_blood_pressure	-0.01684	0	[-0.15577884908689754 ... 0.12209747713984104]
2	heart_rate	-0.22242	1	[-0.36514250221551486 ... -0.08726617598877628]
3	highest_heart_rate	-0.18711	1	[-0.328275040842954 ... -0.05039871461621545]
4	lowest_heart_rate	-0.28554	1	[-0.4326427482093862 ... -0.1547664219826476]
5	lowest_urine_output	0.00651	0	[-0.13243232433459837 ... 0.1454440018921402]
6	highest_ph	-0.03069	0	[-0.16963756268982486 ... 0.10823876353691372]
7	hematocrit	-0.07858	0	[-0.217679316784274 ... 0.0601970094424646]
8	platelet_count	0.18356	1	[0.046726560436002346 ... 0.32460288666274095]
9	potassium	0.10723	0	[-0.031294353018923915 ... 0.24658197320781466]
10	lowest_mean_arterial_pressure	0.07760	0	[-0.06118421974320945 ... 0.21669210648352913]
11	highest_mean_arterial_pressure	-0.03533	0	[-0.1742826917641055 ... 0.10359363446263309]
12	temperature	-0.04904	0	[-0.188021084631102 ... 0.08985524159563657]

Fig.6. Pair coefficients results.

5. Task formulation for regression, multivariate correlation.

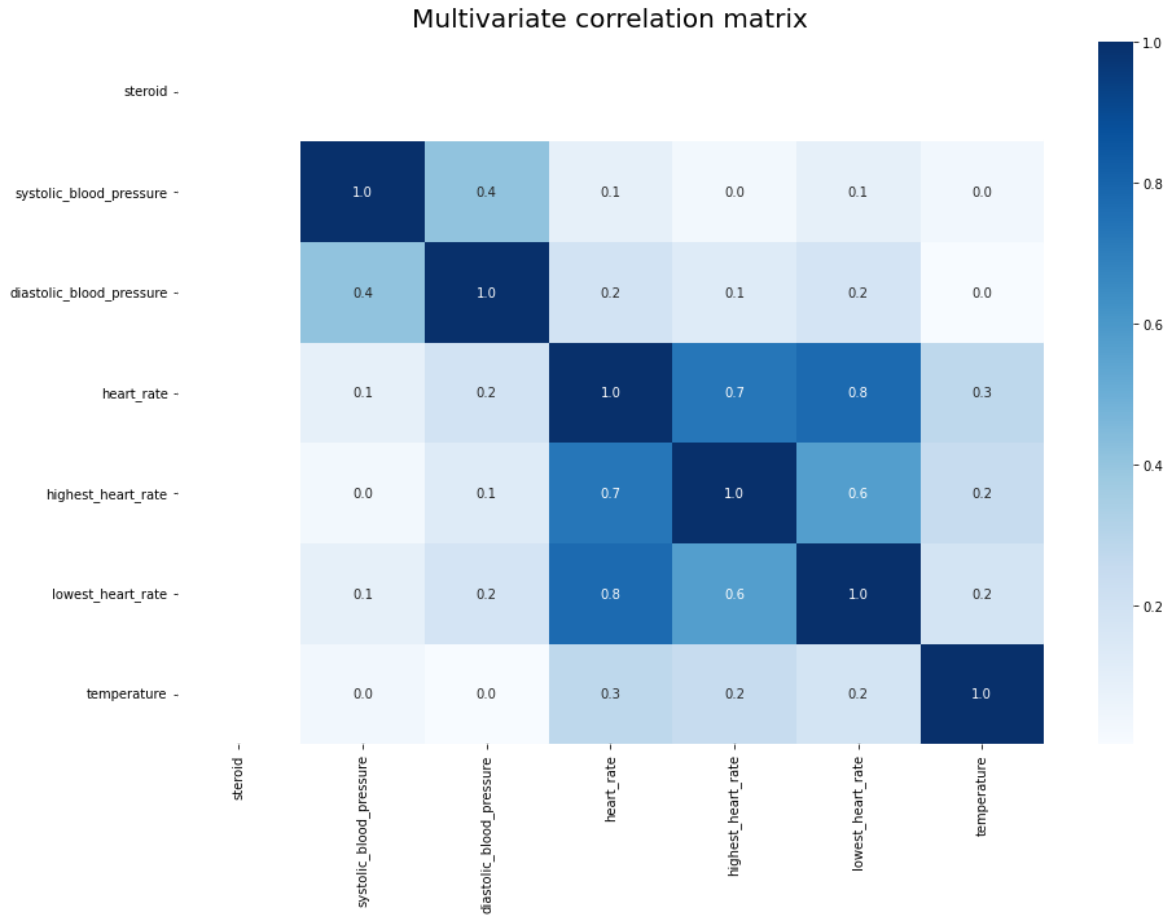

Fig.7. MVCM.

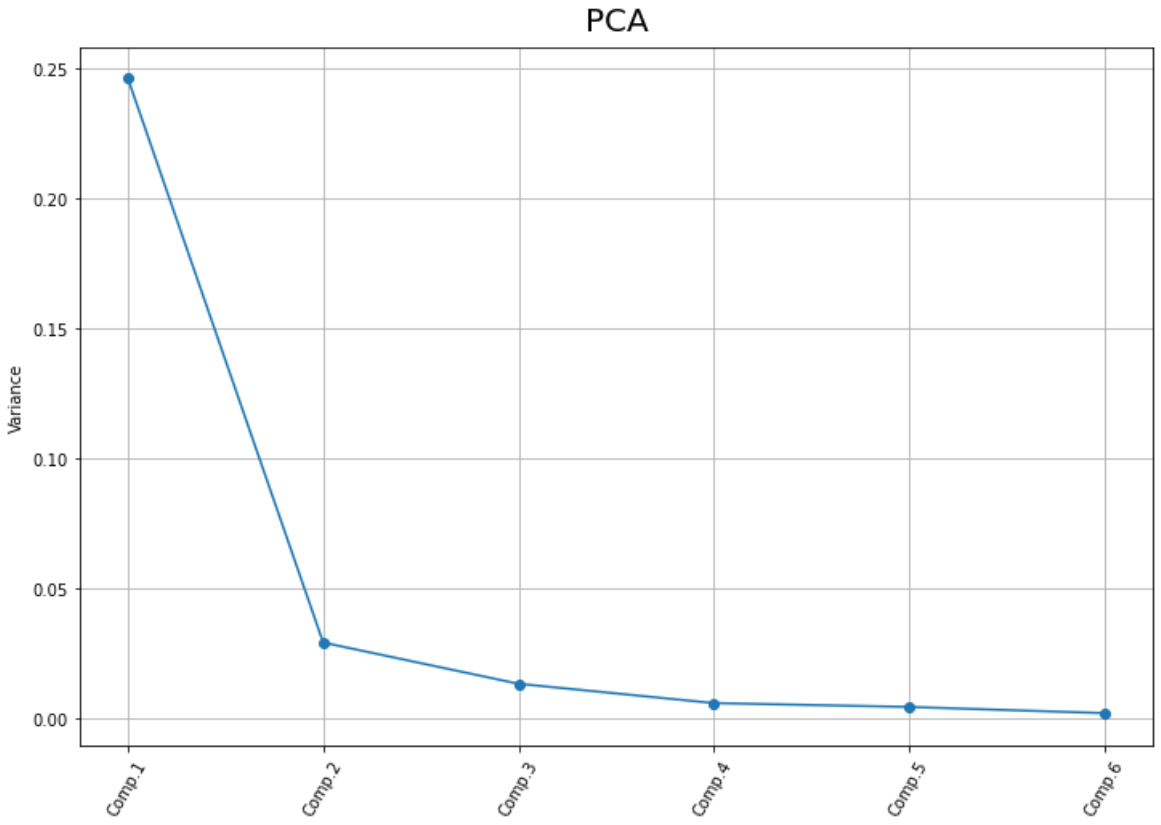

Fig.8. PCA analysis.

PCA algorithm was used in order to reduce feature dimensionality. When the number of components goes from 1 to 2, the decrease in the variance is significant and more variables are not descriptive.

So, the number of chosen variables for the regression problem should be 2.

6. Regression model, multicollinearity and regularization (if needed).

Type	Alpha	MSE	MAE	VAR
Least Squares model	-	84.231	6.271	0.239
Best Lasso model	0.058	84.170	6.235	0.239
LassoLarsIC	-	84.356	6.217	0.238

Fig.9. LSM, Lasso and Ridge models.

7. Quality analysis.

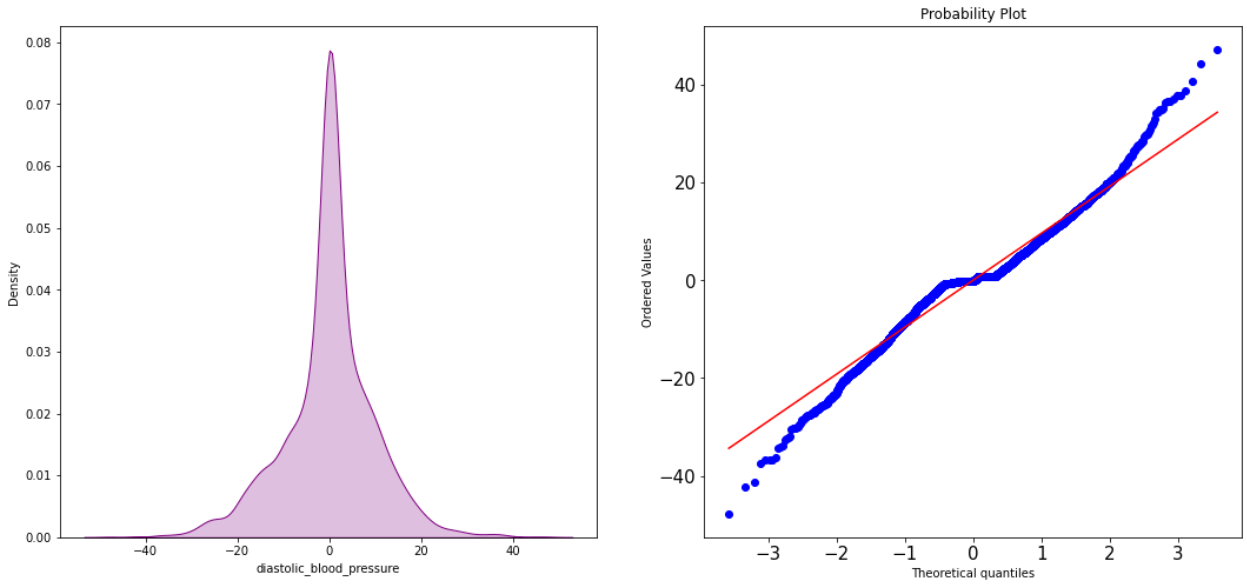

Fig.10. Results visual analysis.

```
Statistic: 70.607
+------+-------+-------------------------------------------------+
|  SL  |   CV  |                       H0                        |
+------+-------+-------------------------------------------------+
| 15.0 | 0.575 | data doesn't look normal (fail to reject H0) |
| 10.0 | 0.655 | data doesn't look normal (fail to reject H0) |
| 5.0  | 0.786 | data doesn't look normal (fail to reject H0) |
| 2.5  | 0.917 | data doesn't look normal (fail to reject H0) |
| 1.0  | 1.091 | data doesn't look normal (fail to reject H0) |
+------+-------+-------------------------------------------------+
KstestResult(statistic=0.39490478262197626, pvalue=0.0)
Residuals are not distributed normally
```

Fig.11. Mathematical results.

Sourcecode

- The full repository with all the labs: https://github.com/RazinAleksandr/M-M-MSA-ITMO
- The repo with Datasets and additional used Data info: https://github.com/RazinAleksandr/M-M-MSA-ITMO/tree/main/Datasets
- The Lab2 ipynb file: https://github.com/RazinAleksandr/M-M-MSA-ITMO/tree/main/Lab_2/lab_2.ipynb

Furthermore, you can find README file with links for every lab folder on the main GitHub repository.