Data Pre-processing and Visualization for Machine Learning Models and its Applications in Education

Konstantin Borodkin[1], Marat Nurtas[1], Aizhan Altaibek[1], Yevgeniya Daineko[1] and Temirlan Otepov[1]

[1]*International Information Technology University, Manas St. 34/1, Almaty, 050040, Kazakhstan*

Abstract

This research study explores the role and degree of influence of data pre-processing techniques in the development and application of Machine Learning models for solving prediction tasks within the domain of education. Effective data visualization techniques are essential for understanding trends, patterns, and relationships within the data, aiding in feature selection, model evaluation, and interpretation. The research study deals with various techniques for improving data quality, such as data cleaning, working with missing values, and data selection. We assume that data quality and the use of different preprocessing techniques can have a significant impact on some machine learning models performance and quality.

Keywords

Data pre-processing, data analysis, machine learning, smart education, data-driven education

1. Introduction

In the modern world, digitalization is reaching huge proportions. Both state-owned and private companies are transferring their business to an online format. Consequently, every day a huge amount of information is generated digitally from users. Every year the amount of information is growing rapidly. All this information must be transported, stored and processed. There is also no single format for data, it can be stored in different forms and structures. This requires huge computing power and the use of the latest technologies in the field of data engineering [1]. At the same time, failures occur, which leads to inaccuracies in the data or deterioration of their quality.

The performance and efficiency of machine learning algorithms [2] are intrinsically linked to the characteristics and quality [3] of the input data. In the era of big data [4], as the volume and variety of available data sources continue to expand, the importance of preparing this data becomes increasingly evident.

Private companies are now very interested in introducing the latest technologies using machine learning and artificial intelligence [5] for their needs. Recently, startups in the field of artificial intelligence and machine learning have been actively opening, huge amounts of money are being invested in the development of these industries and this trend will only continue. Some of the key advantages of automation using artificial intelligence and machine learning algorithms are increased productivity, time and economic efficiency, reduction of human errors, acceleration of business decision-making, forecasting customer preferences and maximizing sales [6]. Big companies that need to analyze and work with a large amount of data contain entire teams that monitor and maintain data quality. This once again proves the importance of data quality for future use.

DTESI 2023: Proceedings of the 8th International Conference on Digital Technologies in Education, Science and Industry, December 06–07, 2023, Almaty, Kazakhstan

kongreat13@gmail.com (K.. Borodkin); m.nurtas@iitu.edu.kz (M. Nurtas); a.altaibek@iitu.edu.kz (A. Altaibek); y.daineko@iitu.edu.kz (Y. Daineko); oneplusone113@gmail.com (T. Otepov)

0000-0003-4351-0185 (M. Nurtas); 0000-0001-8431-7950 (A. Altaibek); 0000-0001-6581-2622 (Y. Daineko)

In addition, data preprocessing cannot yet be fully automated, as it is a rather complex process that may include different techniques, algorithms and must take into account the specifics of the data and the task in order to select methods and achieve the best results. [7].

There are some articles in which scientists have investigated the impact of data quality for various machine learning models in different fields. However, in this paper, the research will be carried out in relation to the field of education and related data.

When predicting diabetes using machine learning models, the authors managed to improve the effectiveness of the model using data preprocessing techniques, such as: inserting missing values and selecting features [8].

In the study [9] about working with neural networks to predict the state of the indoor environment, the authors conclude that separate forecasting of several variables without data preprocessing can give the same accurate forecasts as simultaneous forecasting with data preprocessing, however, the computational costs of training several neural networks for separate forecasting should be taken into account.

In another article [10], scientists decided to find out how data processing will affect machine learning models in prediction tasks. As a result, it turned out that data processing can have a strong positive impact on the results and quality of the forecast, but it can also have a negative impact on the effectiveness of the forecast using machine learning methods.

The authors of another paper [11] studied the possibilities of 6 different algorithms, as well as methods for preprocessing data in the task of classifying the electroencephalogram signal to determine the drowsiness of the driver when driving using machine learning. As a result, they came to the conclusion that the type of algorithm used to solve the problem has a higher impact on the results than preprocessing. However, data processing also improves simulation results. Also, in situations where data preparation is not possible, it is preferable to use tree-based machine learning algorithms.

Another work [12] used machine learning algorithms to predict air pollution. The impact of data preprocessing and feature selection was also evaluated. As a result, when using these methods, better accuracy and efficiency of the models were achieved.

In another study [13] about preprocessing of near-infrared spectra, the researchers concluded that data preprocessing has a big impact on small data sets. With an increase in the amount of data, preprocessing methods lose their effectiveness. Models trained on a large amount of data are more accurate.

One more article [14] investigated the impact of data processing in corporate data analysis. Various preprocessing methods, as well as various machine learning algorithms, were reviewed. As a result, it was empirically proved that some of the preprocessing algorithms have a significant impact on the accuracy of forecasting.

In the case of processing unstructured medical data [15], the researchers also analyzed the most influential methods of data preprocessing. As a result, it was found out that for the analysis of handwritten text, the most effective stages were normalization and correction of errors.

Machine learning in the field of education continues to evolve. New ways of application are found for these technologies. This research study will examine the use of machine learning to analyze and predict students' grades on exams, depending on the indicators of their life and family. Machine learning can also be used for research, automation of management, improvement of online learning, personalization of learning, creation of smart applications.

Data visualization [16] is also an important step in the process of solving a problem using machine learning algorithms. Visualization is applied at various stages in the process of solving the problem.

Despite the clear importance of data preprocessing, a comprehensive understanding of its real-world impact remains a dynamic area of research and application. The complexity of this issue arises from the interplay of diverse preprocessing techniques, the unique characteristics of different datasets, and the specifics of machine learning models. As such, the impact of data preprocessing is not one-size-fits-all but is, instead, context-dependent.

Moreover, there is limited insight into how various preprocessing methods influence the resilience of machine learning models in the face of noisy data, outliers. This dearth of knowledge

can impede the broader adoption of best practices in data preprocessing, limiting the potential of machine learning in real-world applications.

In this regard, the topic of the impact of data quality on machine learning models in education needs more research.

2. Problem statement

This research study assesses the impact of data quality and various data preparation algorithms on the primary quality metrics of machine learning models. It also evaluates their performance in forecasting tasks within the field of education while analyzing how data visualization contributes to problem-solving using machine learning models. We endeavor to provide valuable insights and empirical evidence that guide data scientists, researchers, and practitioners in making informed decisions regarding data preprocessing, ultimately enhancing the effectiveness of machine learning applications across domains.

3. Data pre-processing and analysis

To investigate the issue that stands in this research study, datasets from open source were used.

The dataset [17], which will be used for the prediction task using machine learning algorithms, contains various information about students. The dataset used in the study is an extended version of the original “Students Exam Scores” dataset [18]. It contains a large number of columns and records. It also includes inaccuracies in the data, such as missing values and not informative columns. This dataset is used to train machine learning models to predict student grades. The dataset's dimensions are 30641 rows by 14 columns. The target variable in this study that will be predicted is the result of students' exam in mathematics.

Main attributes of this dataset:

1. Gender: this attribute indicates the gender of students (male or female)
2. Race/ethnicity: students are categorized into groups based on their race or ethnicity, labeled as groups A, B, C, D, E
3. Parental Education: this attribute represents the highest level of education achieved by the students' parents, with categories such as "high school", "some college", "associate's degree", "bachelor's degree", "master's degree"
4. Lunch Type: this attribute describes the type of lunch that students receive, with options for "standard" or "free/reduced"
5. Test Preparation Course: it indicates whether a student completed a test preparation course. Options include "completed" or "none"
6. Parent Marital Status: married/single/widowed/divorced
7. Practice Sport: frequency of student's sports activities never/sometimes/regularly
8. Is First Child: is this student first child in family or not - yes/no
9. Number of Siblings: from 0 to 7
10. Transport Means: what kind of transport does the student use to get to the place of study school bus/private
11. Weekly Study Hours: number of hours of self-study during the week
12. Math Score: this is the score a student achieved on the math portion of the exam
13. Reading Score: this is the score a student achieved on the reading portion of the exam
14. Writing Score: this is the score a student achieved on the writing portion of the exam.

Most of the variables are categorical. This means that in the process of data preparation, they will need to be digitized for further training of machine learning models.

This dataset was generated in order to study the relationship between student demographics, preparation, and academic performance.

The dataset allows researchers and data scientists to explore various aspects of student performance and understand how demographic factors and preparation influence exam scores. It can be used for tasks like predictive modeling, clustering, and statistical analysis. Researchers

often use this dataset to examine disparities in student performance based on demographic attributes and to develop insights into factors that can improve student outcomes.

The dataset is valuable for educational research. It is an excellent resource for exploring educational data and conducting various analyses.

Table 1
Example of dataset

	Unnamed: 0	Gender	EthnicGroup	ParentEduc	LunchType	TestPrep	ParentMaritalStatus	PracticeSport	IsFirstChild	NrSiblings	TransportMeans	WklyS
0	0	female	NaN	bachelor's degree	standard	none	married	regularly	yes	3.0	school_bus	
1	1	female	group C	some college	standard	NaN	married	sometimes	yes	0.0	NaN	
2	2	female	group B	master's degree	standard	none	single	sometimes	yes	4.0	school_bus	
3	3	male	group A	associate's degree	free/reduced	none	married	never	no	1.0	NaN	
4	4	male	group C	some college	standard	none	married	sometimes	yes	0.0	school_bus	
...	...	...	...	...	...	...	...	...	...	...	...	
30636	816	female	group D	high school	standard	none	single	sometimes	no	2.0	school_bus	
30637	890	male	group E	high school	standard	none	single	regularly	no	1.0	private	
30638	911	female	NaN	high school	free/reduced	completed	married	sometimes	no	1.0	private	
30639	934	female	group D	associate's degree	standard	completed	married	regularly	no	3.0	school_bus	
30640	960	male	group B	some college	standard	none	married	never	no	1.0	school_bus	

At the stage of studying data, it is useful to use visualization tools in order to better understand the task and ways to solve the problem.

Analyzing the ethnicity column using pie-chart [19] reveals insights into the diversity of the student population and allows for the exploration of potential disparities in academic performance related to race and ethnicity. According to race, students are distributed as follows:

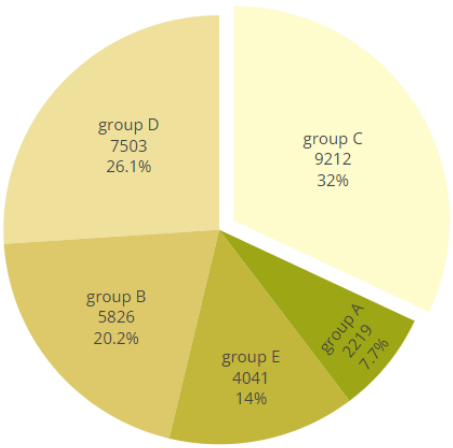

Figure 1: Distribution of students by race

The "ParentEduc" column in the dataset provides information about the highest level of education achieved by the parents of the students in the dataset. Analyzing this column with bar-chart [20] reveals insights into the educational background of the students' parents and its potential influence on student performance. We can see that a smaller part of the students' parents had a higher education.

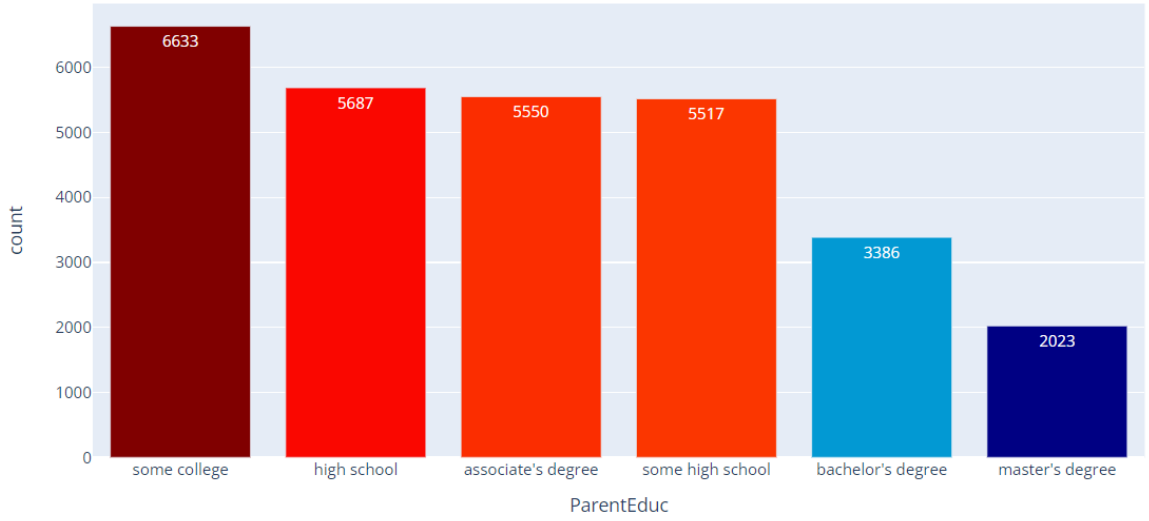

Figure 2: Distribution of students by parental level of education

Further, in the process of data analysis, 2 fields are added that help to better characterize the overall performance of students. The first field is the student's total score for all 3 exams, the second field is the percentage of the total score scored from the maximum. The maximum value for each exam is 100. It is worth noting that the dataset has an almost equal distribution of students by gender: 15424 females and 15217 males. Thus, we can estimate the overall average academic performance depending on various values, for example, gender:

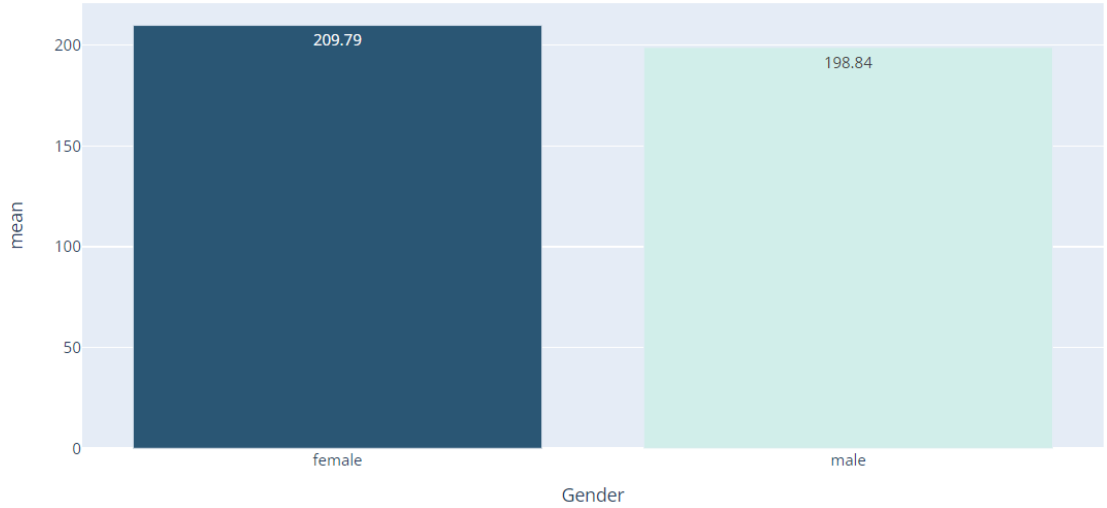

Figure 3: Average overall academic performance depending on gender

Using the correlation matrix [21], it is possible to determine how strongly the numerical variables correlate with each other. Based on the matrix in Figure 4, it can be established that all types of tests have a strong positive correlation relative to each other. For example, this means that students who wrote a written exam well also got a good score in mathematics. Negative values in the matrix mean the inverse relationship of the two variables. There are no large values in this data. Values close to 0 mean that the variables are poorly correlated with each other, there is no direct or inverse relationship between them. Total Score and Total Pct are essentially one indicator needed for data analysis, so their correlation is 1.

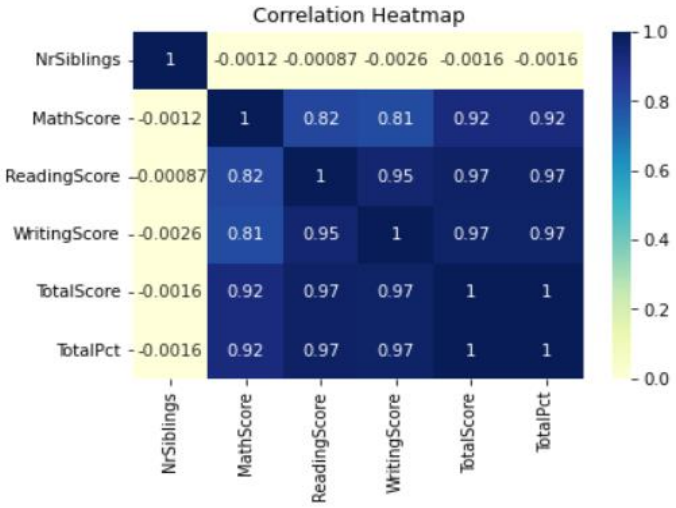

Figure 4: Correlation heatmap

Further, in the process of data analysis, some patterns also emerged that may affect the methods of solving the problem. The mean overall score among all students was 204 points. Students of race "E" were more successful on all types of tests. They scored an average of 18 points more in all subjects combined. The level of education of parents also has an impact on students' academic performance. On average, students whose parents had a master's degree received 20 points more. Students who preferred a standard lunch set scored 29 points more than students with a free lunch. Students who completed preparatory courses received 20 points more than others. Also, students who studied more hours a week on average had slightly higher academic performance compared to the rest. The other features did not have such a big impact on student performance.

A graphical representation of these patterns is shown in Figure 5.

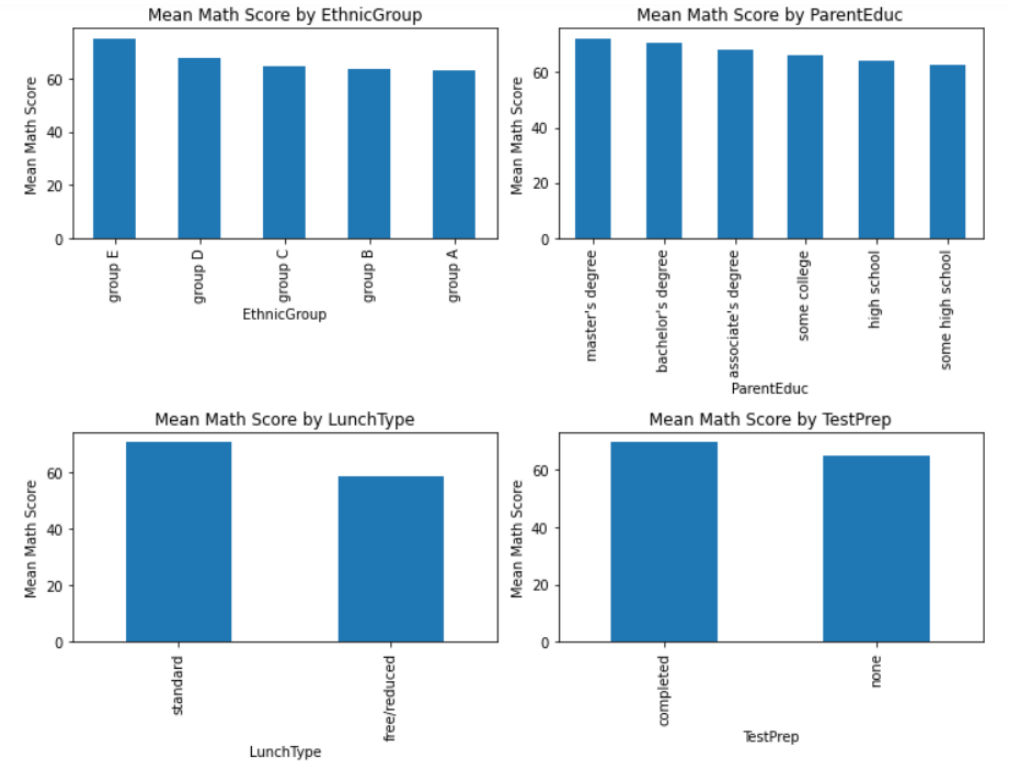

Figure 5: Graphs of variables that had the biggest impact on Math Score

Analyzing the histograms [22] from Figure 6 below, we can notice a significant difference in academic performance between men and women in different subjects. Based on the information, it can be concluded that men show the best performance in mathematics, and women in writing and reading exams.

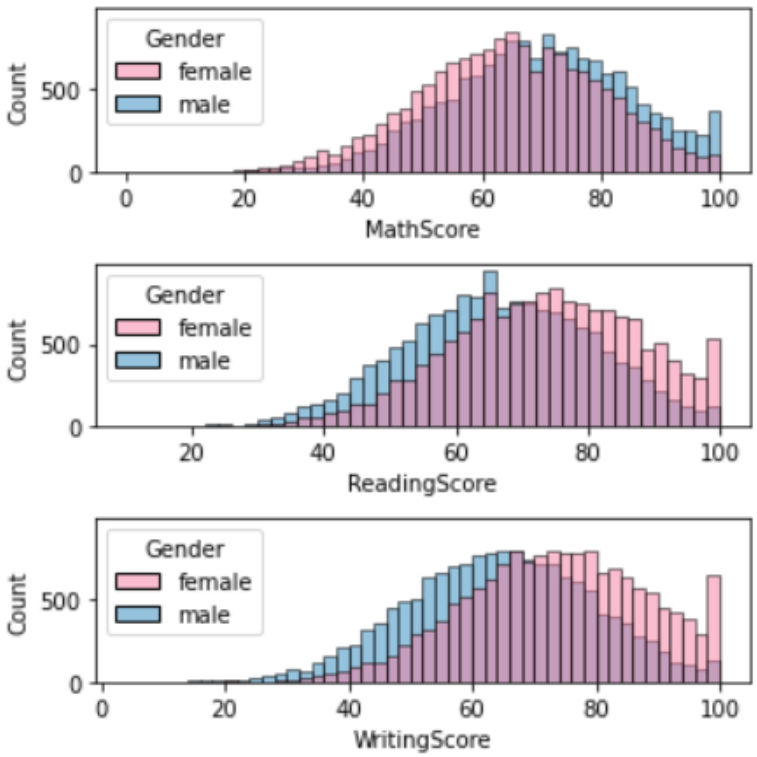

Figure 6: Comparison of male and female scores in different exams

Next, it is necessary to assess the quality of the available data. To do this, it is necessary to check for data duplication, missing values, and analysis of outliers in the data.

Since we do not have a unique student ID, we need to check the data for duplicate rows across all columns. Duplication of data was not detected.

Further, using the interquartile range [23], outliers [24] in the data among numerical variables were checked. For Interquartile Range the formula was used:

$$IQR = Q_3 - Q_1 \quad (1)$$

where Q_3 - upper (75th) quartile, representing the value above which 25% of the data falls, Q_1 - lower (25th) quartile, representing the value above which 75% of the data falls.

Typically, values outside the range (Q1 - 1.5 * IQR, Q3 + 1.5 * IQR) are considered outliers.

The number of outliers turned out to be insignificant: 90 lines for Reading Score and 106 lines for Writing Score. Also, they cannot be fully called outliers, since they are results of students' performance on these types of exams.

There was a certain amount of missing data in the dataset. For each column, it was no more than 11% of the total amount of data. In general, the number of rows with at least one missing value is 11398, which is a percentage of 37.2%. Most of the missed data was for categorical columns.

During the analysis, it turned out that there is one extra column that contains only a number. It was not described in documentation in any way. It looks like it's just the entry number in the dataset. It also did not contain useful information. Therefore, this column has been deleted.

Further, in the process of data preparation, it is necessary to translate categorical variables into numerical form. To do this, we have assigned a unique numeric value for each category. For

example, a field with a student's gender is indicated as follows: 1 - male, 2 - female. The other fields were processed in the same way.

Gender	0
EthnicGroup	1840
ParentEduc	1845
LunchType	0
TestPrep	1830
ParentMaritalStatus	1190
PracticeSport	631
IsFirstChild	904
NrSiblings	1572
TransportMeans	3134
WklyStudyHours	955
MathScore	0
ReadingScore	0
WritingScore	0
TotalScore	0
TotalPct	0

Figure 7: Null values by columns

Thus, in the process of data analysis using visualization tools, it was possible to detect interesting and useful patterns and also problems in the data, which in the future can help in solving certain tasks with this dataset.

4. Methods and research

To solve the problem of predicting students' grades in mathematics based on their available data, a large number of machine learning algorithms with supervised learning were selected. Thus, in the process of training models and evaluating their results, it will be possible to compare the impact of various processing methods on the accuracy of different models. This research study did not pay much attention to the selection of model parameters. Default parameters were used, the values of which can be found in the documentation [25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38] for these algorithms. List of algorithms used for data analysis and evaluation of results:

- LGBMRegressor [25]
- XGBRegressor [26]
- GradientBoostingRegressor [27]
- RandomForestRegressor [28]
- DecisionTreeRegressor [29]
- MLPRegressor [30]
- KNeighborsRegressor [31]
- SVR [32]
- CatBoostRegressor [33]
- LinearRegression [34]
- Lasso [35]
- Ridge [36]
- ElasticNet [37]

Then the initial data set was copied several times. Certain data manipulations were carried out on each of the copies in order to then assess the degree of influence of various data processing methods on the final result. Since most of the values are missed in the categorical columns, it would not be correct to use measures such as the average or median value to fill in these values.

- Dataset 1 - all rows with missing values were deleted. Thus, the amount of data for model training has been significantly reduced to 19243 rows
- Dataset 2 - a mode [38] from each column was inserted into the missing values
- Dataset 3 - a new category has been added for the missing values

- Dataset 4 - 1000 empty values were randomly added to the numerical columns, which have a strong correlation with the target variable. Then the empty values are replaced by the average value of the column
- Dataset 5 - removed columns that, as a result of the analysis, did not have a strong impact on the results of students in the math exam: 'TransportMeans', 'NrSiblings', 'IsFirstChild', 'ParentMaritalStatus'

Further, the datasets were divided into sets for training and prediction in a ratio of 80% to 20%. All types of models were trained on these sets and the results were entered into a common table for further analysis.

To assess the impact of data processing techniques on the change in the load on the computing system during the training of the models, the load on the Central Processing Unit (CPU) [39] was measured. The training process took place on a local machine with a processor with 4 physical and total 8 virtual cores.

The following metrics were used to evaluate the accuracy of predictions.

Mean Absolute Error (MAE) [40] - measures the average absolute difference between actual and predicted values.

$$MAE = \frac{1}{n}\Sigma|y_i - \hat{y}_i| \tag{2}$$

where n - number of errors, y_i - actual values, $\hat{y}_i$ - predicted values.

Square Root of Mean Quadratic Error (RMSE) [40] - also measures the difference between actual and predicted values, but it penalizes larger errors more than MAE.

$$RMSE = \sqrt{\frac{1}{n}\Sigma(y_i - \hat{y}_i)^2} \tag{3}$$

where n - number of errors, y_i - actual values, $\hat{y}_i$ - predicted values.

Coefficient of Determination (R2) [41] - measures the proportion of the variance in the dependent variable explained by the model. It ranges from 0 to 1, where 1 means the model perfectly fits the data, and 0 means the model doesn't explain anything.

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \underline{y}_i)^2} \tag{4}$$

where n - number of errors, y_i - actual values, $\hat{y}_i$ - predicted values, $\underline{y}_i$ - mean of actual values.

5. Results

During the training of machine learning models, measurements of the load on the CPU were also made. As a result, it turned out that some of the models did not put any load on the computing system at all, or its value was so small that monitoring tools could not track the load. Thus, taking into account the indicators with a zero value, the average load for five datasets can be seen in Table 2.

In general, the spread of values turned out to be small. The second dataset showed the highest load on the processor, where the missing values were replaced by a mode.

The lowest load on CPU was exerted by the 5th data set, where the missing values were excluded, and the most informative columns for forecasting were selected. This helped to reduce the load on the computing system because the total amount of processed data has significantly decreased.

Table 2
Average CPU Usage for every dataset (including zeros)

Dataset number	Mean Value
1	19.79
2	20.15
3	18.77
4	19.7
5	19.67

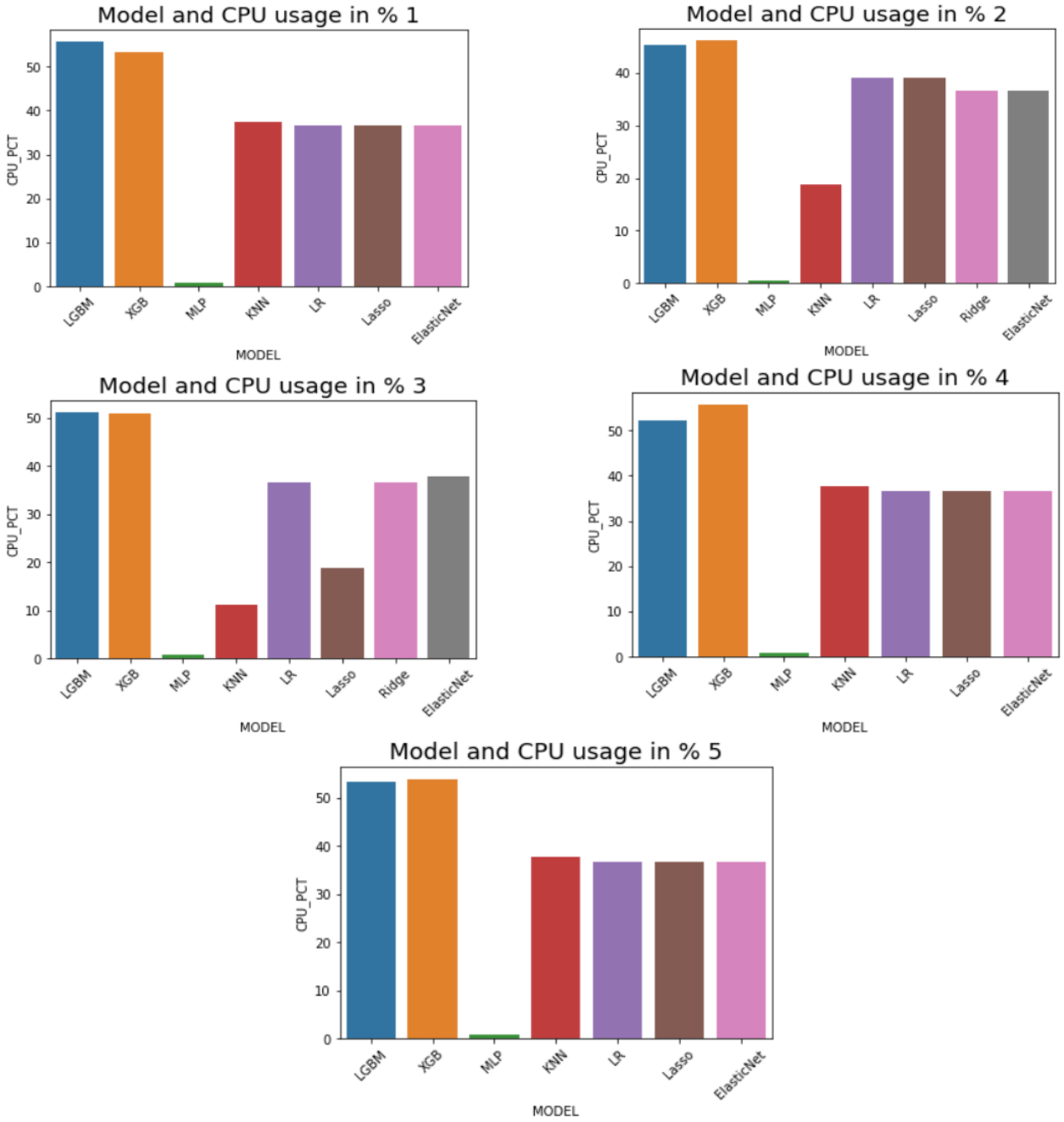

Figure 8: Average CPU usage by algorithm

For visualization and subsequent analysis of processor usage, results with processor usage above zero were selected among the algorithms. There are 2 algorithms at the top for CPU usage in the learning process: LGBM and XGB. The minimum resource consumption, if we do not take

into account zero values, is noticed in the MLP algorithm. Significantly, depending on the data set, the CPU consumption of the KNN algorithm also changed. At the same time, there is no direct relationship between the amount of data and the consumption of resources by this algorithm. The lasso algorithm utilized significantly fewer resources in the dataset with the introduction of a new parameter for the missing values. The rest of the algorithms showed the highest load with a large amount of data replacing the missing values with a mode.

After training all models for 5 subsets of data, the information was recorded in an additional table. Since the dataset size is not large, the training of models took not a long time. These results are not bad and allow the model to predict the math exam scores fairly accurately.

Table 3
Metrics Values

Metrics name	Mean Value	Max Value	Min Value
MAE	5.26	7.55	4.39
RMSE	6.59	9.34	5.49
R2	0.81	0.87	0.64

The spread in accuracy between the algorithms is small. The best average value of R2 among all regression algorithms turned out to be in sample number 5, where rows with empty values and uninformative columns were removed. The average R2 value is 0.827. The worst value turned out to be in sample number 4 where inaccuracies in numerical variables were allowed - 0.77. In accordance with the R2 indicator, we can also see changes in MAE and RMSE. This shows us that as the accuracy of R2 decreased, the values of model errors increased.

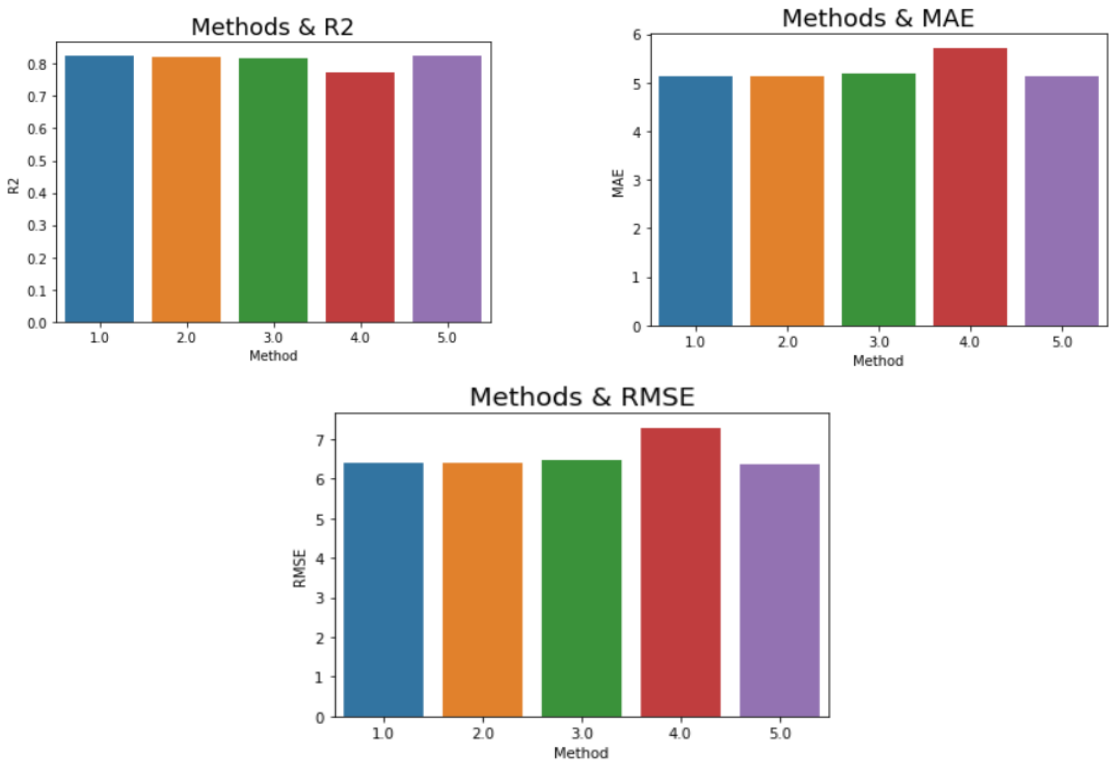

Figure 9: Models' average metrics on every method

If we analyze the accuracy of the models in the context of each algorithm, then we can see on the graphs in Figure 10 that most of them have approximately similar accuracy. Based on the results, we can conclude that the SVR algorithm performed the worst on all sets of data. Also, the algorithms LGBM, XGB, Gradient Boost, Random Forest, Decision Tree, MLP, Cat Boost, Linear Regression, Ridge showed consistently high accuracy. Their R2 score on all datasets was above 0.8. The other algorithms also showed stable results with average accuracy.

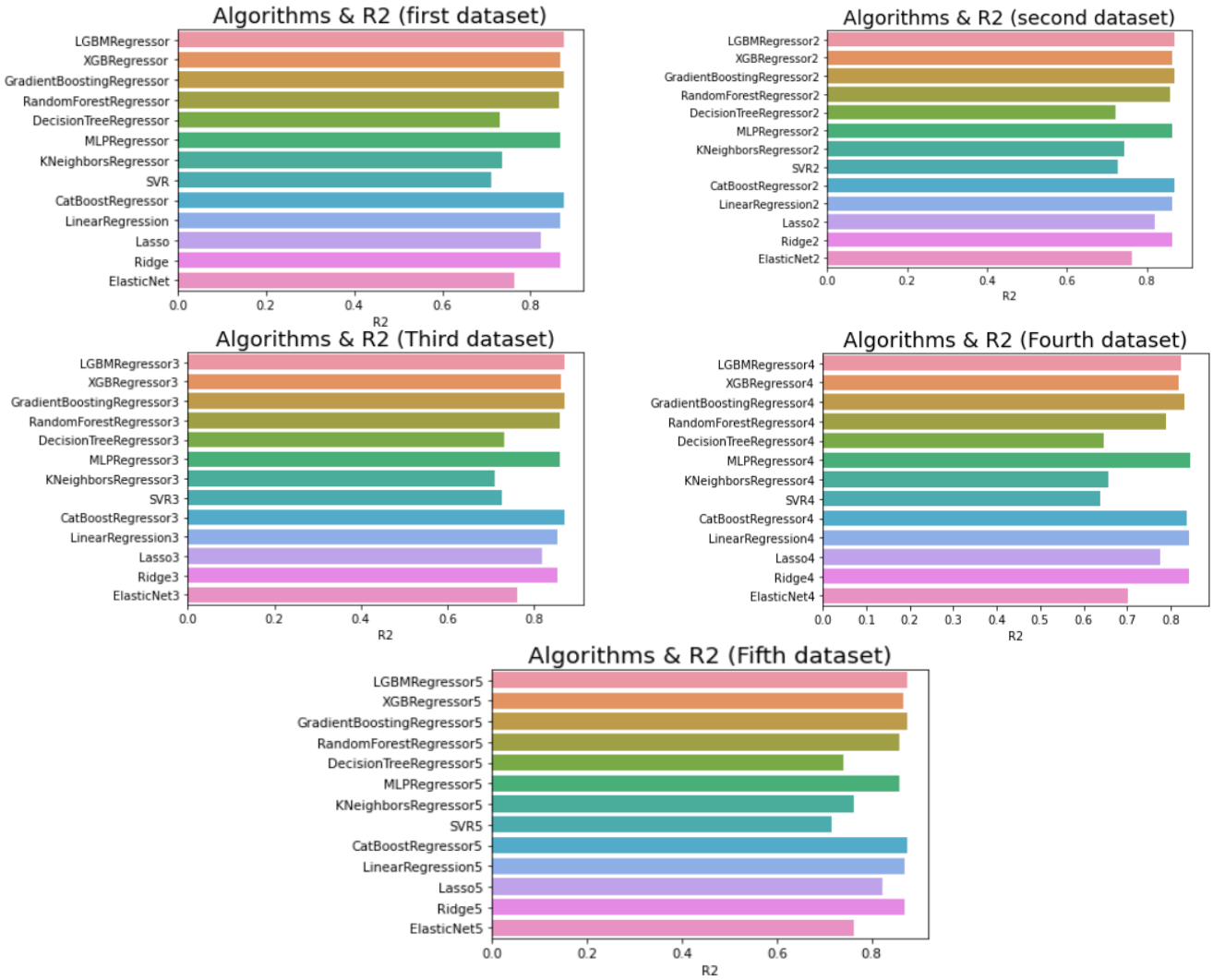

Figure 10: Algorithms R2 score on every dataset

6. Conclusion

While conducting this research study, we considered the problem of the impact of data quality and data preparation techniques on machine learning models. To do this, we found a dataset with errors in the data, analyzed it using visualization methods, applied several different algorithms for data pre-processing, trained models of several machine learning algorithms and compared the main metrics with each other.

As a result, it turned out that for the specified data and for the prediction task, the best result was obtained by removing missing values and uninformative columns. Most of all, missing data in strongly correlated numerical variables have the greatest negative impact on the results of models, regardless of the algorithms of machine learning models. Otherwise, different algorithms of data preprocessing showed similar results on the machine learning model in prediction tasks regardless of the data sets processed.

The CPU load for some of the algorithms varied along with the amount of data being processed. For some of the other algorithms, the amount of data did not affect the consumption of the processor's resources.

Also, we were able to show the importance of data visualization at each of the stages of machine learning model training.

In future works, it is also possible to assess the impact on the consumption of other computing resources, to consider the impact of data preparation methods for solving other tasks, to consider

from what percentage of inaccuracies the influence of missing data increases and also to evaluate the influence of the parameters of individual models on the results of the accuracy of predictions.

7. References

[1] J.M. Conejero, J.C. Preciado, A.J. Fernandez-Garcia, A.E. Prieto, R. Rodriguez-Echeverria. (2021). Towards the use of Data Engineering, Advanced Visualization techniques and Association Rules to support knowledge discovery for public policies, Expert Systems with Applications 170. Doi: 10.1016/j.eswa.2020.114509.

[2] R. Lao, A Beginnner's Guide to Machine Learning, 2018. URL: https://medium.com/@randylaosat/a-beginners-guide-to-machine-learning-dfadc19f6caf.

[3] T. Ram, What is Data Quality in Machine Learning, 2023. URL: https://www.analyticsvidhya.com/blog/2023/01/the-role-of-data-quality-in-machine-learning/.

[4] B. Botelho, S. J. Bigelow, Big Data, 2022. URL: https://www.techtarget.com/searchdatamanagement/definition/bigdata#:~:text=Big%20data%20is%20a%20combination,and%20other%20advanced%20analytics%20applications.

[5] Great Learning Team, 2020. URL: https://medium.com/@mygreatlearning/what-is-artificial-intelligence-how-does-ai-work-and-future-of-it-d6b113fce9be.

[6] N. Soni, E. K. Sharma, N. Singh, A. Kapoor. (2020). Artificial Intelligence in Business: From Research and Innovation to Market Deployment, Procedia Computer Science. Pp. 2200-2210. doi: 10.1016/j.procs.2020.03.272.

[7] K. Maharana, S. Mondal, B. Nemade. (2022). A review: Data pre-processing and data augmentation techniques, Global Transitions, 3. Pp.91-99. doi: 10.1016/j.gltp.2022.04.020.

[8] C.C. Olisah, L. Smith, M. Smith. (2022). Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective, Computer Methods and Programs in Biomedicine 220. doi: 10.1016/j.cmpb.2022.106773.

[9] Q. Zhou, R. Ooka. (2021). Influence of data preprocessing on neural network performance for reproducing CFD simulations of non-isothermal indoor airflow distribution, Energy and Buildings 230. doi: 10.1016/j.enbuild.2020.110525.

[10] J. Huang, Y. F. Li, M. Xie. (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation, 67. Pp.108-127. doi: 10.1016/j.infsof.2015.07.004.

[11] F. Farhangi. (2022). Investigating the role of data preprocessing, hyperparameters tuning, and type of machine learning algorithm in the improvement of drowsy EEG signal modeling, Intelligent Systems with Applications. 15. doi: 10.1016/j.iswa.2022.200100.

[12] I. Aksangur, B. Eren, C. Erden. (2022). Evaluation of data preprocessing and feature selection process for prediction of hourly PM10 concentration using long short-term memory models, Environmental Pollution. 311. doi: 10.1016/j.envpol.2022.119973.

[13] M. Schoot, C. Kapper, G. H. van Kollenburg, G. J. Postma, G. van Kessel, L. M. C. Buydens, J. J. Jansen. (2020). Investigating the need for preprocessing of near-infrared spectroscopic data as a function of sample size, Chemometrics and Intelligent Laboratory Systems. 204. doi: 10.1016/j.chemolab.2020.104105.

[14] S. F. Crone, S. Lessmann, R. Stahlbock. (2006). The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing, European Journal of Operational Research. 173. Pp.781-800. doi: 10.1016/j.ejor.2005.07.023.

[15] M. Kashina, I. D. Lenivtceva, G. D. Kopanitsa. (2020). Preprocessing of unstructured medical data: the impact of each preprocessing stage on classification, Procedia Computer Science. 178. Pp.284-290. doi: 10.1016/j.procs.2020.11.030.

[16] K.V. Balaji, What is Data Visualization and Why Is It Important?, 2020. URL: https://medium.com/analytics-vidhya/what-is-data-visualization-and-why-is-it-important-

3c2ccb108945#:~:text=Data%20visualization%20is%20the%20visual,communicate%20it%20to%20your%20peers.
[17] Kaggle Team, Students Exam Scores: Extended Dataset, 2023. URL: https://www.kaggle.com/datasets/desalegngeb/students-exam-scores/data.
[18] R. Kimmons, Exam Scores dataset, 2012. URL: http://roycekimmons.com/tools/generated_data/exams.
[19] T. Moriarty, The Right Way to Make a Pie Chart, 2013. URL: https://medium.com/eyeful/the-right-way-to-make-a-pie-chart-7852f568eaa9.
[20] K. Bhalla, Bar Charts: What they are, when to use them & Guidelines for creating, 2021. URL: https://medium.com/@komal.bhlla/bar-charts-what-they-are-when-to-use-them-guidelines-for-creating-64f0720a88d1.
[21] S. Wagavkar, Introduction to the Correlation Matrix, 2023. URL: https://builtin.com/data-science/correlation-matrix.
[22] J. Chen, G. Scott, P. Rathburn, How a Histogram Works to Display Data, 2023. URL: https://www.investopedia.com/terms/h/histogram.asp
[23] S. Thomas, What Is the Interquartile Range (IQR)?, 2023. URL: https://articles.outlier.org/what-is-the-interquartile-range.
[24] P. Flom, Outliers: An Introduction, 2019. URL: https://towardsdatascience.com/outliers-an-introduction-e07445c8f430.
[25] LightGBM Core Team, LGBMRegressor, 2023. URL: https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRegressor.html.
[26] XGBoost Core Team, Python API Reference, 2022. URL: https://xgboost.readthedocs.io/en/stable/python/python_api.html.
[27] SKLearn Core Team, Gradient boosted trees, 2023. URL: https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting.
[28] SKLearn Core Team, Random forests and other randomized tree ensembles, 2023. URL: https://scikit-learn.org/stable/modules/ensemble.html#forest.
[29] SKLearn Core Team, Decision Trees, 2023. URL: https://scikit-learn.org/stable/modules/tree.html#tree.
[30] SKLearn Core Team, MLPRegressor, 2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html.
[31] SKLearn Core Team, Nearest Neighbors Regression, 2023. URL: https://scikit-learn.org/stable/modules/neighbors.html#regression.
[32] SKLearn Core Team, SVR, 2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html.
[33] Cat Boost Core Team, CatBoostRegressor, 2023. URL: https://catboost.ai/en/docs/concepts/python-reference_catboostregressor.
[34] SKLearn Core Team, Linear Regression, 2023. URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html.
[35] SKLearn Core Team, Lasso, 2023/ URL: https://scikit-learn.org/stable/modules/linear_model.html#lasso.
[36] SKLearn Core Team, Ridge Regression and Classification, 2023. URL: https://scikit-learn.org/stable/modules/linear_model.html#ridge-regression.
[37] SKLearn Core Team, Elastic-Net, 2023. URL: https://scikit-learn.org/stable/modules/linear_model.html#elastic-net.
[38] S. Manikandan. (2011). Measures of central tendency: Median and mode, J Pharmacol Pharmacother. 2. doi: 10.4103/0976-500X.83300.
[39] H.M. Deitel, B. Deitel, Chapter 3 – The Processor, An Introduction to Information Processing (1986) 46-71. doi: 10.1016/B978-0-12-209005-9.50009-6.
[40] T. O. Hodson, Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not, Geosci. Model Dev., 15 (2022): 5481–5487, doi: 10.5194/gmd-15-5481-2022.
[41] W. Rowe, Mean Square Error & R2 Score Clearly Explained, 2018. URL: https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/.