See You in the Branches: Correlation Trees and Forests in Predictive Data Analysis

Adam Dudáš

Department of Computer Science, Faculty of Natural Sciences,
Matej Bel University, 974 01 Banská Bystrica, Slovakia;
adam.dudas@umb.sk

Abstract: Correlation analysis is one of the statistical methods extensively used in the processes of data analysis. Since this method is focused on the identification of the prediction potential of an attribute pair in a dataset, it creates connections between multiple types of data analysis - from descriptive, through explorative to predictive. Yet, despite its importance, the utilization of visual models in conjunction with correlation analysis remains underexplored. Therefore, the main objective of this work is the design and implementation of graphical models for correlation analysis purposes called correlation trees and correlation forests. These models are focused on the visual presentation of attribute sequences of a dataset that bear strong potential for predictive analysis purposes. After the implementation of the proposed models, correlation trees and forests are constructed over three benchmarking datasets, and the approach is evaluated from three points of view. Firstly, visualization of the correlation trees and forests is evaluated. Then the utilization of correlation trees in regression analysis is verified with the use of LOESS and SVR regressors. And lastly, the main advantages and disadvantages of the proposed approach are identified. Experiments conducted on the considered datasets show that the use of correlation trees and forests is beneficial in the interpretation of correlation analysis results and in lowering regression model error.

Keywords: Correlation trees, Correlation analysis, Visual analysis, Regression analysis, Predictive analysis

1 Introduction

The utilization of visualization techniques in data analysis processes is crucial since it transforms complex data into comprehensible visual representations, aiding analysts in better understanding and interpretation of datasets and models built on them [1]. However, since it is common for the studied datasets to be multidimensional, techniques and methods for the selection of appropriate portions of datasets for further analysis purposes are necessary [2].

Through correlation analysis, it is possible to identify the strength and direction of the functional relationship between the values of two attributes in a common dataset.

Such a relationship, which can also be referred to as the prediction potential, is employed by decision-making models to solve regression or classification problems. In correlation analysis, the strength and direction of the relationship between a pair of attributes is measured by the correlation coefficient. Since this coefficient is conventionally measured only between two attributes, the approach of measuring the correlation coefficient for all possible pairs of attributes is required in multidimensional datasets [3].

The graphical models of correlation analysis, referred to as correlation structures, presented in [4] use the pseudo-transitivity of the predictive potential embedded in the correlation coefficients to build such a sequence of correlation pairs that can be used in predictive analysis. In other words, pseudo-transitivity is a process of predicting the values of attribute B based on the values of attribute A, then predicting the values of attribute C based on the predicted values of B, and so on.

The main objective of this work is the design and implementation of novel visual-correlation analysis models based on graphical structures called correlation trees and correlation forests. The focus of the proposed approach is on the visualization of sequences of attribute pairs of the studied dataset which demonstrate the strongest capacity for prediction potential utilized in predictive data analysis models. The main contributions of the approach presented in this work can be summarized in the following points:

- Design of novel correlation structure models of correlation trees and correlation forests. Both of these models are implemented using *Python* programming language while utilizing several specialized tool packages.
- Study of visualization of the proposed structures on three benchmarking open-access datasets - Iris dataset, Abalone dataset, and Wine dataset.
- Study of prediction error reduction, while using sequences of attribute pairs identified by the correlation trees, in the context of regression analysis.

The work presented in the scope of this study continues with a brief description of results reached in the related research works and then is divided into three main sections. In Section 2 the concept of correlation analysis is introduced as a basis for the design of the models of correlation trees and correlation forests. Then, in Section 3 these models are experimentally verified and evaluated with the use of three benchmarking datasets. The evaluation conducted in this work focuses on visualization of the models, use of the proposed models in regression analysis, and identification of advantages and disadvantages of use of correlation trees and forests. Lastly, Section 4 contains a summarization of the presented models and results, and suggests several future work areas in the studied field.

1.1 Related Work

Since the models proposed in the scope of this work are focused on visual analysis and visualization of correlation analysis with the use of graphical models, in this section, we offer a brief overview of works from these areas.

In [5], authors present a visualization approach designed to explore small-world networks represented by graphs called *ProtEGOnist*. This model uses the technique of ego-graphs which are visualized in an aggregated state as a so-called glyph where the size encodes the size of the neighborhood in the graph, and in a detailed version where the original network nodes can be explored. The focus of the proposed approach is on the reduction of the visual complexity of a graph, which in turn allows explorative analysis of the data and facilitates pattern recognition – the primary task of any type of data analysis [6]. The authors of the work demonstrate the applicability of the ego-graph approach in the context of real-world datasets – a co-author network and two protein-protein interaction networks.

Work explored in [7] focuses on insufficiencies present in conventional visual analysis tools when expressing the potential information behind the qualitative data. The authors of the study suggest a two-step methodology for the solution of this issue. The first step of this method focuses on the design and implementation of a text generation technology for adding text descriptions to visual data representations. Then, a new method of combining text and visualization is proposed to generate an interactive data analysis report, which helps users to explore data visualization through the interaction between text and visual components of the analysis. After its implementation, the authors present the proposed methodology in the context of manufacturing quality inspection and conclude that the text and visualization enrich each other, enhance the visualization effect, and the interactive analysis report optimizes the user's visual exploration of data.

Visual analysis in the context of text classification tasks is explored in [8], where authors focus on the semantic structures of datasets and decision-making explanation in the context of such datasets. The proposed approach called *SemLa* is a visual analysis system for comprehension of complex semantic structures in a dataset and visualizing nuances in the meaning of text samples in order to explain the decision-making process of the used model. As proposed, the *SemLa* design allows contrastive analysis at various levels by exploration of lexical and conceptual patterns including biases and artifacts in data. In their work, authors include human expert feedback on the proposed model and case studies, which serve as an evaluation and verification of the model in the visual analysis of text data.

Both correlation analysis and visualization are closely related to the modern area of explainable artificial intelligence. In [9] authors investigate the capability of the most popular explainable artificial intelligence techniques, such as LIME or SHAP [10], to explain the decisions of linear additive models. In the conducted experiments, authors measure the accuracy of additive and non-additive explanations while working with three regression models and 40 datasets. Authors focus on several aspects of the datasets, which can influence such explanation model's performance, eg. explanation sample size, similarity metric, used pre-processing technique, the number of numerical or categorical attributes, and correlation of attributes. It is the correlation analysis that plays a significant role in the work, and the authors conclude that rank-based measures, such as the Spearman rank correlation, are the most appropriate measures of similarity in the studied domain.

2 Correlation Trees and Forests

Correlation between a pair of attributes A and B originating in a common dataset measures the strength and orientation of prediction potential relationships between these attributes. Simply put, with the use of correlation analysis, analysts measure the amount of functional dependency of the values of one attribute on the other [11].

The core of correlation analysis is the correlation coefficient, described as:

$$corr(A,B) \in < -1, 1 > \tag{1}$$

where $corr(A,B)$ denotes generalized form of correlation coefficient value measured between attributes A and B.

For measuring of specific correlation cases, one of the following basic correlation coefficient types can be used [11, 12]:

- Pearson correlation coefficient (r) is fitting for measuring the strength and orientation of linear relationship between a pair of attributes in the following way [12]:

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

  where $\mu(A)$ and $\mu(B)$ denote mean values of attributes A and B and n is number of entities in the analyzed dataset.

- Spearman rank correlation coefficient (ρ) can be used as one of the non-linear alternatives to correlation coefficients. Using attribute value ranking - the position of an attribute value in the sorted array of its values - this coefficient is computed as [4]:

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

  where $rank(A)$ and $rank(B)$ are the rankings of studied attributes and n is the number of instances over which are the attributes measured.

- Kendall rank correlation coefficient (τ) represents a less frequently used alternative to the non-linear correlation measures. Instead of simple attribute values rankings, the so-called concordant and discordant pairs of attribute values are used in its computation [13]:

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

where n_c denotes the number of concordant pairs of values in the dataset, n_d denotes the number of discordant pairs of values, while n is the number of entities in the dataset.

As stated above, all correlation coefficient types acquire a value from $< -1, 1 >$ interval, while the closer the value of the correlation coefficient measured between a pair of attributes is to one of the extremes of this interval, the stronger the prediction potential between these attributes.

In the process of data analysis conducted over multidimensional datasets, the summarization of correlation coefficient values is conventionally done using a correlation matrix. For a dataset of n attributes, the correlation matrix C of size $n \times n$ uses attributes of the dataset for its cross-indexation, while:

$$\forall i, j \in \{1, 2, ..., n\}, C_{i.j} = corr(attr_i, attr_j). \tag{5}$$

Since for larger datasets, this matrix can be hard to interpret, the techniques from the area of data visualization are commonly used in the context of exploratory data analysis. The most conventional method for correlation matrix visualization is correlation heatmap, where the values of correlation coefficients are assigned a color from a given spectrum [4]. Less common visual models, which can be applied in correlation analysis of a dataset are represented by correlation structures - a set of graphical models, which use a correlation matrix and its modifications as an adjacency matrix of a graph [13].

In the context of correlation structures, the correlation graph can be viewed as a superstructure from which all other correlation-based substructures are derived. It consists of a vertices, one corresponding to each attribute in the studied dataset, with weighted edges representing the correlation coefficients measured between pairs of attributes. In this foundation, several commonly used correlation subgraphs can be identified — most notably correlation chains, which capture sequences of pairwise relationships between attributes; correlation n-ptychs, defined as complete subgraphs of the correlation graph; and correlation cycles, which represent cyclic correlation-based relationships among attributes.

The main objective of this work is the design and implementation of novel correlation structures — correlation trees and correlation forests — both of which are grounded in the correlation graph framework.

2.1 Correlation Trees

Both of the correlation structures introduced in this work are based on the general graph structures, described by [14]:

$$G = (V,E), E \subset [V]^2 \; while \; V = \{v_1, v_2, ..., v_n\}, \; E = \{e_1, e_2, ..., e_m\} \quad (6)$$

where G is a graph consisting of a set of vertices V and a set of edges E, where the elements of E represent two-element subsets of V - therefore, edges connect pairs of vertices, which for an edge connecting vertices v_1 and v_2 can be formalized as $e(v_1, v_2)$.

A tree is a specific type of graph, defined as an acyclic, connected graph - a graph in which there are no cycles and which consists of exactly one component. Commonly, one of the vertices of a tree is selected as the so-called root of a tree, the action, which suggests the ordering of the vertices in the tree. Such a tree is often called a rooted tree [14].

In the correlation structure visualizations, the set of attributes of a studied dataset is used as elements of V and the values of correlation coefficients represent edges between such vertices. A correlation tree T is a weighted tree, meaning all edges of T acquire value in the following way:

$$\begin{aligned} &\forall i, j \in \{1, 2, ..., n\}, \forall e(attr_i, attr_j) \in T, \\ &weight(e(attr_i, attr_j)) = corr(attr_i, attr_j) \end{aligned} \quad (7)$$

A correlation tree is rooted in the user-defined input attribute. This input attribute is used to construct the entire correlation tree, which consists of a set of branches - paths originating in $input$ and ending in other attributes of the dataset, in which the value of branch correlation strength (φ) is maximized:

$$\varphi(input, branch_B) = \frac{\sum_{i=1}^{n_{branch_B}-1} |corr(attr_i, attr_{i+1})|}{length(branch_B)} \quad (8)$$

where $input$ is an attribute of the input dataset which is used as the root node of pathfinding, $attr_i$ and $attr_j$ are attributes from the input dataset, and $length(branch_B)$ is the number of correlations measured in the branch B. In this way, the set of correlation branches is constructed (see Figure 1).

Since all of these branches are rooted in $input$, the visualization of the correlation tree is conducted as a union of roots of the set of branches:

$$input_{branch_1} \cup input_{branch_2} \cup ... \cup input_{branch_{n-1}} \quad (9)$$

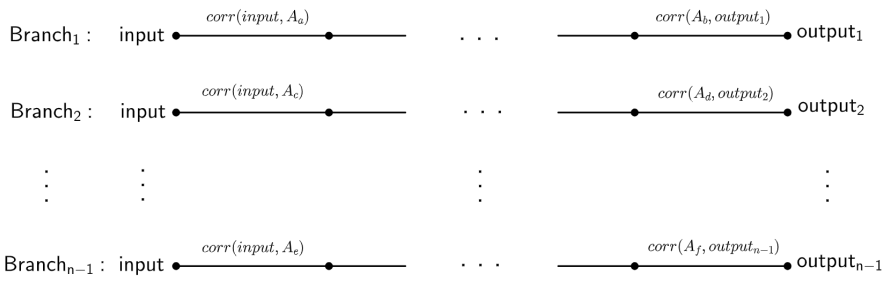


Figure 1
Set of correlation tree branches for a dataset of n attributes

The resulting correlation tree visualization is presented in Figure 2.

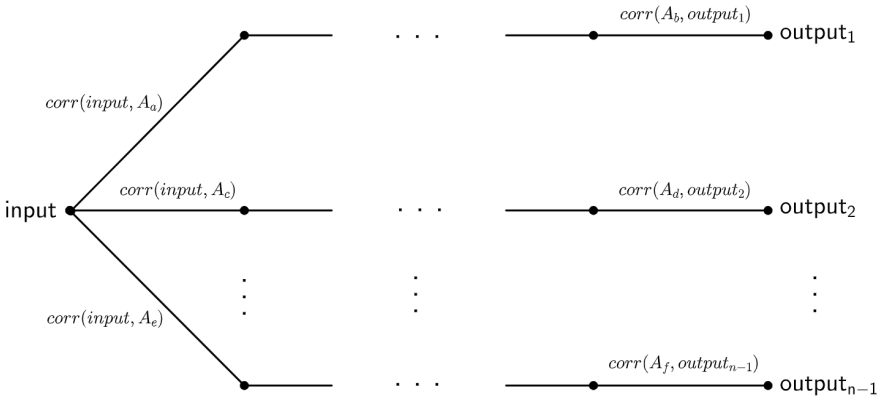


Figure 2
Correlation tree for a dataset of n attributes

Therefore, a correlation tree T is a rooted weighted tree constructed from a set of strongest correlation branches of an analysed dataset. Some noted properties of correlation trees as proposed (for a dataset of n attributes) are:

- Each correlation tree consists of $n-1$ branches. This is caused by the fact, that the model does not measure the correlation between the input attribute and itself $corr(input, input)$.
- Length of a branch of a correlation tree is computed as the number of correlation coefficients (edges) computed between $input$ and $output$ attribute of the branch. The maximal length of a branch in a correlation tree is $n-1$ while the minimal length of such a branch is 1.
- Even though all branches of a correlation tree have a unique structure (sequence of attributes), it is natural that subbranches of high correlation are repeated in the tree.

As can be seen in Figure 2 correlation tree represents an effective visualization of se-

quences of prediction potential relationships between an input attribute of the dataset and all other attributes of said dataset. Since the branches of the correlation tree contain the strongest correlation relationships between input and output attributes, this approach should lead to a lowering of the value of prediction error when conducting regression analysis from the input to an output attribute.

2.2 Correlation Forests

Naturally, the need for an input attribute in the construction of a correlation tree requires some amount of knowledge of the analyzed dataset. In order, to select such an attribute effectively, the analyst needs to familiarize themselves with the dataset of interest, which can originate in a completely unknown field of expertise to the analyst. Therefore, the necessity of an input attribute in a correlation tree can be perceived as a disadvantage, which is negated by the model of correlation forest.

A correlation forest is a set of correlation trees constructed from all attributes of the studied dataset. This leads to the following properties of correlation forests as proposed (for a dataset containing n attributes):

- There is n trees in all correlation forests, one for each $input$ attribute.
- All properties of correlation trees hold for all trees in a correlation forest.
- The trees in correlation forest contain unique branches except for reversed branches:

$$input_{branch_B} \rightarrow ... \rightarrow output_{branch_B} = (output_{branch_B} \rightarrow ... \rightarrow input_{branch_B})^R \tag{10}$$

  This is caused by a commutativity of correlation coefficients, where $corr(attr_1, attr_2) = corr(attr_2, attr_1)$. Such reversed equivalent branches are never present in one correlation tree.

A correlation forest visualization of the input dataset represents a more holistic view of correlation relationships in the analyzed data, where all branches of the highest values of branch correlation strength are used. This approach is fitting for use in the exploration of data and its relationships and can be used to identify correlation trees with the highest potential of use in the next steps of the analysis process.

3 Evaluation of Correlation Trees and Forests in the Process of Data Analysis

The models proposed in this work were implemented in *Python* language using specialized packages. This section of the study presents an evaluation of correlation tree and correlation forest approaches in the context of various aspects of data analysis. Firstly, the visual analysis of strong correlation sequences of a dataset based on Spearman rank correlation is presented, with a focus on the visualization of correlation trees, and correlation forests and a brief analysis of their descriptive properties. Then, the model of correlation trees is applied in the regression analysis with the use of LOcally Estimated Scatterplot Smoothing (LOESS) and Support

Vector Regression (SVR) regressors. Finally, the advantages and disadvantages of both proposed visualization models are discussed.

The visual and regression analysis conducted as a part of the evaluation of the correlation trees and forests proposed in this work uses three open-access benchmarking datasets commonly used in regression tasks:

- Iris dataset contains 5 attributes that describe the physical properties of iris flowers (length and width for petal and sepal leaves) and their classification. For the purposes of correlation tree evaluation, the version of the dataset with label encoded class attribute was used [15, 16].
- Abalone dataset is composed of 8 quantitative attributes describing physical measurements of abalones. Among these properties, the characteristics as sex, length, diameter, and height of the shell, weight of various parts of the abalone body, or the number of rings visible in the shell are measured. Commonly, the dataset is used for the prediction of the age of abalone based on its physical properties [17].
- Wine dataset of 13 attributes focusing on chemical analysis of Vinho Verde wines in Italy was selected as the third dataset for its well-known low correlation values between the chemical properties of wines, such as various forms of acidity of the wine, the amount of sugars, chlorides, sulfur dioxide, or the density, alcohol content, and overall quality of the wine assigned by a human expert from the area [18, 19].

3.1 Visualization of Correlation Analysis Through Correlation Trees and Forests

The first point of view in the context of the evaluation of the proposed correlation tree and forest model is its visualization. The correlation trees and forests are constructed for each of the three selected benchmarking datasets to visualize the sequences of the strongest correlation relationships between pairs of attributes of the dataset and the basic properties of the trees and forests are described.

Fig 3 presents a correlation tree constructed over the Iris dataset with the use of $Sepal\_W$ as an input attribute. As can be seen, the tree contains four branches with an average value of $\varphi = 0.885$ while the length of branches varies between 3 and 4.

In Figure 4 a correlation tree for the Abalone dataset is visualized. In this case, the $Shell$ attribute was selected as an input for computing, and 7 other attributes of the dataset were considered an output of the tree-building process. Compared to the tree of the Iris dataset, we can see more variety in the length of branches of the Abalone tree - in this specific case, the branch length varies from 2 up to 7 with an average φ value of 0.96.

The third dataset used in the visualization of the correlation trees was the Wine dataset, the tree of which is presented in Figure 5. When constructing a correlation tree from the $resSugar$ attribute there are branches of length from 2 to 7 with

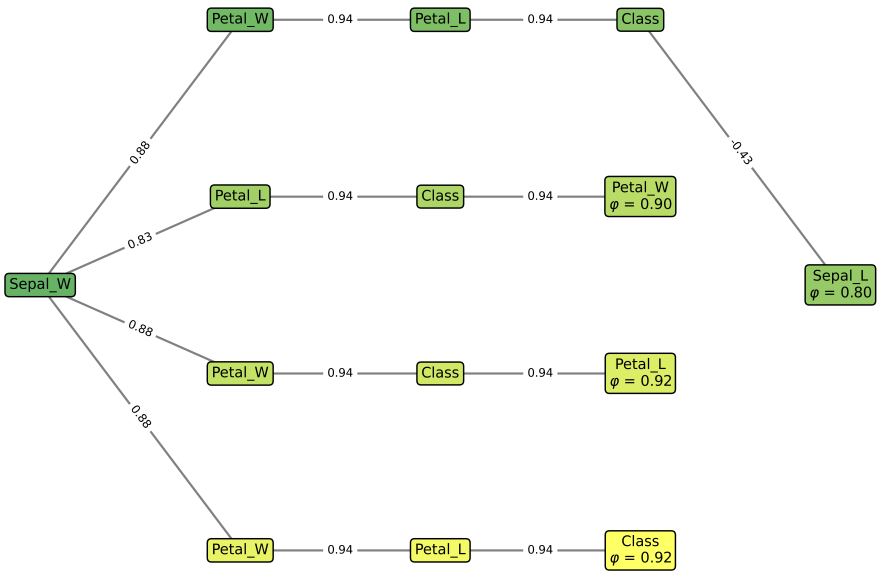


Figure 3
Correlation Tree for Iris dataset with input attribute = *Sepal_W*

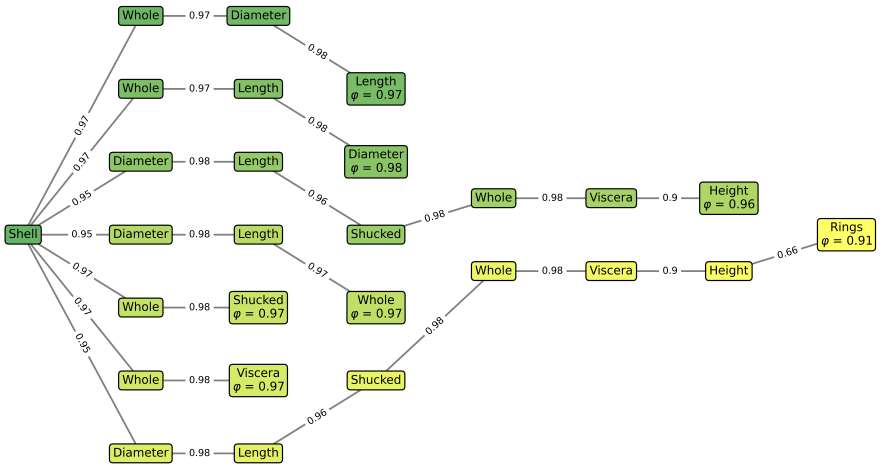


Figure 4
Correlation Tree for Abalone dataset with input attribute = *Shell*

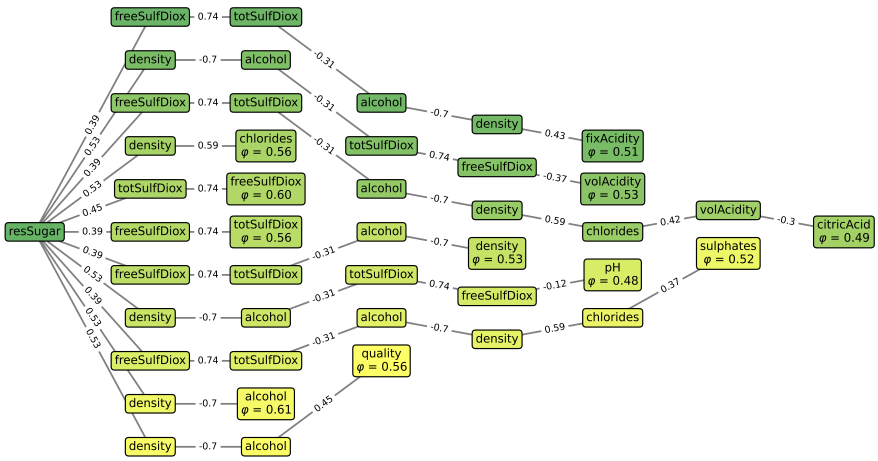


Figure 5
Correlation Tree for Wine dataset with input attribute = *resSugar*

Table 1
Basic Properties of the Visualized Correlation Trees for Iris, Abalone, and Wine Datasets

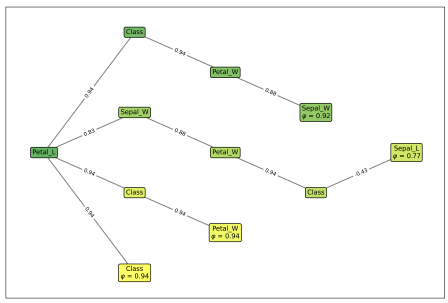
Dataset	**min (length of a branch)**	**max (length of a branch)**	**min(φ)**	**max(φ)**
Iris	3	4	0.8	0.92
Abalone	2	6	0.91	0.98
Wine	2	7	0.48	0.61

an average $\varphi = 0.54$. This low value of branch correlation strength is caused by low correlation coefficient values of the dataset - there are weaker functional relationships between the chemical properties of wines considered in this dataset.

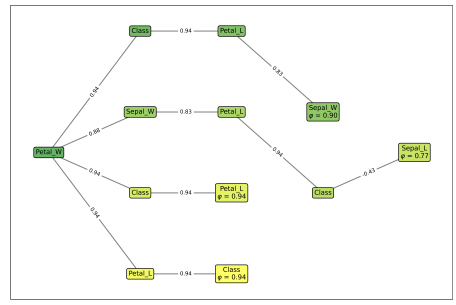
All of the basic properties of the visualized trees - minimal and maximal length of a branch and minimal and maximal value of φ in the tree - are presented in Table 1.

As an example of a correlation forest, we present a forest of the Iris dataset consisting of 5 correlation trees - one for each attribute of the dataset used as an input (see Figure 6 for the visualization of the forest). As seen in the figure, the trees in the forest are quite heterogeneous from the point of view of branch length, which varies from 1 to 4 in some trees of the forest (eg. 6a) but can be consistently 4 in others (eg 6c). However, the value of φ does not vary drastically between the trees, which is natural, since all of the trees of the forest use the strongest correlation portions of the dataset.

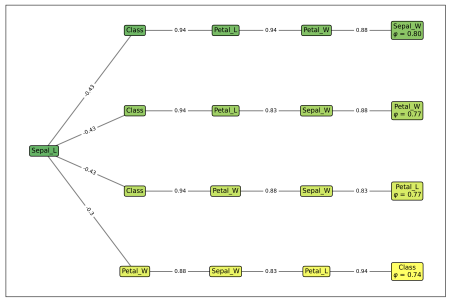
Summarization of the properties of Iris, Abalone, and Wine correlation forests are presented in Table 2. The basic observations of tree properties in the common correlation forest hold for all three datasets, yet the Wine dataset exhibits one interesting characteristic - there are 12 attributes in the Wine dataset, but the maximal length of the branch in the whole forest is 7. Therefore, at most only 8 of the 12 attributes



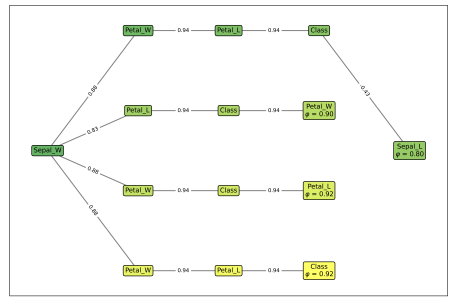
(a)



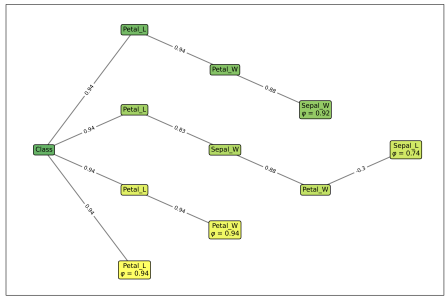
(b)



(c)



(d)



(e)

Figure 6
Correlation Forest for Iris Dataset

Table 2
Basic Properties of the Constructed Correlation Forests for Iris, Abalone, and Wine Datasets

Dataset	Number of trees	min (length of a branch)	max (length of a branch)	min(φ)	max(φ)
Iris	5	1	4	0.74	0.94
Abalone	8	1	7	0.92	0.98
Wine	12	2	7	0.48	0.61

were included in the building of the correlation branch, which suggests there are parts of the dataset (attributes), which might not be significant from the point of view of prediction data analysis. Examples of such attributes of the dataset are *sulphates* or pH, which are not included in any branch of any tree in the forest (except when being the output of the branch).

3.2 Regression Analysis Using Correlation Trees

Regression analysis focuses on the prediction of values of quantitative attributes from continuous intervals based on the patterns and trends identified in data [20]. A large number of algorithms, models, and methods of regression analysis were developed, but in this work, we use two of the most common ones - LOcally Estimated Scatterplot Smoothing regression (LOESS) and Support Vector Regression (SVR).

Regardless of the method used in the regression problem, the quality of the decision-making process is measured by one of the standardized error metrics. For the purposes of this work, the Root Mean Squared Error (RMSE) is utilized as follows [21]:

$$RMSE(A) = \sqrt{\frac{\sum_{i=1}^{n}(p(A_i) - a(A_i))^2}{n}} \tag{11}$$

where A denotes the predicted attribute, $p(A_i)$ is the predctited value of i-th measurement of A, and $a(A_i)$ denotes the actual value of the measurement.

In the scope of the regression analysis, we focus on the comparison of RMSE values between two approaches - direct regression from input attribute to output attribute, and regression based on the sequence of attributes defined by branches of the correlation tree. This comparison is done using all branches of correlation trees presented in Section 3.1. Therefore, 22 regression cases are being solved - 4 for the Iris dataset, 7 for the Abalone dataset, and 11 for the Wine dataset.

Tables 3 and 4 present the results of these experiments for LOESS and SVR regressor respectively, while supplementing the RMSE information with data on the length of branch used in the experiment, φ value of the brach, and direct correlation value between input and output attributes of the regression problem ($\rho(input, output)$).

Table 3
Comparison of Quality of the Constructed Correlation Trees in the Context of Regression Analysis Using LOESS Regression

Dataset	**Input**	**Branch output**	**Branch length**	φ	ρ **(input, output)**	**LOESS RMSE direct**	**LOESS RMSE branch**
Iris	Sepal_W	Sepal_L	4	0.8	-0.16	**0.41**	0.54
		Petal_W	3	0.9	0.88	0.781	**0.45**
		Petal_L	3	0.92	0.83	0.393	**0.243**
		Class	3	0.92	0.8	0.467	**0.207**
Abalone	Shell	Length	3	0.97	0.95	0.033	**0.019**
		Diameter	3	0.98	0.95	0.026	**0.016**
		Height	6	0.96	0.92	**0.022**	0.023
		Whole	3	0.97	0.97	0.141	**0.128**
		Shucked	2	0.97	0.92	0.101	**0.054**
		Viscera	2	0.97	0.94	0.044	**0.028**
		Rings	7	0.91	0.7	**2.546**	2.638
Wine	resSugar	fixAcidity	5	0.51	-0.03	1.27	**1.18**
		volAcidity	5	0.53	-0.06	0.156	0.156
		citricAcid	7	0.49	0.07	0.143	**0.134**
		chlorides	2	0.56	-0.04	0.034	**0.033**
		freeSulfDiox	2	0.6	0.4	15.729	**12.348**
		totSulfDiox	2	0.56	0.45	45.534	**37.313**
		density	4	0.53	0.53	0.0023	**0.0022**
		pH	5	0.48	-0.15	**0.202**	0.203
		sulphates	6	0.52	-0.14	0.149	**0.143**
		alcohol	2	0.61	-0.33	1.095	**0.873**
		quality	3	0.56	-0.02	0.868	**0.776**

In both tables, the lower (and therefore better) RMSE values are presented in bold lettering.

The overall evaluation of RMSE comparison is summarized in Table 5, where the number (#) and percentage (%) of conducted regression cases for individual datasets and regressors are presented. These experiments brought several interesting results:

- The approach of using branches of correlation trees in the regression problem produced lower RMSE values in 77.3% cases when compared to direct regression methods.
- Even though both, the LOESS and SVR regressors, reached lower RMSE values using the correlation tree model in 77.3% cases, in SVR the percentage of cases of correlation branches producing higher RMSE was 13.7%, while using LOESS produced 18.2% of such cases.
- The majority of cases of correlation branches producing higher RMSE when compared to direct correlation have occurred in the Abalone dataset. This might be caused by incremental error inflation in the branches, which be-

Table 4
Comparison of Quality of the Constructed Correlation Trees in the Context of Regression Analysis Using SVR Regression

Dataset	Input	Branch output	Branch length	φ	ρ (input, output)	SVR RMSE direct	SVR RMSE branch
Iris	Sepal_W	Sepal_L	4	0.8	-0.16	0.409	**0.34**
		Petal_W	3	0.9	0.88	0.782	**0.428**
		Petal_L	3	0.92	0.83	0.397	**0.207**
		Class	3	0.92	0.8	0.49	**0.196**
Abalone	Shell	Length	3	0.97	0.95	**0.037**	0.039
		Diameter	3	0.98	0.95	0.032	**0.027**
		Height	6	0.96	0.92	**0.036**	0.039
		Whole	3	0.97	0.97	0.138	**0.129**
		Shucked	2	0.97	0.92	0.104	**0.064**
		Viscera	2	0.97	0.94	0.056	**0.046**
		Rings	7	0.91	0.7	**2.547**	2.635
Wine	resSugar	fixAcidity	5	0.51	-0.03	1.267	**1.156**
		volAcidity	5	0.53	-0.06	0.155	**0.154**
		citricAcid	7	0.49	0.07	0.143	**0.132**
		chlorides	2	0.56	-0.04	0.073	**0.072**
		freeSulfDiox	2	0.6	0.4	15.866	**12.497**
		totSulfDiox	2	0.56	0.45	46.587	**38.094**
		density	4	0.53	0.53	0.019	0.019
		pH	5	0.48	-0.15	0.2	0.2
		sulphates	6	0.52	-0.14	0.148	**0.139**
		alcohol	2	0.61	-0.33	1.103	**0.814**
		quality	3	0.56	-0.02	0.875	**0.802**

comes noticeable when the values of the direct correlation coefficient in a dataset are very high, like in the case of the Abalone dataset.

3.3 Advantages and Disadvantages of Correlation Trees and Forests

Finally, in this section of evaluation of the proposed model, we present the advantages and disadvantages of the correlation trees and forests as proposed. The main, major advantages of the approach are:

- Visualization of the sequences of highly correlated attributes in a dataset. This form of visualization brings high readability of transitive prediction potential in the studied dataset, which is much harder to retrieve from conventional visualization models used in correlation analysis, such as correlation matrix.
- Lower values of RMSE when compared to direct regression approaches. With the use of correlation trees and forests, we are able to build a relatively simple and interpretable regression model which - in most cases - reaches lower error values than its direct regression counterpart. This type of correlation tree-

Table 5
Evaluation of RMSE Values for All Three Considered Datasets and Both Regression Models

Regression method	Parameter	Iris	Abalone	Wine	Overall
LOESS	$\#\, RMSE_{branch} < RMSE_{direct}$	3	5	9	17
	$\%\, RMSE_{branch} < RMSE_{direct}$	75%	71.4%	81.8%	77.3%
	$\#\, RMSE_{branch} = RMSE_{direct}$	0	0	1	1
	$\%\, RMSE_{branch} = RMSE_{direct}$	0%	0%	9.1%	4.5%
	$\#\, RMSE_{branch} > RMSE_{direct}$	1	2	1	4
	$\%\, RMSE_{branch} > RMSE_{direct}$	25%	28.6%	9.1%	18.2%
SVR	$\#\, RMSE_{branch} < RMSE_{direct}$	4	4	9	17
	$\%\, RMSE_{branch} < RMSE_{direct}$	100%	57%	81.8%	77.3%
	$\#\, RMSE_{branch} = RMSE_{direct}$	0	0	2	2
	$\%\, RMSE_{branch} = RMSE_{direct}$	0%	0%	18.%	9%
	$\#\, RMSE_{branch} > RMSE_{direct}$	0	3	0	3
	$\%\, RMSE_{branch} > RMSE_{direct}$	0%	43%	0%	13.7%

based regression model is very flexible since its parts (regression models used in the branches of the correlation tree) can be interchanged freely.

The main disadvantages of the proposed model can be summarized into the following two points:

- The need for an input attribute when working with correlation trees. Since the model requires user-defined input in the form of one of the dataset's attributes, previous knowledge of the dataset is needed for the qualified selection of such an attribute. This issue is negated by correlation forests, which require no input since one tree is built for each of the attributes of the dataset separately.
- The main disadvantage of the proposed approach at the moment is its time and memory consumption when constructing a correlation forest for large datasets. Most of the computation time and memory is dedicated to pathfinding in the correlation matrix, which identifies branches with the highest φ. This problem suggests the use of parallel computing, where each of the trees of the forest could be built concurrently.

4 Conclusion

The correlation tree and correlation forest presented in the scope of this work represent an effective visualization of prediction potential based on correlation analysis in multidimensional datasets. This approach was designed to lower the value of prediction error when conducting regression analysis using explainable, interpretable, and visualizable regressors. Proposed models are, therefore, usable in exploratory and predictive analyses which are crucial from the point of view of data analysis process.

The proposed concept was experimentally verified and evaluated from three main

perspectives. Firstly, the visualization of the correlation trees and forests was examined together with the statistical properties of the proposed correlation structures. Secondly, the use of correlation trees in conjunction with LOESS and SVR regressors was presented in the context of regression analysis. The use of proposed models brought lower RMSE results in 77.3% of testing cases when compared to the basic regression methods. Lastly, the strengths and weaknesses of the correlation tree and forest models were examined.

In this way, correlation trees provide a unified, visualization-based model for identifying all strong predictive sequences within a dataset. From the perspective of correlation structures, a correlation tree is defined as a subgraph of the correlation graph that contains the strongest correlation chains (or paths) between a selected input attribute and all other attributes in the dataset. A correlation forest can then be defined as a set of such correlation tree subgraphs, which are not naturally discernible within the full correlation graph and, as a result, cannot be conveniently analyzed or interpreted without the proposed method for their separate extraction and visualization.

The research in the area of correlation trees and correlation forests presented in the scope of this study suggests several future work directions worthy of examination:

- Type of pruning of a correlation forest for selection of only a given amount of best-performing correlation trees is necessary and possible. This pruning should be based on a measurement for overall correlation tree strength, which can be computed in several ways, eg. the highest φ of the tree, the mean value of φ of all branches, a tree with the lowest deviation of φ measure and so on.

- Alternative visualization layouts for larger trees that use repeated sub-branches of a tree to cluster a set of branches are worthy of implementation. This way, more conventional n-ary trees can be created.

- An examination of the regression error inflation in the branches of correlation trees might be required in the future. This could lead to further improvement of the performance of correlation trees and forests in the visual data analysis and regression problems.

- Since the implementation of the concept used in this work is quite time- and memory-intensive, the optimization of computation itself is open to research.

Acknowledgement
The research presented in this work was supported by the University Grant Agency of Matej Bel University in Banská Bystrica project number UGA-14-PDS-2025.

Code and Data Availability
Python code for the proposed visualization model is available at:

`github.com/AdamDudasUMB/corrTreesAndForests`

Datasets used in the experiments of the presented study are publicly and openly available at `archive.ics.uci.edu`.

References

[1] R.G.C. Maack et al. Uncertainty-aware visual analytics: scope, opportunities, and challenges. Visual Computer, 2023. DOI: 10.1007/s00371-022-02733-6

[2] I. Belkacem et al. Interactive Visualization on Large High-Resolution Displays: A Survey. Computer Graphics Forum, 2024. DOI: 10.1111/cgf.15001

[3] S. S. Skiena. The data science design manual. Springer, 2017. DOI: 10.1007/978-3-319-55444-0

[4] A. Dudáš. Graphical representation of data prediction potential: correlation graphs and correlation chains. Visual Computer, 2024. DOI: 10.1007/s00371-023-03240-y

[5] N. Brich et al. ProtEGOnist: Visual Analysis of Interactions in Small World Networks Using Ego-graphs. Computer Graphics Forum, 2024. DOI: 10.1111/cgf.15078

[6] F. Miranda et al. The State of the Art in Visual Analytics for 3D Urban Data. Computer Graphics Forum, 2024. DOI: 10.1111/cgf.15112

[7] L. Chen, M. Wu. Research on interactive analysis report in data analysis and visualization platform. Multimedia Tools and Applications, 2023. DOI: 10.1007/s11042-022-13652-y

[8] M. Battogtokh et al. Visual Analytics for Fine-grained Text Classification Models and Datasets. Computer Graphics Forum, 2024. DOI: 10.1111/cgf.15098

[9] A.H.A. Rahnama et al. Can local explanation techniques explain linear additive models? Data Mining and Knowledge Discovery, 2024. DOI: 10.1007/s10618-023-00971-3

[10] G. Szepannek, K. Lübke. Explaining Artificial Intelligence with Care. Künstliche Intelligenz, 2022. DOI: 10.1007/s13218-022-00764-8

[11] L.B. Iantovics. Avoiding Mistakes in Bivariate Linear Regression and Correlation Analysis, in Rigorous Research. Acta Polytechnica Hungarica, 2024. DOI: 10.12700/APH.21.6.2024.6.2

[12] M.F. Ikhwan et al. Pearson Correlation and Multiple Correlation Analyses of the Animal Fat S-Parameter. TEM journal, 2024. DOI: 10.18421/TEM131-15

[13] A. Dudáš. Correlation *n*-ptychs of Multidimensional Datasets. Lecture Notes in Networks and Systems, Vol. 990, 2024. DOI: 10.1007/978-3-031-60328-0_15

[14] R. Diestel. Graph Theory. Springer, 2016. DOI: 10.1007/978-3-662-53622-3

[15] R.A. Fisher. Iris. UCI Machine Learning Repository, 1988. DOI: 10.24432/C56C76.

[16] A. Unwin, K. Kleinman. The Iris data set: In search of the source of virginica. Significance, 2021. DOI: 10.1111/1740-9713.01589

[17] W. Nash et al. Abalone. UCI Machine Learning Repository, 1995. DOI: 10.24432/C55C7W

[18] S. Aeberhard, M. Forina. Wine. UCI Machine Learning Repository, 1991. DOI: 10.24432/C5PC7J

[19] P. Cortez et al. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 2009. DOI: 10.1016/j.dss.2009.05.016

[20] C. Igel, S. Oehmcke. Remember to Correct the Bias When Using Deep Learning for Regression!. Künstliche Intelligenz, 2023. DOI: 10.1007/s13218-023-00801-0

[21] T. Schmid et al. The AI Methods, Capabilities and Criticality Grid. Künstliche Intelligenz, 2021. DOI: 10.1007/s13218-021-00736-4