



Graphical representation of data prediction potential: correlation graphs and correlation chains

Adam Dudáš¹

Accepted: 14 December 2023
© The Author(s) 2024

Abstract

The correlation of the set of attributes is a crucial statistical value for the measuring of prediction potential present in a dataset. The correlation coefficient, which measures the correlation between the values of two attributes, can be used in order to measure the prediction potential between two-element subsets of a dataset containing a high number of attributes. In this way two common summary visualizations of prediction potential in datasets are formed—correlation matrices and correlation heatmaps. Both of these visualizations are focused on the presentation of correlation between pair of attributes but not much more regarding the context of correlations in the dataset. The main objective of this article is the design and implementation of graphical models usable in a visual representation of data prediction potential—correlation graphs and correlation chains—which emphasize the pseudo-transitivity of prediction potential in a dataset.

Keywords Prediction potential · Correlation · Data visualization · Graph · Chain · Data analysis

1 Introduction

The fundamental statistical values such as minimum, maximum or mean are important indicators which can be used to describe a dataset. However, from the point of view of building machine learning models the correlation analysis is a more relevant metric [1].

Let us have two attributes of dataset D labeled as A_1 , A_2 . These attributes correlate when attribute A_1 has prediction potential for attribute A_2 . Such prediction potential—positive (correlation) or negative (anticorrelation)—speaks of the presence of trends and patterns in the dataset and the possibility of building analytical models that work with the data [2, 3]. In the case a dataset contains more than one numerical attribute, we can measure the correlation between the two-element subsets of this dataset [4].

Even though correlation itself is not considered to be a transitive metric, the prediction potential measured by this correlation is. In the case, we can strongly predict the value of attribute A_1 on the basis of the value of attribute A_2 : $A_2 \rightarrow A_1$ (there is a strong correlation or anticorrelation

between these two attributes) and the similar situation is between attributes A_2 and A_3 : $A_3 \rightarrow A_2$, we can use values of A_3 to predict values of A_2 which in turn can be used to predict the values of A_1 : $A_3 \rightarrow A_2 \rightarrow A_1$. In this article, we refer to this phenomenon as the pseudo-transitivity of prediction potential in data.

This article is focused on the design and implementation of graphical models usable in the visual representation of data prediction potential which visualize the pseudo-transitivity of prediction potential in the dataset and can be used in order to find patterns and trends when analyzing data. The article presents the concept and implementation for two visualization models of prediction potential—correlation graphs and correlation chains. The contribution of the research presented in this article can be summarized as follows:

- Presentation of original visualization models which can be used in correlation analysis and subsequent predictive analysis of large and multidimensional datasets. These models are based on principles from the area of graph theory and are called correlation graphs and correlation chains.
- Implementation of proposed graphical representations of predictive potential stored in datasets in the form of freely available Python code.

✉ Adam Dudáš
adam.dudas@umb.sk

¹ Department of Computer Science, Faculty of Natural Sciences, Matej Bel University, Banská Bystrica, Slovakia

- Evaluation of the proposed graphical models built on two datasets - Iris dataset as a representation of standard dataset and original graph property dataset, which consists of a higher number of attributes and records.

The body of the presented article is structured into three sections. In the first section, we present a basic overview of correlation coefficient types, common methods of visualization of correlation in datasets and some interesting modern uses of correlation analysis and prediction potential. The second section of the body of work consists of the description and design of proposed visualization models of correlation graphs and correlation chains while the third section contains an evaluation of the implementation of these models on two datasets of various sizes and structures.

2 Correlation coefficients, matrices and heatmaps

We measure the correlation of two variables using the correlation coefficient $corr(A_1, A_2)$, which represents the amount of how much the attribute A_1 is a function of the attribute A_2 and vice versa [2]. In general, we use three standard methods to analyze correlations and measure correlation coefficients— Pearson correlation coefficient (r), Spearman rank correlation coefficient (ρ) and Kendall correlation coefficient (τ) [5, 6].

The correlation coefficient—measured between two attributes A_1, A_2 —of any type can obtain values from the $[-1, 1]$ interval while:

- 1 indicates the complete correlation of two attributes - in the case, the value of attribute A_1 increases, the value of attribute A_2 also increases. If there is a complete correlation between the values of two attributes, we identify a strong prediction potential and thus these attributes are suitable for mutual prediction.
- 0 indicates the worst situation from the point of view of the correlation of two values, which we refer to as non-correlation. In the case when the correlation coefficient between two attributes is close or equal to 0, these are independent values that are unusable from the point of view of building analytical models.
- -1 is the opposite of complete correlation, called complete anticorrelation. In this case, we can identify a trend in which as the value of attribute A_1 increases, the value of attribute A_2 decreases, or vice versa. As in the case of complete correlation, this is a satisfactory condition for building analytical models.

The correlation coefficient shows how much it is possible to predict the values of attribute A_2 in the selected data sample

based on attribute A_1 . The closer the value of this coefficient is to the extremes of the considered interval (that is, to the value 1 or -1), the more suitable the given attribute A_1 is for predicting the values of the optional attribute A_2 [2].

Most of the time, we speak of two attributes as strongly correlated when the value of the correlation coefficient measured between them reaches values higher than 0.8. There is a strong anticorrelation between the two attributes if the correlation coefficient reaches a value lower than -0.8 . This limit of acceptability of the prediction potential can be relaxed closer to the values of 0.7 or -0.7 , but more is not recommended [7].

2.1 Pearson correlation coefficient

Pearson correlation coefficient is focused on the linear prediction of values and the relationship between attributes A and B . It is described using the following relation [8]:

$$r = \frac{\sum_{i=1}^n (A_i - \mu(A))(B_i - \mu(B))}{\sqrt{\sum_{i=1}^n (A_i - \mu(A))^2} \sqrt{\sum_{i=1}^n (B_i - \mu(B))^2}} \quad (1)$$

where $\mu(A)$ is the mean of attribute A , similarly $\mu(B)$ is the mean value of attribute B , and n is the number of measurements (vertical size of the dataset). The obvious dependence on the mean value of attributes brings the biggest disadvantage of the Pearson correlation coefficient - sensitivity to outliers.

Using the Pearson correlation coefficient, we are looking for a linear function that describes the values of the given attributes. Therefore, the Pearson correlation coefficient can be used when attributes A and B contain linear relationships, normal (Gaussian) distribution and no outliers.

2.2 Spearman rank correlation coefficient

As a way of dealing with datasets that contain nonlinear relationships with outliers, we can use a different type of correlation coefficient—specifically the Spearman rank correlation coefficient. This method of measuring the correlation between attributes creates a hierarchy (ranking) of individual attribute values for its functionality.

$$\text{IF } A_1 > A_3 > A_0 > A_2 \text{ THEN } \text{rank}(A_1) = 1, \\ \text{rank}(A_2) = 4, \text{ and so on} \quad (2)$$

In this way, we actually measure the monotonicity of the values within the attribute, and therefore we can say that the Spearman rank correlation coefficient is most suitable for datasets with monotonic relationships between attributes—in the case one attributes value increases, the other never

decreases or vice versa. On the other hand, this type of correlation coefficient is not recommended to be used if there are repeated values (same rank) in the dataset. This effect is attenuated with increasing dataset size. Spearman rank correlation coefficient is computed as [9]:

$$\rho = 1 - \frac{6 \sum (rank(A_i) - rank(B_i))^2}{n(n^2 - 1)} \tag{3}$$

where A_i and B_i are considered attributes of the dataset and n is the number of measurements of these attributes in the dataset.

2.3 Kendall correlation coefficient

Similar to Spearman Rank Correlation Coefficient, Kendall (Rank) Correlation Coefficient is a nonparametric measure of the correlation between two attributes in a dataset which is not dependent on the normal distribution of data or outliers. Other than ranking the attributes as in the previous case, all combinations of ranks are considered and the number of so-called concordant and discordant pairs of ranks is used for the computation of the coefficient [10]:

$$\tau = \frac{n_c - n_d}{\frac{n(n-1)}{2}} \tag{4}$$

where n_c is the number of concordant ranking pairs, n_d is the number of discordant ranking pairs and n is the number of measurements of considered attributes. Concordance of a combination of ranking pairs can be described as monotonicity of rankings in the combination (when $rank(A_1)$ in the first ranking of given pair descends in comparison to $rank(B_1)$, then $rank(A_1)$ in second ranking in the pair also descends in comparison to $rank(B_1)$ and vice versa). Discordance is the opposite situation to concordance.

2.4 Correlation matrices and heatmaps

Generally, we need to measure the correlation between larger sets of attributes. Therefore, for a dataset of size n we need to measure $(n(n - 1))/2$ correlation coefficient values. For the presentation of these measurements, we use a correlation matrix - a table containing a correlation coefficient measured between all possible pairs of attributes in the dataset. In Table 1, we see the correlation coefficient measured between the attributes A_1, A_2, \dots, A_n .

This matrix has two natural properties—it is symmetric along the diagonal and this diagonal always contains the values of the correlation coefficient equal to 1 - the correlation of the i -th attribute A_i with itself is always $corr(A_i, A_i) = 1$ regardless of the method used, which is natural since the

Table 1 Example of correlation matrix for dataset containing attributes A_1, A_2, \dots, A_n

	A_1	A_2	..	A_n
A_1	$corr(A_1, A_1)$	$corr(A_1, A_2)$..	$corr(A_1, A_n)$
A_2	$corr(A_2, A_1)$	$corr(A_2, A_2)$...	$corr(A_2, A_n)$
..
A_n	$corr(A_n, A_1)$	$corr(A_n, A_2)$...	$corr(A_n, A_n)$

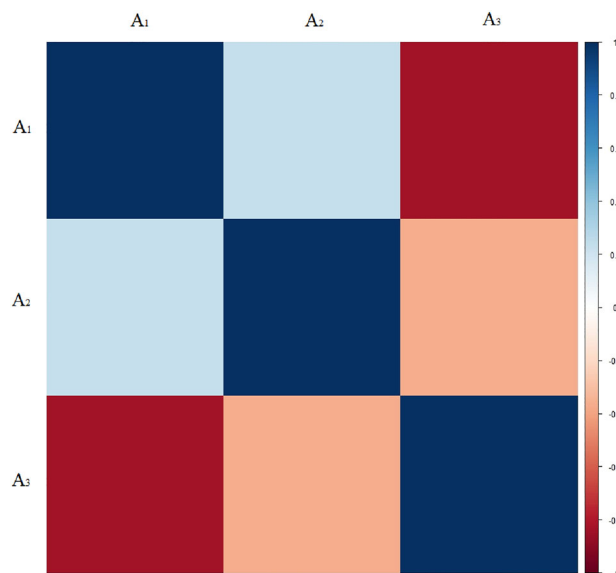


Fig. 1 Example of correlation heatmap for dataset containing attributes A_1, A_2, A_3

value of attribute A_i is always fully dependent on the value of attribute A_i .

It is obvious that for a dataset composed of dozens of attributes, such a matrix would be confusing and hard to read. Therefore it is often replaced by a correlation heatmap or correlation plot (see Fig. 1).

Such a correlation heatmap is a simple projection of the correlation matrix into a color grid, in which the color of the field is defined by the value of the correlation coefficient for the given pair of attributes [6]. For better readability, the scale (right) containing the interval of possible values for the correlation coefficient is indicated in the correlation heatmap. Instead of looking for numbers close to the extremes of the $[-1, 1]$ interval in the correlation matrix, in the correlation heatmap, we look for dark red or dark blue grid fields that indicate the same property and are easier to identify.

Both of these methods have some insufficiencies [11, 12]:

- The readability and interpretability of correlation values lowers with the growing horizontal size of the given dataset. This lowering of readability is becoming present on quite small sizes of correlation matrices. The corre-

lation heatmap holds its readability until we work with circa two to three dozen attributes.

- In both methods it is hard to find sets of correlated attributes. The correlation matrix and correlation heatmap are both focused on the presentation of correlation between a pair of attributes but not much more.
- An Additional weakness of correlation heatmap is that this method is not standalone. Even though we can use it to identify correlation extremes of given datasets or to find strong correlations and anticorrelations, it is common to use a correlation matrix in order to clear some ambiguities in correlation coefficient values.

2.5 Uses of correlation analysis and prediction potential identification

Modern approaches to correlation analysis and the use of predictive potential stored in multidimensional datasets presented in the literature vary significantly. There is number of interesting and novel correlation analysis approaches in the form of either the use of novel methods and visualizations proposed for the problem or the use of powerful methods in interesting problems and areas of expertise.

In [13], multi-label image classification is studied with the use of a novel channel correlation network which is fully based on a convolutional neural network using correlation analysis approaches to measure prediction potential between individual labels assigned to the image. The authors propose a new module for the convolution of the image features obtained by convolutional neural network in order to obtain the correspondence between the label and the channel-wise feature map. Then, number of transformations is used to eliminate the irrelevant information to better explore the label correlation.

Authors of [14] propose an efficient method for stereo matching called area-based correlation and non-local attention network. The proposed method, which is based on correlation maps, uses area-based correlation to capture more local similarity in cost volume. Identified correlated features are then, transformed into a four-dimensional area-based cost volume.

In order to improve the performance of shallow convolutional neural networks, authors of [15], designed two methods: Weight Correlation Reduction and Features Normalization. The formal method is designed to eliminate weight redundancy, while the latter is used to increase the sparsity of learned deep features.

Interesting results were reached in [16], authors of which work on extending the notion of High-Utility Pattern Mining, specifically, authors introduce a new framework that allows for novel classes of utility criteria. As part of the research, the authors also present support for recent extensions of Answer

Set Programming with external functions for fast and effective encoding and testing of the new framework.

The study presented in [17] is of great interest to the research presented in this article. Authors of this research explore three techniques that are representative of different strategies to visualize correlations in multivariate data: either juxtaposing all locations for a given time step, or juxtaposing all time steps for a given location; and encoding thematic attributes either using symbols overlaid on top of map features, or using visual channels of the map features themselves. The study presents visualization's effectiveness depending on the task to be carried out. Based on the findings authors present a set of design guidelines for geo-temporal visualization techniques and identification of prediction potential in such data.

On the other hand the work [18] focuses on visualization of prediction potential stored in datasets through a correlation color map of the transformed or pseudodata used to show clusters of correlated variables in the task of identification of true and false alarms. In the correlation color map correlation and redundancy information can be easily found and used to improve the alarm settings, and statistical methods such as singular value decomposition techniques can be applied within each cluster to help design multivariate alarm strategies.

3 Correlation graphs and correlation chains

Graph G is described as pair of sets V (vertices) and E (edges) while [19]

$$G = (V, E), E \subseteq V^2 \quad (5)$$

This type of structure is ideal for any transitive or pseudo-transitive phenomena where the direct and indirect influence of values can be evaluated. When working with a correlation of two attributes, with the use of graphs, we can see the direct and indirect correlation (and prediction potential) influence of sets of attributes on one of the attributes in a dataset.

In the case, we are constructing a complete graph for a dataset containing n attributes the graph consists of n vertices corresponding to individual attributes and $(n(n-1))/2$ edges between these vertices while each edge contains correlation coefficient value measured between the two considered vertices. Such a graph would be - even with a comparatively low number of attributes (vertices) - potentially hard to read. Since this work is focused on prediction potential the task-specific two-phase pruning of such a graph is proposed:

Phase 1: Visualization of edges with maximal correlation values - since we are focused on the prediction potential of

Fig. 2 Visualization of edges with maximal correlation values

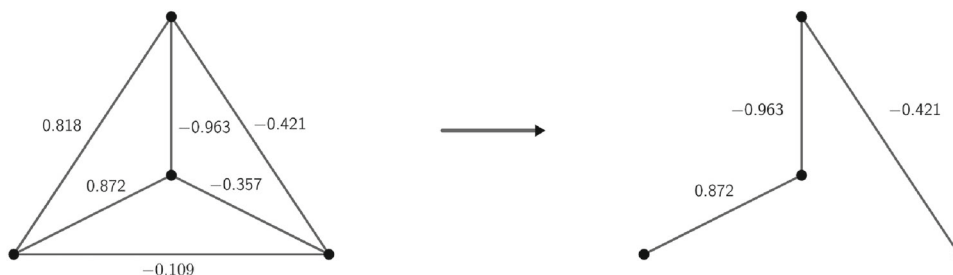


Table 2 The value of σ border in relation to the relationship explored in dataset

Explored relationship	σ value
Weak relationship	$\sigma \leq 0.3$
Moderate relationship	$\sigma \leq 0.7$
Strong relationship	$\sigma \leq 1$

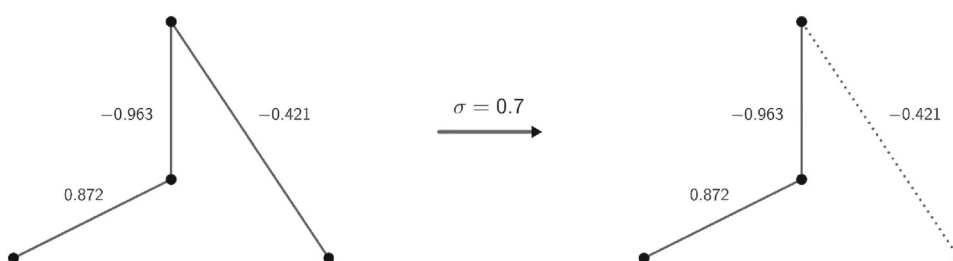
attributes in the dataset and relationships between these prediction potentials, the correlation graph contains only edges which maximize the correlation coefficient values between each of the (at the moment unused) attributes and all other attributes of the dataset (see Fig. 2).

Phase 2: Since some of the attributes are generally weakly correlated with the rest of the dataset, in the second phase of correlation graph pruning, we can take into account only edges containing correlation values greater than the set border. This border can be set to a value from the $[0, 1]$ interval, depending on the level of correlation coefficient values in the dataset we are examining. From a statistical point of view, we can identify three levels of correlation coefficient values [20]:

- Values between 0 and 0.3 (0 and -0.3) indicate a weak relationship.
- Values between 0.3 and 0.7 (0.3 and -0.7) indicate a moderate relationship.
- Values between 0.7 and 1.0 (-0.7 and -1.0) indicate a strong relationship.

Based on these levels of correlation coefficient values, it is recommended to set the σ parameter to the appropriate values listed in Table 2.

Fig. 3 Selection of edges containing correlation value greater than set border ($\sigma = 0.7$)



For this pruning, see Fig. 3 while the σ parameter denotes the set border of the absolute value of correlation.

Therefore, correlation graph G is undirected, weighted graph, which for dataset D containing n attributes consists of:

- n vertices, each corresponding to one of the attributes from D ,
- at least one connected edge incident with each of n vertices - this edge denotes the strongest correlation or anticorrelation between the chosen pair of vertices (attributes of D).

The undirected, weighted subgraph of G containing a subset of vertices of G incident with edges, where correlation is greater than σ or anticorrelation is lower than $-\sigma$ is called the correlation chain of the graph (Fig. 4).

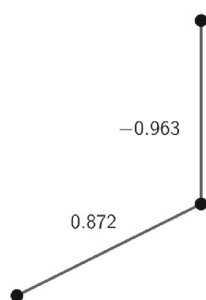
4 Evaluation of proposed graphical model

This section of the article is focused on the evaluation of the concept of correlation graphs and chains on two datasets of varying size and structure with the criteria of interpretability of visualization, self-sufficiency of visualization, and visualization of pseudo-transitivity of the prediction potential.

The proposed visualization model is implemented in the software form for Python language with the following user-defined input parameters:

- Dataset - the first of input parameters consists of two subparameters - the name of the file containing the dataset of interest and the separator symbol used in the file.

Fig. 4 Correlation chain of example correlation graph with $\sigma = 0.7$



- **Border(σ)** - for the implementation of Phase 2 of the proposed correlation graph pruning, the software needs information about the border of acceptability of correlation measured between attributes of the input dataset.
- **Method of measuring correlation coefficient** - the proposed model is capable of working with the Pearson correlation coefficient, Spearman rank correlation coefficient and Kendall correlation coefficient.
- **Type of model** - there are two graphical representations in the created package, either correlation graph or correlation chain.
- **Attribute of interest** - in order to identify the direct and indirect correlation (and prediction potential) influence of sets of attributes on one of the attributes in a dataset, the software expects a label of one interesting attribute (often called descriptor variable).

In the evaluation of the proposed model, we focus on three criteria for the visualization relevant from the point of view of this article. These three criteria are compared on four considered visualization methods—correlation matrix, correlation heatmap, correlation graph and correlation chain. The criteria are:

- **Interpretability of the visualization** which combines evaluation of clarity, readability and scalability of visualization. In this way, we explore the ability of users to understand the information presented in the visualization, while assessing how well the visualization model works with large datasets.
- Through **self-sufficiency of the visualization**, we evaluate whether the visualization model in question can be used as a standalone tool for specific tasks.
- **Visualization of pseudo-transitivity of the prediction potential** can be understood as a domain-specific criterion. This property of a visualization model is focused on assessing its ability to visualize the local and global correlation structure of the explored dataset.

In the conclusion of this section, we present strengths, weaknesses and feedback on the proposed model from a small user sample.

Table 3 Correlation matrix for the Iris dataset

	Sepal_W	Sepal_L	Petal_W	Petal_L	Class
Sepal_W	1	-0.11	0.87	0.82	0.78
Sepal_L	-0.11	1	-0.42	-0.36	-0.42
Petal_W	0.87	-0.42	1	0.96	0.95
Petal_L	0.82	-0.36	0.96	1	0.96
Class	0.78	-0.42	0.95	0.96	1

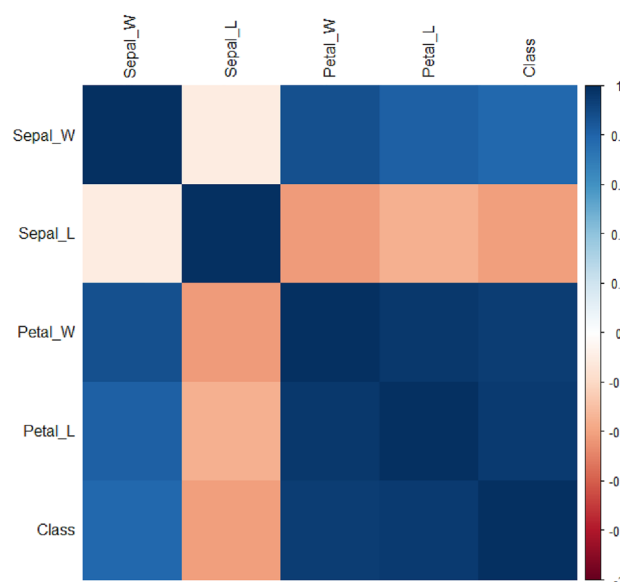


Fig. 5 Correlation heatmap for Iris dataset

4.1 Correlation graphs and chain constructed on iris dataset

The first dataset used for purposes of testing of the proposed models was the Iris dataset [21]. This dataset is one of the standard tools for the evaluation of machine learning and data analysis models.

The dataset consists of five attributes measured over 150 entities - in this case, Iris flower individuals. Attributes contained in this dataset are composed of measurements of the length (denoted by L) and width (denoted by W) of two types of leaves of the Iris flowers (Sepal_W, Sepal_L, Petal_W and Petal_L) and Class attribute for classification of the flowers into one of three considered species [22].

The correlation matrix of the Pearson correlation coefficient is presented in Table 3 and its projection into correlation heatmap is in Fig. 5.

Both of these visualizations of correlations and prediction potential are quite readable which is caused mainly by the small size of the dataset. We can see that the correlation heatmap itself is a little less interpretable since there are a lot of similar correlation coefficient values in this dataset. This correlation matrix would be difficult to interpret in printed

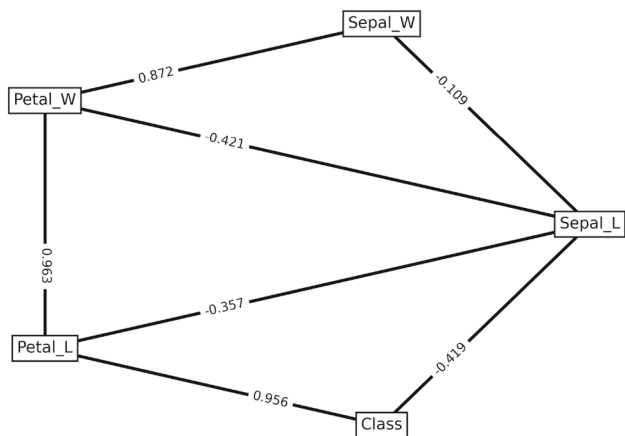


Fig. 6 Correlation graph constructed on Iris dataset (full graph)

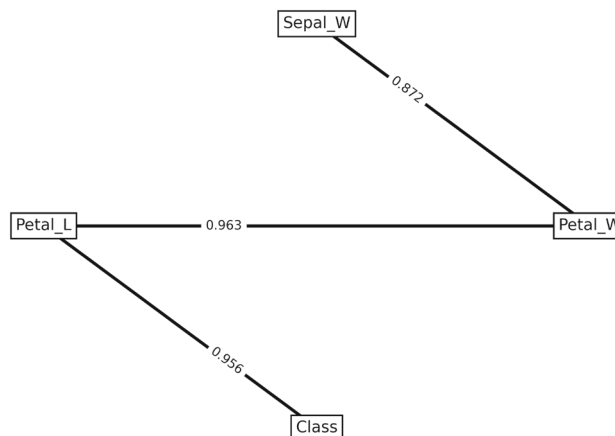


Fig. 8 Correlation chain constructed on Iris dataset - $\sigma = 0.8$

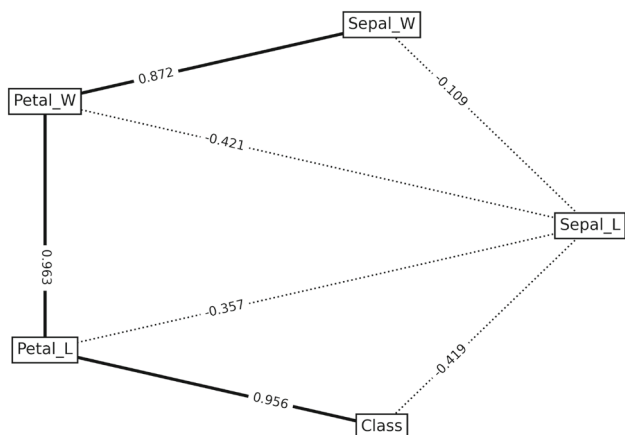


Fig. 7 Correlation graph constructed on Iris dataset - $\sigma = 0.8$ (strong correlation/anticorrelation subgraph)

```
['Sepal_L', 'Petal_L']
Indirect correlation influence:
['Petal_W', 'Sepal_W']
```

```
Correlation chain for 'Class' (border = 0.8):
Direct correlation influence: ['Petal_L']
Indirect correlation influence: ['Petal_W']
```

This output contains information on which attributes of the dataset influence the value of *Class* attribute directly - Petal length in correlation chain complemented by Sepal length in correlation graph. For the computed correlation chain, the indirect influence on the *Class* attribute is measured on the Petal width attribute. For the correlation graph, this indirect influence consists of Petal width and Sepal width.

form or when presented on lower-resolution projectors or monitors.

Figure 6 presents a correlation graph for the Iris dataset with no set border. This graph contains all five attributes of the dataset with the strongest correlation measured for each of the present attributes. It is evident, that from the point of view of prediction potential analysis, the Sepal Length (*Sepal_L*) attribute is unusable - the correlation between the values of this attribute and the rest of the dataset is unsatisfactory.

In Fig. 7 we present a correlation graph of the dataset with the border set to $\sigma = 0.8$. Figure 8 shows the correlation chain extracted from the correlation graph of the Iris dataset.

Graphical representation of the prediction potential in the dataset is complemented by the identification of direct and indirect correlation influences for the chosen attribute in the dataset. For the purposes of evaluation, the *Class* attribute was selected to be of interest. The output of the implemented package consists of the following:

```
Correlation graph for 'Class' (no border):
Direct correlation influence:
```

4.2 Correlation graphs and chain constructed on cubic graph property dataset

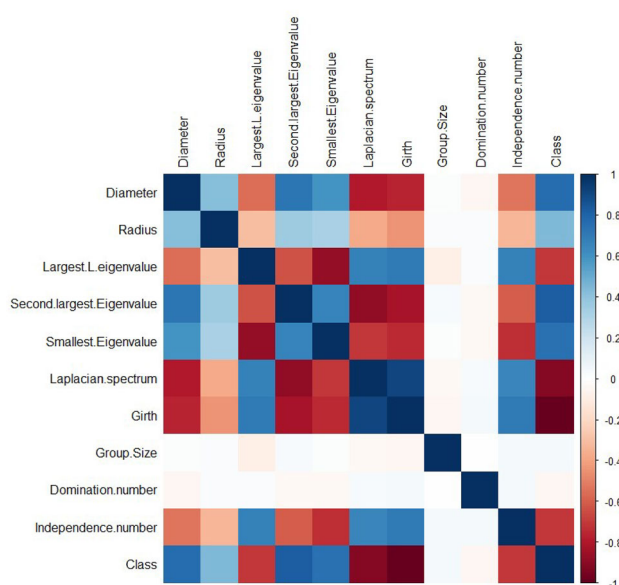
The second dataset used for the evaluation of the proposed model is a dataset containing measurements of properties of 36-vertex cubic graphs. The dataset contains 500 measurements, each containing 11 graph properties [23].

Table 4 contains the Pearson correlation matrix of the cubic graph property dataset. For better readability, we rounded up the correlation coefficient numbers to two decimal points and shortened the labels of attributes from this dataset as follows: Diameter (D), Radius (R), Largest L Eigenvalue (LLE), Second Largest Eigenvalue (SLE), Smallest Eigenvalue (SE), Laplacian spectrum (LS), Girth (G), Group size (GS), Domination number (DN), Independence number (IN) and Class (C).

Other than seeing the elements on the diagonal of this matrix are always equal to 1, the symmetry of the matrix, and being able to find some simple maxima or minima in individual rows/columns there is not much knowledge to be gotten from this type of correlation matrix. This is the reason

Table 4 Correlation matrix for the dataset of cubic graph properties

	D	R	LLE	SLE	SE	LS	G	GS	DN	IN	C
D	1	0.42	-0.56	0.72	0.59	-0.8	-0.77	0.01	-0.04	-0.54	0.77
R	0.42	1	-0.31	0.36	0.32	-0.38	-0.45	0.03	0.03	-0.33	0.45
LLE	-0.56	-0.3	1	-0.63	-0.87	0.68	0.7	-0.09	0.03	0.67	0.7
SLE	0.72	0.36	-0.63	1	0.66	-0.88	-0.83	0.03	-0.03	-0.61	0.83
SE	0.59	0.32	-0.87	0.66	1	-0.71	-0.74	0.01	-0.03	-0.73	0.74
LS	-0.8	-0.38	0.68	-0.88	-0.71	1	0.91	-0.03	0.04	0.65	-0.91
G	-0.77	-0.45	0.7	-0.83	-0.74	0.91	1	-0.04	0.04	0.71	-1
GS	0.01	0.03	-0.09	0.03	0.01	-0.03	-0.04	1	0.00	0.05	0.04
DN	-0.05	0.03	0.03	-0.03	-0.03	0.04	0.04	0.00	1	-0.05	-0.05
IN	-0.54	-0.33	0.67	-0.61	-0.73	0.65	0.71	0.05	-0.05	1	-0.71
C	0.77	0.45	0.7	0.83	0.74	-0.91	-1	0.04	-0.05	-0.71	1

**Fig. 9** Correlation heatmap for cubic graph property dataset

for the common interpretation of the correlation matrix in the form of a correlation heatmap presented in Fig. 9

Even though the correlation heatmap is much more interpretable and readable when compared to the correlation matrix, the aforementioned weaknesses of this method are still present. The readability of the heatmap lasts up to a certain number of attributes and even for the heatmap of the size presented in Fig. 9 the plot is not self-sufficient. For precise reading of the information, we need to complement the heatmap with the correlation matrix.

Figure 10 contains a correlation graph for the graph property dataset with no border set yet ($\sigma=0$). As described in Sect. 2 this graph contains all attributes of the dataset with the strongest correlation for each of the attributes. As evident from the graph itself, there is a number of weakly correlated attributes in the dataset. These attributes and their relation-

ship to the rest of the dataset can be highlighted with the use of higher value for the border.

Figure 11 presents a correlation graph with two subgraphs divided by σ parameter set to 0.7. Full edges of the graph denote correlation $\geq \sigma$ or $\leq -\sigma$ between two attributes of the dataset, dotted edges denote other (potentially less interesting) correlations in the dataset.

After the removal of correlations which do not satisfy the set σ border condition, we extract the correlation chain as proposed in Sect. 2. Figure 12 presents the correlation chain of the graph property dataset for $\sigma = 0.7$.

Other than these graphical representations of the prediction potential in the dataset proposed function also identifies direct and indirect correlation influences for the chosen attribute in the dataset. For evaluation purposes, we chose the attribute of *Class* to be of interest. The output of the implemented software consists of the following:

Correlation graph for 'Class' (no border):

```
Direct correlation influence:
    ['Second largest Eigenvalue', 'Girth']
Indirect correlation influence:
    ['Diameter' 'Domination number'
     'Group Size' 'Independence number'
     'Laplacian spectrum'
     'Largest L-eigenvalue' 'Smallest Eigenvalue']
```

Correlation chain for 'Class' (border = 0.7):

```
Direct correlation influence:
    ['Second largest Eigenvalue', 'Girth']
Indirect correlation influence:
    ['Diameter' 'Independence number'
     'Laplacian spectrum']
```

From this output, we can identify attributes, which directly influence the value of *Class* attribute - Second Largest Eigenvalue and Girth of a graph - and attributes the value of which influences these two attributes - Diameter of a graph, Independence number of a graph, the Laplacian spectrum of

Fig. 10 Correlation graph constructed on cubic graph property dataset (full graph)

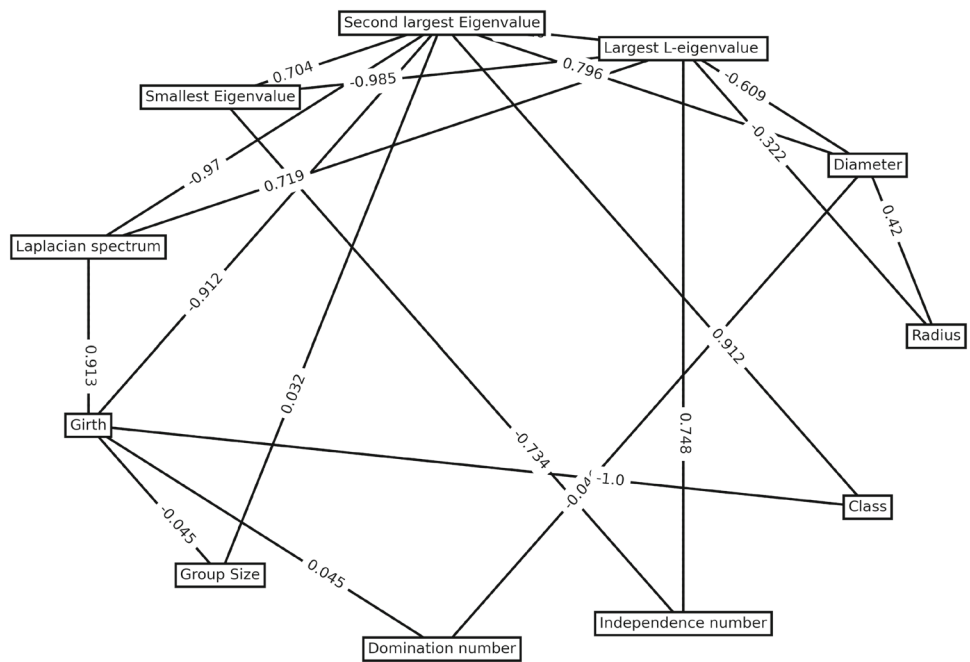
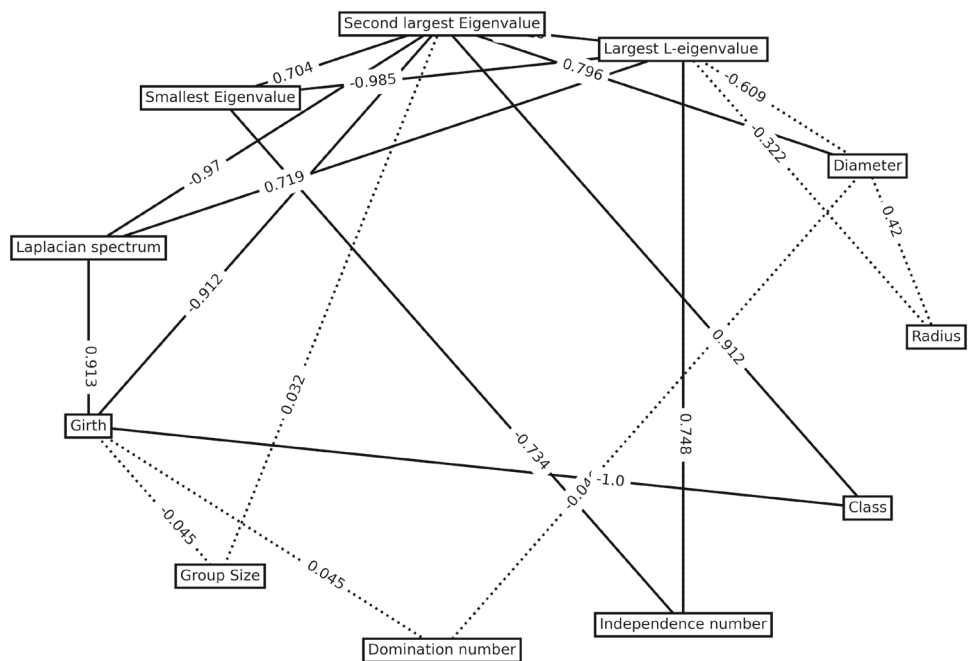


Fig. 11 Correlation graph constructed on cubic graph property dataset - $\sigma = 0.7$ (strong correlation/anticorrelation subgraph)



a graph for created correlation chain with the addition of Domination number, Group size, Largest L-eigenvalue and Smallest eigenvalue of an adjacency matrix of a graph for correlation graph.

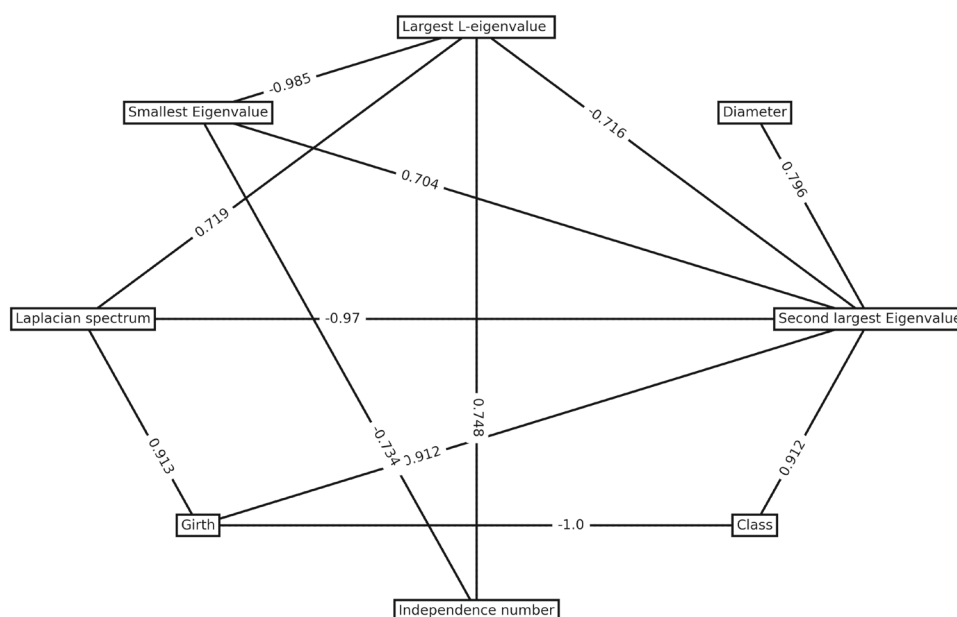
4.3 Comparison of correlation visualization methods

As can be seen in the previous subsections of this section, each of the prediction potential visualization methods has its

strong and weak aspects. We focused on three criteria for visualization relevant from the point of view of this article - the most important of these criteria is the interpretability of the visualization itself which is complemented by the self-sufficiency of the visualization, and since we are focusing on the prediction potential itself, the visibility of pseudo-transitivity of the prediction potential.

Table 5 contains the comparison of the correlation matrix, correlation heatmap and proposed graphical representations

Fig. 12 Correlation chain constructed on cubic graph property dataset - $\sigma = 0.7$



of prediction potential evaluated by a three-level scale measured from LOW (worst) to HIGH (best).

With the increasing size of the dataset examined for possible predictive analysis, the interpretability of the results obtained using conventional methods of visualization of predictive potential decreases significantly. Even in the case of the datasets examined in this study, we can see that the interpretability of the correlation matrix and correlation heatmap decreases drastically with the increasing number of dimensions (attributes) of the dataset.

Proposed models provide a different view of the studied dataset, with the help of which we can maintain the interpretability of the visualization at a good level for a much higher number of attributes present in the studied datasets. In this respect, the correlation chain is the most readable visualization of the prediction potential in the datasets. This is caused by the advantage of a certain dimensionality reduction compared to the other methods - within this method, not all attributes contribute to the resulting visualization.

Regarding the self-sufficiency of the visualization model used to identify predictive potential in datasets, all compared methods offer satisfactory results except the correlation heatmap. This model can be used to quickly explore the values of correlation coefficients and prediction potential in datasets, but without the addition of one of the other models, it is not possible to make specific decisions based on this visualization. Since the correlation matrix by itself is very difficult to interpret in large datasets, it is advisable to use the correlation graph or correlation chain model, which implement the readability of graphs to represent the predictive potential stored in the data being studied.

Pseudo-transitivity of prediction potential is very difficult to follow with the use of conventional methods for visualization of values of correlation coefficients. This fact is natural since these visualizations are a form of summarizing the values of the prediction potential in the examined datasets. Nowadays, however, it is necessary to be able to determine the possibility of predicting the values of two completely independent, not even correlated, attributes in multidimensional datasets. Therefore, visualizing relationships and sequences of strong correlation coefficient (and prediction potential) values in the dataset is vital for modern correlation analysis with the objective of predictive data analysis. Such visualization of this prediction potential pseudo-transitivity is native to the proposed models.

4.4 User evaluation of correlation graph and correlation chain model

As an addition to the evaluation of the proposed correlation graph and chain model from the point of view of interpretability, self-sufficiency and visibility of (pseudo)transitivity of prediction potential presented in the previous subsection of the article, we focus on user evaluation of the model.

For the needs of user evaluation of the work, we were able to secure a sample of ten individuals of two different experience levels:

- Five computer scientists working in the field of data processing and analysis, big data, artificial intelligence, and high-performance computing.

Table 5 Comparison of correlation visualization methods

	Interpretability (Iris)	Interpretability (Graph property)	Self-sufficiency	Pseudo-transitivity
Matrix	Middle	Low	High	Low
Heatmap	Middle	Low	Low	Low
Graph	High	Middle	High	High
Chain	High	High	High	High

Table 6 Frequency table of the user answers regarding potential use of the correlation graph or chain model

Answer	Frequency	Explanation of the answer
Yes	6	–
Maybe	3	–
No	1	User does not use correlation analysis in their research

- Five students of computer science in the master’s degree studies focused on the analysis of large and multidimensional data.

The evaluation of the presented tool was carried out using the compilation of correlation graph and chain with various settings for σ border with ongoing interpretation and description of the used concepts. After the presentation of the model (and in several cases the independent work of the user), we asked the users two open questions and asked for feedback for the tool:

- What are the strengths and weaknesses of the correlation graph/chain model in your opinion?
- Would you use the presented model in your research (for computer science researchers) or your master’s degree thesis (for students)?
- Please provide any feedback for improvement of the model.

Since the answers to the questions were anonymous and the questions were open, we present a summary of the answers without knowledge of whether the respondent is a computer scientist or a computer science student:

- **Strengths of the correlation graph/chain model:** Low time complexity (comparable with correlation heatmap), ease of use, visibility of transitivity of correlation coefficient values, reduction of dimensionality of the dataset.
- **Weaknesses of the correlation graph/chain model:** dependence on the user-defined border of correlational acceptability, (unnecessary) intersection of some components of the graph.
- **Usability of the model.** This question was posed as open, but the majority of the users chose to answer with YES, MAYBE or NO only. There was only one answer con-



Fig. 13 Proposed change for correlation graph and chain node visualization

taining an explanation of the answer. The distribution of answers is presented in Table 6.

- **Feedback:** In larger datasets, the labels and components of the compiled graph intersect. One of the participants proposed to modify the visualization of the model in the manner indicated in Fig. 13.

The feedback from the user exploration of the model was incorporated and the node visualization was updated as proposed in Fig. 13. Wrapping the name of the attribute in the node of the graph to some extent reduced the number of unnecessary intersections of parts of the correlation graph/chain with its labels.

The necessity for a user-defined correlation border (σ) is at the moment still present in the model. This border can be computed in number of ways - either by the selection of correlation value from Table 2 or, in the case, the user needs to identify the strongest correlation and anticorrelation values in the dataset, as a third quartile of the correlation coefficient values for the attributes of interest:

$$\sigma = \frac{\mu(\text{corr}(A_i, A_j)) + \max(\text{corr}(A_i, A_j))}{2} \tag{6}$$

where A_i and A_j are i -th and j -th attributes in the studied dataset, $\mu(\text{corr}(A_i, A_j))$ is the mean value of the correlation coefficient for these attributes and $\max(\text{corr}(A_i, A_j))$ is the maximal value of correlation coefficient for the studied attributes.

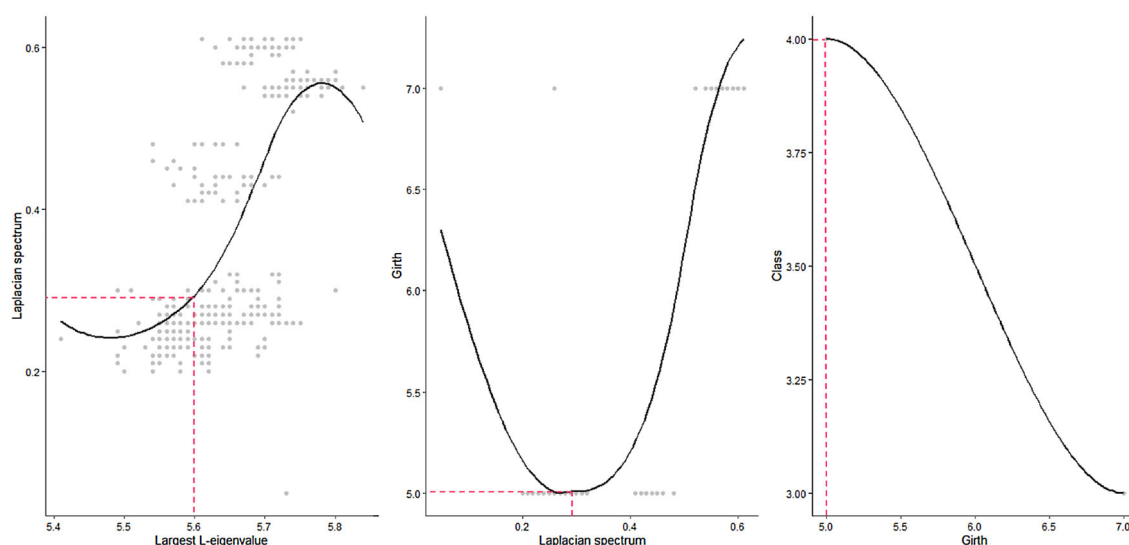


Fig. 14 Regression curves for the subset of graph properties in the cubic graph property dataset

5 Conclusion

The method of visualization of prediction potential stored in a dataset with the use of correlation graph and correlation chain models was proposed and implemented as a Python programming language software. This model uses the concept of pseudo-transitivity of prediction potential to identify direct and indirect correlation influences of attribute values among each other.

Other than the design and implementation of this method, we compared the visualization of two differently sized and structured datasets with the use of conventional visualization methods, such as correlation heatmap, with the proposed methods. We conclude, that the model of correlation graphs and correlation chains is more interpretable than standard visualization techniques in the area while being completely self-sufficient. The interpretability and self-sufficiency of visualization hold with the growing size of the dataset.

The use of correlation graphs and chains in the context of predictive data analysis can be demonstrated on a regression problem in the cubic graph property dataset from Section 4.2. The most important of the graph properties in the dataset is the attribute *Class* - the value of this attribute marks whether the graph is edge 3-colorable or not. The computation of edge 3-colorability of a graph is one of the well-known instances of NP-complete problems and, therefore, is fitting for precise approximation instead of standard computation of the property value.

The edge 3-colorability of a cubic graph is measured by chromatic index property (marked as *Class* in the dataset), which can acquire value 3 or 4 based on the number of colors needed for the edge coloring of a specific cubic graph. This number can be estimated with the use of a regression model.

When applying regression models, we use strong correlation and anticorrelation values for the determination of fitting predictors - in this case, the best predictor for the *Class* property is the attribute *Girth*, where $r(\text{Class}, \text{Girth}) = -1$. If the *Girth* of a specific graph is not measured, we would need to use a different, less fitting, graph property in the specified task. But, based on the correlation chain presented in Fig. 12, we can use the property of the Laplacian spectrum to impute the values of *Girth* for the graph, then use *Largest L-eigenvalue* for imputation of Laplacian spectrum and so on. In this case, we use the *Largest L-eigenvalue* to predict the value of the Laplacian spectrum and then use this predicted value to estimate the *Girth* of the graph, we can use the predicted *Girth* to determine the *Class* of graph (see Fig. 14). In the cubic graph property dataset, this simple approach using the concept of prediction potential (pseudo)transitivity reached $RMSE = 6.756e^{-15}$. For the comparison, the prediction of *Class* on the basis of the second best correlated attribute (Second largest eigenvalue) reached higher $RMSE = 0.2975$.

In the future, we plan to study the use of correlation graphs and correlation chains in the process of meta-analysis of decision-making models. One of the most well-regarded methods in this area is the Shapley Additive Explanation method based on the Shapley values from the theory of games.

With the use of Shapley values, we measure how individual attributes or sets of attributes contribute to the overall quality of the created decision-making model. The input for the Shapley Additive Explanations model is a trained machine or deep learning model, therefore the Shapley values are specific for that input model, which in turn needs to be trained to perform specific task.

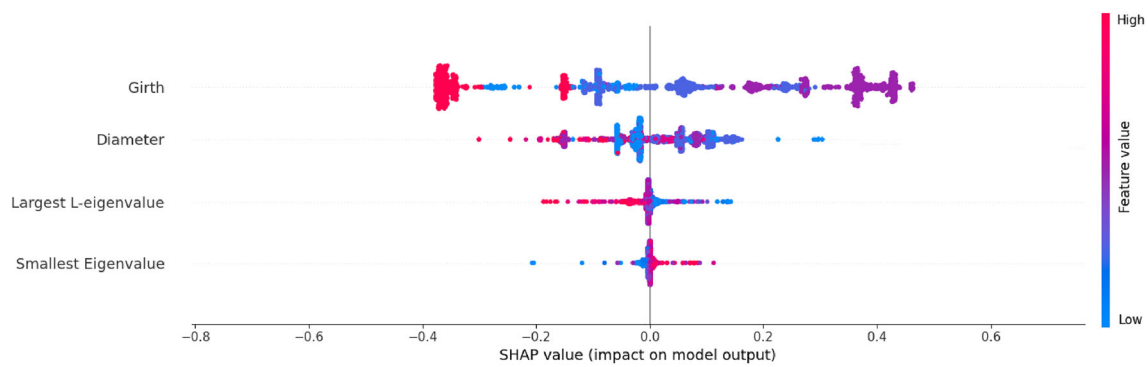


Fig. 15 Shapley Additive Explanations (SHAP) plot for the cubic graph dataset

Figure 15 presents a beeswarm plot of Shapley values measured on the Multilayer perceptron neural network with two hidden layers, which was trained for the classification of cubic graphs into classes based on the data presented in Section 4.2. This method identified the importance of individual graph properties in the studied dataset and determined the hierarchy of importance as shown in Fig. 15. The graph properties which reached the highest Shapley values are present in the correlation graph and chain models in a way of direct and indirect correlation influence - specifically, Girth of a graph was determined as a direct correlation influence on the classification of cubic graphs and the Diameter of a graph as an indirect one.

Since the datasets that are analyzed often contain a higher number of attributes, correlation graphs will not meet the condition of graph planarity - and thus will always contain intersecting elements (edges or vertices). In fact, the largest complete correlation graph that can be visualized on a plane in this way can examine the correlations between at most 4 attributes [24]. However, this shortcoming can be solved by appropriately plotting the graph in three-dimensional space in such a way that the strongest correlations are displayed on one plane and the weaker correlations on others.

Therefore, future work in the area also consists of visualization parameter tuning for the created implementation (mainly crossing of labels), including the correlation graph and chain package to the Python Package Index and creating an equivalent R language package. The possibility of design and implementation of the three-dimensional correlation graph and chain package is also strong.

Acknowledgements The author of the work would like to thank all the user study participants from the Department of Computer Science of Matej Bel University.

Funding Open access funding provided by The Ministry of Education, Science, Research and Sport of the Slovak Republic in cooperation with Centre for Scientific and Technical Information of the Slovak Republic.

Data availability Python code for the created 'Correlation Graph and Chain' software is freely available on <https://github.com/AdamDudasUMB/correlationGraphsAndChains>. The link also contains the dataset used in Section 4.2 of the presented article. In the case of any questions, please contact the author via e-mail: adam.dudas@umb.sk.

Declarations

Conflicts of interests The author declares no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Molnar, C.: *Interpretable Machine Learning*. Published independently. (2019). ISBN 979-8411463330
2. Skiena, S.S.: *The Data Science Design Manual*. Springer (2017). ISBN 978-3-319-55443-3
3. Kvet, M.: *Covering Undefined and Untrusted Values by the Database Index. Lecture Notes in Networks and Systems* **470**, 473-483. Springer. (2022). https://doi.org/10.1007/978-3-031-04829-6_42
4. Custode, L.L., Iacca, G.: Evolutionary learning of interpretable decision trees. *IEEE Access* **11**, 6169-6184 (2023). <https://doi.org/10.1109/ACCESS.2023.3236260>
5. Kutsanedzie, F., Achio, S., Ameko, E.: *Practical Approaches to Measurements*. Science Publishing Group, Sampling Techniques and Data Analysis (2016) ISBN 978-1-940366-58-6
6. Ramasubramanian, K., Singh, A.: *Machine Learning Using R*. Springer. (2019). ISBN 978-1-4842-4214-8
7. Fröhlich, K., Kundrata, I., Blaho, M., et al.: Performance of HfO_x- and TaO_x-based Resistive Switching Structures for Realization of Minimum and Maximum Functions. *MRS Adv.* **3**, 3427-3432 (2018). <https://doi.org/10.1557/adv.2018.377>

8. Nettleton, D.: *Commercial Data Mining*. Elsevier. (2014). ISBN 978-0-12-416602-8
9. Bon-Gang, H.: *Performance and Improvements of Green Construction Projects*. Elsevier. (2018). ISBN 978-0-12-815483-0
10. Weier, D.R., Basu, A.P.: An investigation of kendall τ modified for consored data with applications. *J. Stat. Plan. Inference* **4**, 381–390 (1980). [https://doi.org/10.1016/0378-3758\(80\)90023-3](https://doi.org/10.1016/0378-3758(80)90023-3)
11. Maack, R.G.C., Scheuermann, G., Hagen, H., et al.: Uncertainty-aware visual analytics: scope, opportunities, and challenges. *Vis. Comput.* (2022). <https://doi.org/10.1007/s00371-022-02733-6>
12. Earnshaw, R.A.: A new renaissance for creativity in technology and the arts in the context of virtual worlds. *Vis. Comput.* (2021). <https://doi.org/10.1007/s00371-021-02182-7>
13. Xue, L., Jiang, D., Wang, R., et al.: Learning semantic dependencies with channel correlation for multi-label classification. *Vis. Comput.* (2020). <https://doi.org/10.1007/s00371-019-01731-5>
14. Li, X., Fan, Y., Lv, G., et al.: Area-based correlation and non-local attention network for stereo matching. *Vis. Comput.* (2022). <https://doi.org/10.1007/s00371-021-02228-w>
15. Song, C., Wu, J., Zhu, L., et al.: Weight correlation reduction and features normalization: improving the performance for shallow networks. *Vis. Comput.* (2022). <https://doi.org/10.1007/s00371-021-02125-2>
16. Cauterruccio, F., Terracina, G.: Extended high-utility pattern mining: an answer set programming-based framework and applications. *Theory Pract. Logic Program.* (2023). <https://doi.org/10.1017/S1471068423000066>
17. Pena-Araya, V., Pietriga, E., Bezerianos, A.: A comparison of visualizations for identifying correlation over space and time. *IEEE Trans. Visual. Comput. Graph.* **26**(1), 375–385 (2019). <https://doi.org/10.48550/arXiv.1907.06399>
18. Yang, F., Shah, S.L., Xiao, D., Chen, T.: Improved correlation analysis and visualization of industrial alarm data. *ISA Trans.* **51**(4), (2021). <https://doi.org/10.1016/j.isatra.2012.03.005>
19. Caro, Y., Petruszewski, M., Skrekovski, R.: Remarks on proper conflict-free colorings of graphs. *Disc. Math.* **346**, 2 (2023). <https://doi.org/10.1016/j.disc.2022.113221>
20. Liu, H., Chen, C.h., Li Y., et al.: *Characteristic and Correlation Analysis of Metro Loads. Smart Metro Station Systems*, Elsevier, Pages 237–267 (2022). <https://doi.org/10.1016/B978-0-323-90588-6.00009-3>
21. Fisher, R.A.: *Iris. UCI Machine Learning Repository*. (1988). <https://doi.org/10.24432/C56C76>
22. Szűcs, G.: Multiclass classification by min-max ECOC with hamming distance optimization. *Vis. Comput.* (2022). <https://doi.org/10.1007/s00371-022-02540-z>
23. Dudáš, A., Modrovičová B.: Decision trees in proper edge k-coloring of cubic graphs. In *Proceedings of 33rd Conference of FRUCT Association*, pp. 21–29. (2023). ISSN 2305-7254
24. Yang, Y., Lin, J., Dai, Y.: Largest planar graphs and largest maximal planar graphs of diameter two. *J. Comput. Appl. Math.* **144**(1–2), 349–358 (2002). [https://doi.org/10.1016/S0377-0427\(01\)00572-6](https://doi.org/10.1016/S0377-0427(01)00572-6)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Adam Dudáš was born in Banská Bystrica, Slovakia, in 1992. He received the BS and MS degrees in computer science from Matej Bel University, Banská Bystrica, in 2015 and 2017, respectively, and the PhD degree in computer science from the University of Zilina, Slovakia, in 2020. Since 2020, he has been an Assistant Professor with the Department of Computer Science, Faculty of Natural Sciences, Matej Bel University. He is the author or coauthor of more than 25 research articles.

His research interests include exploratory and predictive data analysis, information visualization and big and irregular data and databases.