Orange Software Usage in Data Mining Classification Method on The Dataset Lenses

Aulia Ishak[1], Khawarita Siregar[2], Aspriyati[3], Rosnani Ginting[4], Muhammad Afif[5]

[1,2,4,5]Industrial Engineering Department, Faculty of Engineering, Universitas Sumatera Utara
[3]Public Health Faculty, Universitas Sumatera Utara, Medan, Indonesia

E-mail: muhammadafif2603@gmail.com aulia.ishak@usu.ac.id

Abstract. Data Mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from various large databases. Decision tree is a very interesting classification method that involves the construction of a decision tree consisting of decision nodes which are connected by branches from the root node to the leaf node (end). The problem that will be examined in this study is about classification of dataset lenses obtained from using orange software. The method used in this paper is the method of classification process is performed on a decision tree using the orange software. Tree construction begins with the formation of roots (located at the top). Then the data is divided based on attributes that are suitable to be used as leaves. Decision rule information is to make decision rules from trees that have been formed.

1. Introduction

Data mining has attracted a lot of attention in the community in recent years, being able to convert large amounts and large amounts of data into useful information and knowledge. The information and knownledge obtained can be used to apply such as market analysis, fraud detection, and customer retention, for production control and exploration science.

Data mining is a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and related knowledge from various large databases [1]. Data mining is a series of processes to explore the added value of a data set in the form of knowledge that has not been knwn manually [2]. On the off chance that the consequences of QC tests can't satisfy the acknowledgment models, the aftereffects of examination of the entire arrangement of the estimations on that day must be eliminated or should be re-dissected, and an incomplete or full re-approval of the strategy considered [12].

There are many data mining methodologies, one of which is poplar is the decision tree. Decision tree is a very interesting classification method that involves the construction of a decision tree consisting of decision nodes which are connected by branches from the root node to the leaf node (end). At the decision node the attribute will be tested, and each result will produce a branch. Each branch will be directed to another node or to the end node to produce a decision [3].

Contact lenses have become part of the lifestyle of modern society today. Contact lenses are very popular, especially in big cities. Many people, especially women, use contact lenses not only as visual aids but are also used as cosmetic tools to beautify the eye with a variety of attractive colors [4].
To build classification algorithms on datasets, open source data mining software is needed. The software used is a tool from Orange based on phyton programming [5]. Orange is a data mining tool which is useful for visual programming and explorative data analysis. It can be written in phyton. Orange has multiple components are known as widgets. This data mining tool supports macOS, Windows and Linux [6].
The decision tree has several versions, namely ID3, C4.5, J48, C5.0. However, Orange Data Mining uses ID3 which not only acts a classification, but can also perform regression, as in the CART method [7]. The classification results of Decision Tree produce two rules of supplier selection model to classify suppliers so thet companies can easily choose suppliers according to the criteria desired by the company [8]
The problem that will be examined in this study is about he classification of dataset lenses obtained from using orange software. Where the dataset lenses is obtained from the UCI Machine Learning Repository website and is used to analyze contact pairs. By doing this research, is expected to produce a knowledge for patients who require or do not require contact lenses.

2. Research Methodology

The study was conducted using Orange software with classification method performed on the decision tree. The data used is secondary data, data collected second hand or from other sources that were available before the study was conducted [9]. The data source obtained comes from the UCI Machine Learning Repository website. The steps taken are entering the dataset to be checked into the Orange software. After entering the dataset, then arrange the widgets in such a way as to install the decision tree on Orange. By installing Tree Viewer, will bring up results obtained decision tree. The data used in this study were taken from UCI with 24 data and 4 attributes.

2.1. Attributes on Dataset

Attribute is something that is important to the accuracy of the procces, so it is necessary to know the main attributes [10]. The "type" attribute becomes the class attribute. The set lenses classes in this dataset include:

- The patient should be fitted with hard contact lenses (1)
- The patient should be fitted with soft contact lenses (2)
- The patient should not be fitted with contact lenses (3)

And 4 other attributes as much as 4 pieces, namely:

- Age of the patient: (1) young, (2) pre-presbyopic, (3) presbopic
- Spectacle prescriptions: (1) myope, (2) hypermetrope
- Astigmatic: (1) no, (2) yes
- Tear production rate: (1) reduced, (2) normal

2.2. Missing value on Dataset

Missing value is information that is not available for an object (case). The missing value due to the information about the object is not given, it is difficult to find, or indeed the information is not ther, it will cause a decrease in the accuracy and the quality of the data as it is processed [11]. Missing value in the dataset is nothing (0).

2.3. Instances on Dataset

Instances are the number of records in the dataset. The number of instances in the dataset is 24.

2.4. Characteristics of Attibutes on Dataset

The characteristic of the attributes in this dataset are chategorical.

2.5. Dataset Display

Here is a table of dataset lenses.

Table 1. Dataset lenses display

No.	Age of the patient	Spectacle prescription	Astigmatic	Tear production rate	Type
1	1	1	1	1	3
2	1	1	1	2	2
3	1	1	2	1	3
4	1	1	2	2	1
5	1	2	1	1	3
6	1	2	1	2	2
7	1	2	2	1	3
8	1	2	2	2	1
9	2	1	1	1	3
10	2	1	1	2	2
11	2	1	2	1	3
12	2	1	2	2	1
13	2	2	1	1	3
14	2	2	1	2	2
15	2	2	2	1	3
16	2	2	2	2	3
17	3	1	1	1	3
18	3	1	1	2	3
19	3	1	2	1	3
20	3	1	2	2	1
21	3	2	1	1	3
22	3	2	1	2	2
23	3	2	2	1	3
24	3	2	2	2	3

In this study, expected after doing this research will produce a science that can be used for the benefit of government and society.

3. Result and Discussion

Here is an initial stage until the end of the classification decision tree by using Orange:

- Open the Orange software, then select File and drag it to the worksheet section

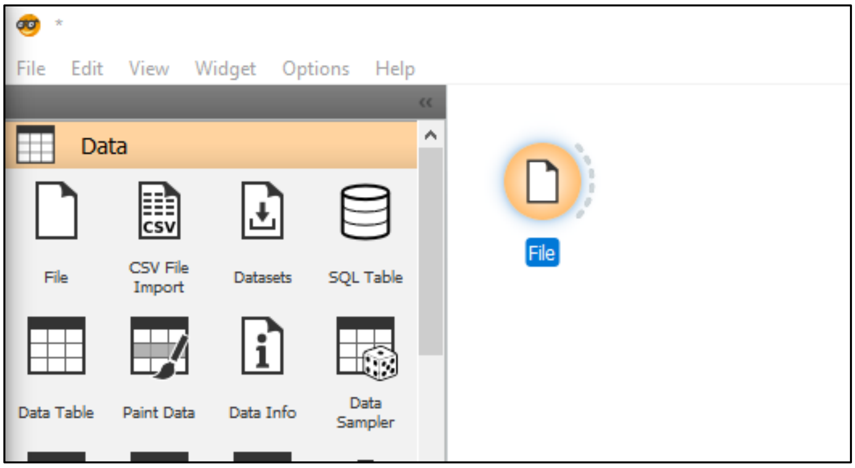
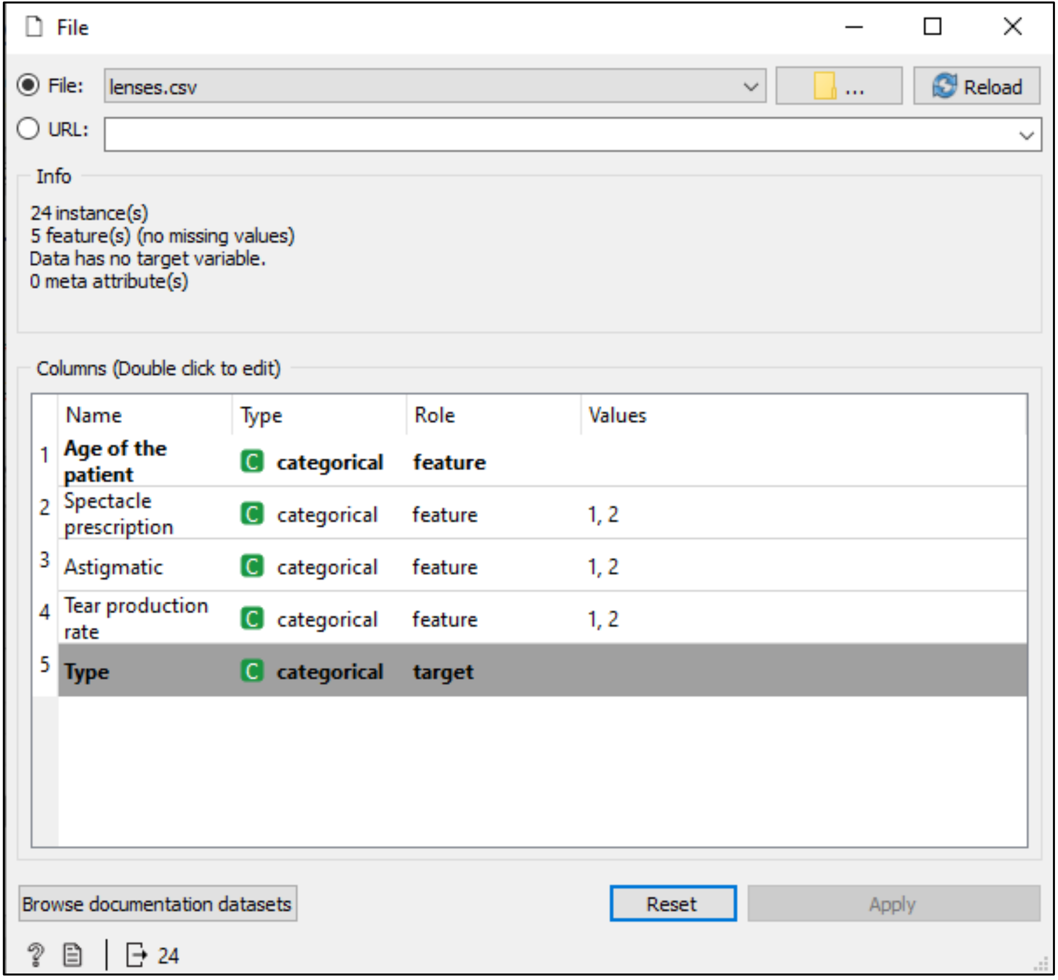


Figure 1. Display widget file in orange software

- Double-click the File symbol on the worksheet, and select the dataset to be examined. So the new window appears as shown below



● File: lenses.csv

○ URL:

Info

24 instance(s)
5 feature(s) (no missing values)
Data has no target variable.
0 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	**Age of the patient**	C **categorical**	**feature**	
2	Spectacle prescription	C categorical	feature	1, 2
3	Astigmatic	C categorical	feature	1, 2
4	Tear production rate	C categorical	feature	1, 2
5	**Type**	C **categorical**	**target**	

e

Figure 2. Display file properties in orange software

- Drag the Data Table on the worksheet section to view the selected data, then connect it.

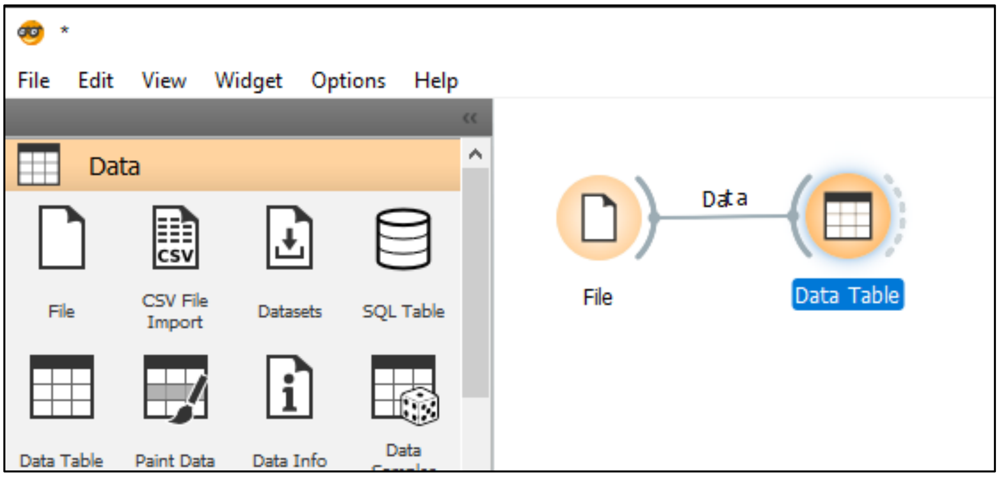
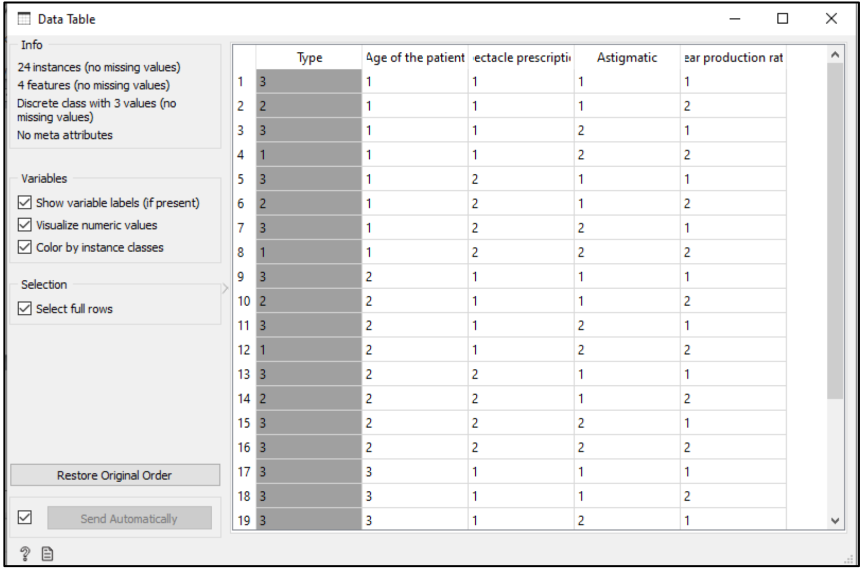


Figure 3. Display data table widget in orange software

- Double-click on the Data Table on the worksheet to see the input data, then a menu like this will appear



Info

24 instances (no missing values)
4 features (no missing values)
Discrete class with 3 values (no missing values)
No meta attributes

Variables

☑ Show variable labels (if present)
☑ Visualize numeric values
☑ Color by instance classes

Selection

☑ Select full rows

☑ Send Automatically

	Type	Age of the patient	ectacle prescripti	Astigmatic	ear production rat
1	3	1	1	1	1
2	2	1	1	1	2
3	3	1	1	2	1
4	1	1	1	2	2
5	3	1	2	1	1
6	2	1	2	1	2
7	3	1	2	2	1
8	1	1	2	2	2
9	3	2	1	1	1
10	2	2	1	1	2
11	3	2	1	2	1
12	1	2	1	2	2
13	3	2	2	1	1
14	2	2	2	1	2
15	3	2	2	2	1
16	3	2	2	2	2
17	3	3	1	1	1
18	3	3	1	1	2
19	3	3	1	2	1

Figure 4. Display properties data table in orange software

- Insert the Select Columns widget on the canvas then connect with the widget file

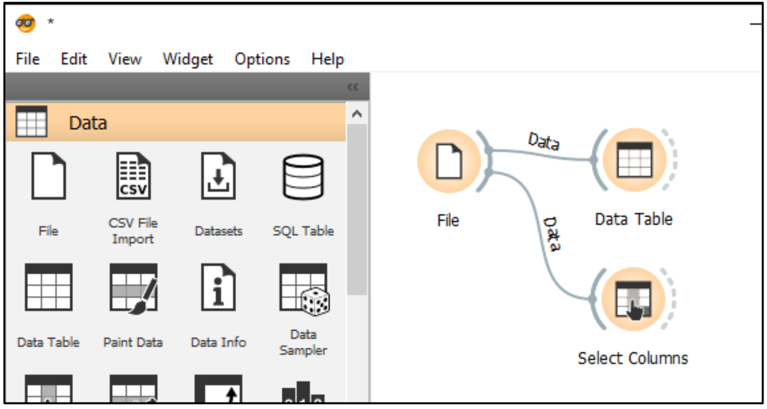


Figure 5. Display select columns widget linked to file widgets in orange software

- To do the classification process on existing data, you can select the Tree widget. And then connect with the Select Column Widget

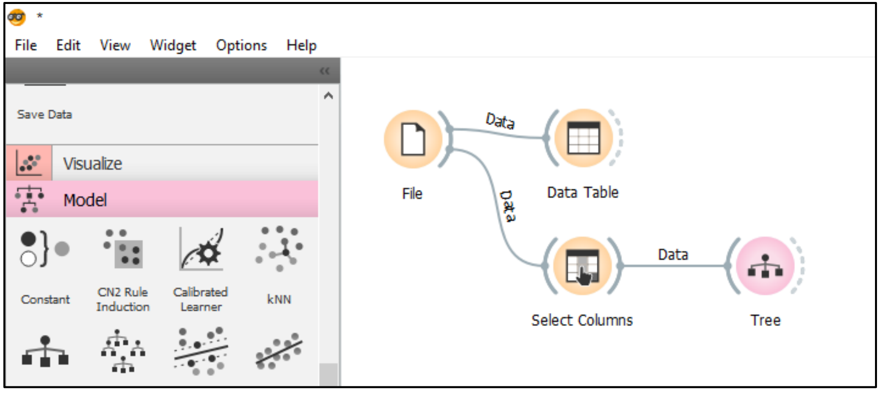


Figure 6. Display of select columns widget linked to the widget tree in orange software

- Drag the Tree Viewer widget in the worksheet section to see the result obtained, then connect it.

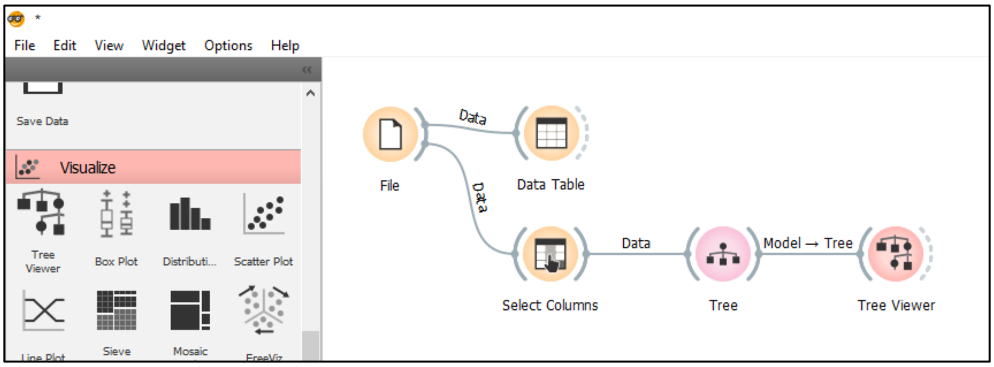
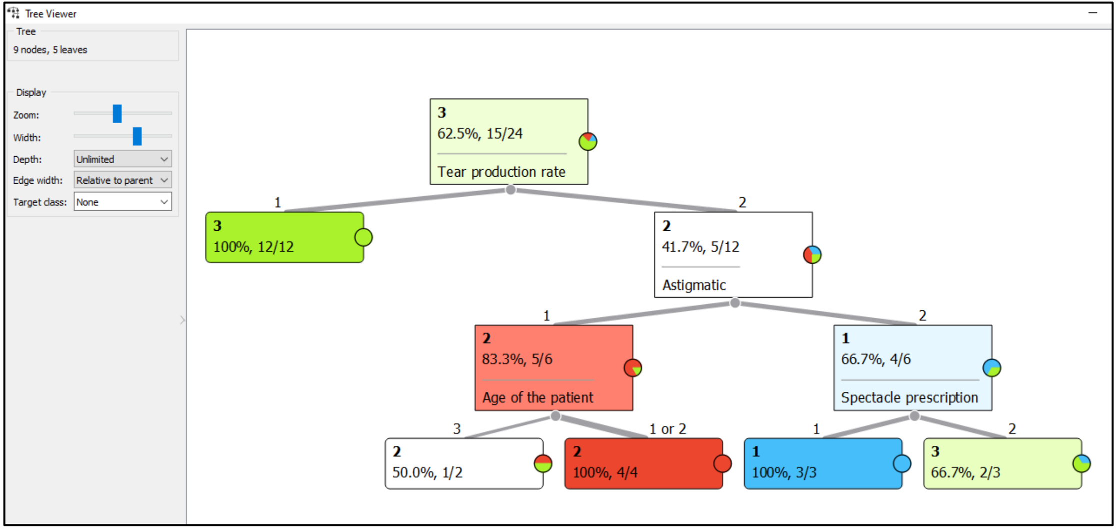


Figure 7. Display widget tree linked to widget tree viewer in orange software

- The next step is to double-click on the Tree Viewer widget to see the existing output. The output is shown in the image below.


Tree Viewer

Tree
9 nodes, 5 leaves

Display
Zoom:
Width:
Depth: Unlimited
Edge width: Relative to parent
Target class: None

Figure 8. Display tree viewer properties in orange software

- To see the evaluation results from the dataset it can be done by selecting the Predictions widget on the evaluate tab. Drag on the canvas and connect with the widget tree and file.

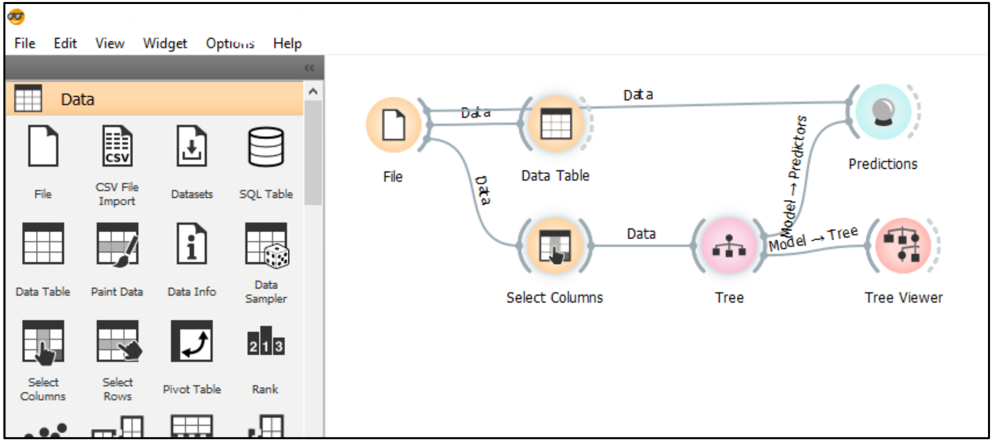
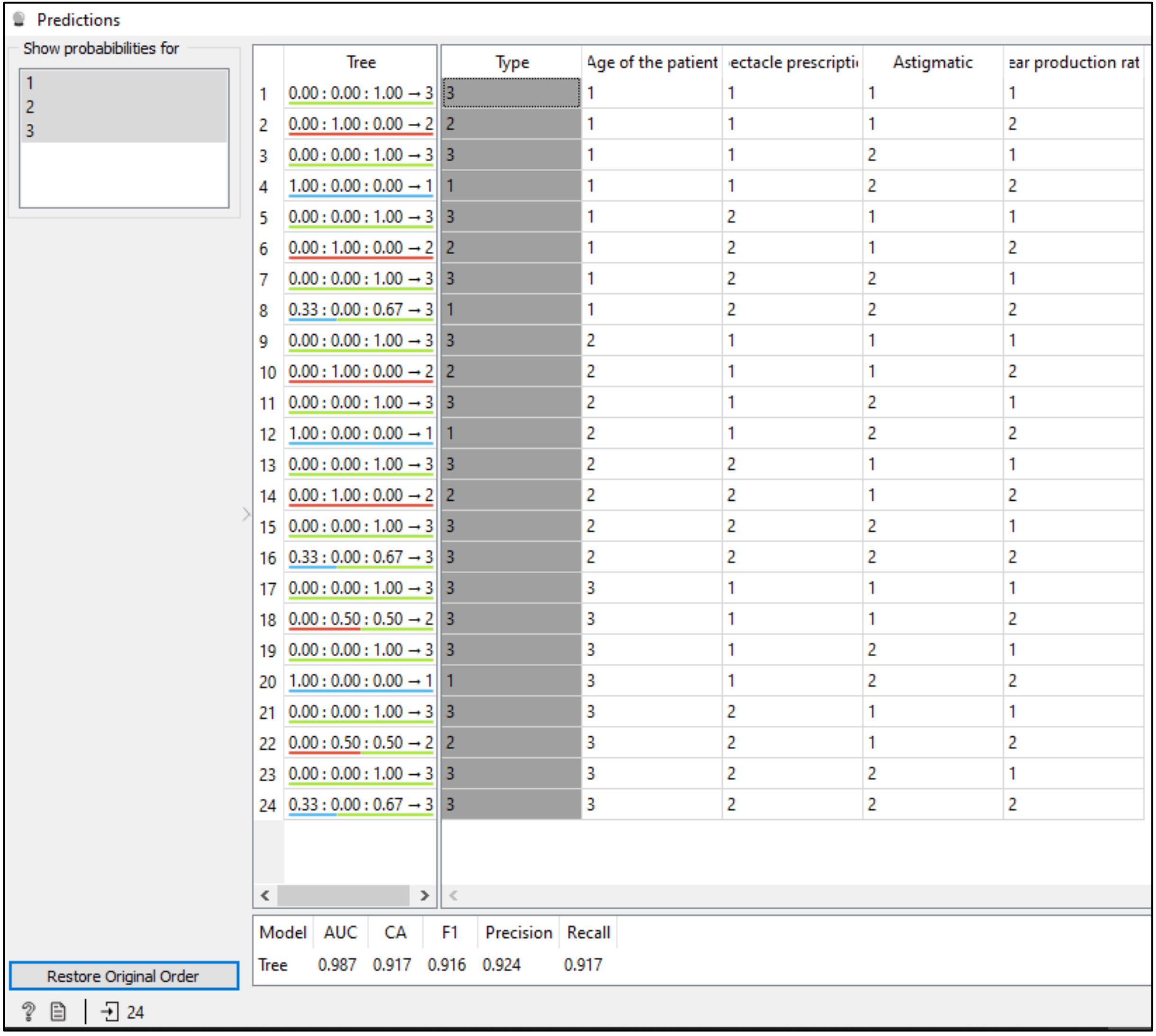


Figure 9. Display widget predictions that are connected

- Double-click on the Predictions widget to see the evaluation values that came out.



	Tree	Type	Age of the patient	ectacle prescripti	Astigmatic	ear production rat
1	0.00 : 0.00 : 1.00 → 3	3	1	1	1	1
2	0.00 : 1.00 : 0.00 → 2	2	1	1	1	2
3	0.00 : 0.00 : 1.00 → 3	3	1	1	2	1
4	1.00 : 0.00 : 0.00 → 1	1	1	1	2	2
5	0.00 : 0.00 : 1.00 → 3	3	1	2	1	1
6	0.00 : 1.00 : 0.00 → 2	2	1	2	1	2
7	0.00 : 0.00 : 1.00 → 3	3	1	2	2	1
8	0.33 : 0.00 : 0.67 → 3	1	1	2	2	2
9	0.00 : 0.00 : 1.00 → 3	3	2	1	1	1
10	0.00 : 1.00 : 0.00 → 2	2	2	1	1	2
11	0.00 : 0.00 : 1.00 → 3	3	2	1	2	1
12	1.00 : 0.00 : 0.00 → 1	1	2	1	2	2
13	0.00 : 0.00 : 1.00 → 3	3	2	2	1	1
14	0.00 : 1.00 : 0.00 → 2	2	2	2	1	2
15	0.00 : 0.00 : 1.00 → 3	3	2	2	2	1
16	0.33 : 0.00 : 0.67 → 3	3	2	2	2	2
17	0.00 : 0.00 : 1.00 → 3	3	3	1	1	1
18	0.00 : 0.50 : 0.50 → 2	3	3	1	1	2
19	0.00 : 0.00 : 1.00 → 3	3	3	1	2	1
20	1.00 : 0.00 : 0.00 → 1	1	3	1	2	2
21	0.00 : 0.00 : 1.00 → 3	3	3	2	1	1
22	0.00 : 0.50 : 0.50 → 2	2	3	2	1	2
23	0.00 : 0.00 : 1.00 → 3	3	3	2	2	1
24	0.33 : 0.00 : 0.67 → 3	3	3	2	2	2

Model	AUC	CA	F1	Precision	Recall
Tree	0.987	0.917	0.916	0.924	0.917

Figure 10. Display properties predictions in orange software

Tree construction begins with the formation of roots (located at the top). Then the data is divided based on attributes that are suitable to be used as leaves. Decision rule formation is to make decision rules from trees that have been formed. The rule can be in the form of if then derived from the decision tree by tracing from root to leaf.
Based on the dataset processing using the tree method using Orange software, the accuracy of the Decision Tree classification process is obtained, the value of pecision is 92,4% which means Decision Tree method is good.

4. Conclusion

The conclusions that can be obtained from the results of the description and discussion can be seen in Figure 8.

- If tear production rate is 1, then type of patient is 3
- If tear production rate is 2, then 41,7% type of patient is 2
- If tear production rate is 2 and astigmatic is 1, then 83,3% type of patient is 2
- If tear production rate is 2 and astigmatic is 2, then 66,7% type of patient is 1
- If tear production rate is 2, astigmatic is 1, and age of patient is 3, then 50% type of patient is 2
- If tear production rate is 2, astigmatic is 1, and age of patient is 1 or 2, then type of patient is 2
- If tear production rate is 2, astigmatic is 2 and spectacle prescription is 1, then type of patient is 1
- If tear production rate is 2, astigmatic is 2 and spectacle prescription is 2, then 66,7% type of patient is 3

For each node and its branch will be given in if, while the value of the value of the leaf will be written in then. After all the rules are made, the rules can be simplified or combines

Acknowledgements

The author would like to thank all those who helped, assisted and guided in the making of this research in a journal.

References

[1] Hendrian S 2018 *Faktor Exacta* **11** (3) pp 267-274
[2] Mardi Y 2016 *Edik Informatika* **2** pp 213-219
[3] Meilina P 2014 Penerapan Data Mining Dengan Metode Klasifikasi Menggunakan Decision Tree dan Regresi *Teknologi* **7** (1) pp 11-20
[4] Pietersz E, Sumual V and Rares L 2016 Penggunaan Lensa Kontak dan Pengaruhnya Terhadap Dry Eyes Pada Mahasiswa Fakultas Ekonomi Sam Ratulangi *e-CliniC* **4** (1)
[5] Oktanisa I and Supianto A 2017 *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)* **5** (5) pp 567-576
[6] Kukasvadiya M and Divecha N 2017 *International Journal of Engineering Development and Research*, **5** (2) pp 1836-1840
[7] Ambarsari E, Khotijah S and Sunarmintyastuti L 2019 *STRING (Satuan Tulisan Riset dan Inovasi Teknologi)* **4** (1) pp 9-17
[8] Aulia Ishak and Tommy Wijaya 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **801** 012118
[9] Herviani V and Febriansya A 2016 *Riset Akuntansi* **8** (2) pp 19-27
[10] Utomo D and Mesran 2020 *Jurnal Media Informatika Budidarma* **4** (2) 437-444
[11] Irawan N, Wijono and Setyawati O 2017 Perbaikan Missing Value Menggunakan Pendekatan Korelasi Pada Metode K-Nearest Neighbor *Jurnal Infotel* **9** (3) pp 305-311
[12] Indrayanto G 2018 *Natural Product Communications*, **13** (12)