

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»**

Інститут Комп'ютерних наук та інформаційних технологій

Кафедра Програмної інженерії та інтелектуальних технологій управління

Спеціальність 122 Комп'ютерні науки

Освітня програма Комп'ютерні науки та інтелектуальні системи

ЛАБОРАТОРНА РАБОТА №3 за курсом

«Інтелектуальний аналіз даних та видобування знань»

Тема лабораторної роботи Класифікація за допомогою WEKA

Виконав студент 5 курсу, групи КН-М422

Захар ПАРАХІН

(підпис, прізвище та ініціали)

Перевірила Оксана ІВАЩЕНКО

(підпис, прізвище та ініціали)

Харків 2022

ЗМІСТ

Вступ	3
1 Хід виконання роботи	4
1.1 Класифікація	4
1.2 WEKA Classification	5
1.2.1 Classification with Decision Tree	6
1.2.2 Classification with Naive Bayes	10
Висновки	15
Список джерел інформації	16

Вступ

Інтелектуальний аналіз даних у загальних рисах означає пошук або глибоке копання в даних, які знаходяться в різних формах, щоб отримати шаблони та отримати знання про цей шаблон. У процесі інтелектуального аналізу даних великі набори даних спочатку сортуються, потім визначаються закономірності та встановлюються зв'язки для виконання аналізу даних і вирішення проблем.

Класифікація: це завдання аналізу даних, тобто процес пошуку моделі, яка описує та розрізняє класи та поняття даних. Класифікація — це проблема ідентифікації, до якої з набору категорій належить нове спостереження, на основі навчального набору даних, що містить спостереження та приналежність до категорій які відомі.

1 Хід виконання роботи

1.1 Класифікація

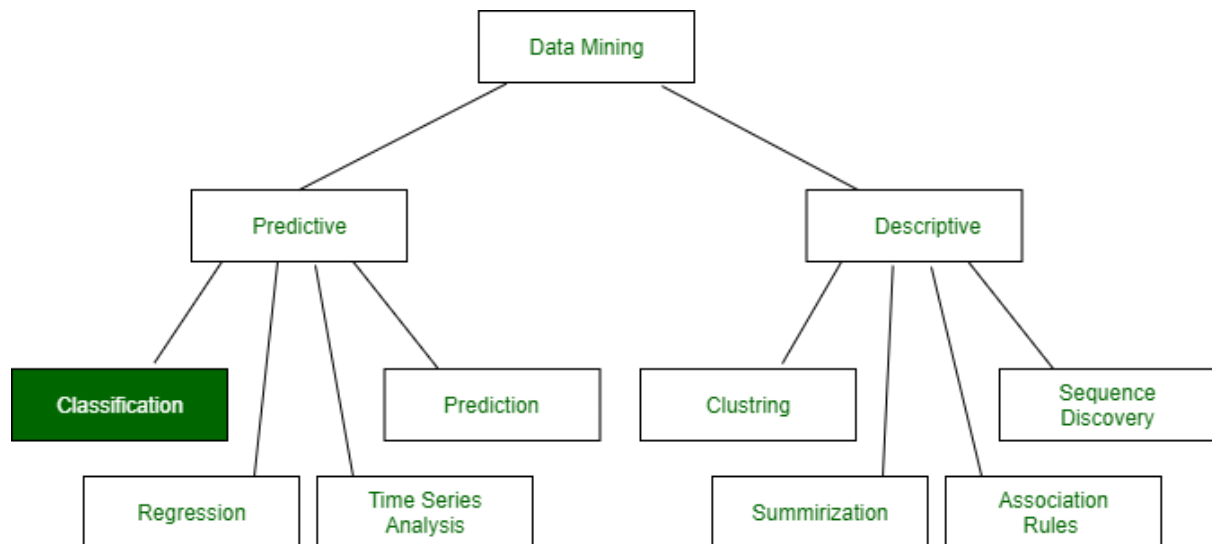


Рисунок 1 - розміщення Класифікації серед інших складових Data Mining

Розглянемо дерево рішень. Алгоритм дерева рішень відноситься до категорії навчання під наглядом. Вони можуть бути використані для розв’язання задач регресії та класифікації.

Дерево рішень використовує представлення дерева для вирішення проблеми, у якій кожен вузол відповідає мітці класу, а атрибути представлені у внутрішньому вузлі дерева. Можна представити будь-яку булеву функцію на дискретних атрибутах за допомогою дерева рішень.

C4.5: Цей алгоритм є наступником алгоритму ID3. Цей алгоритм використовує підсилення інформації або коефіцієнт підсилення для визначення атрибута класифікації. Це пряме вдосконалення алгоритму ID3, оскільки він може обробляти як безперервні, так і відсутні значення атрибутів.

Naive Bayes classifiers — це набір алгоритмів класифікації на основі теореми Байеса. Це не один алгоритм, а сімейство алгоритмів, де всі вони мають загальний принцип, тобто кожна пара ознак, що класифікуються, не залежить одна від одної.

1.2 WEKA Classification

Вгорі вкладки "Classify" (Класифікація) знаходиться область вибору класифікатора "Classifier". Вибір та налаштування параметрів класифікатора подібні до вибору фільтра попередньої обробки даних. Результат застосування вибраного класифікатора буде протестований згідно з параметрами, заданими в області Test Options.

На панелі «Test options» визначається метод тестування отриманого класифікатора: на навчальній вибірці (use training set), на тестовій вибірці з окремого файлу (supplied test set), по блоках (cross-validation), за допомогою розділення вихідної вибірки на навчання та контроль (Percentage split). При виборі деяких опцій доведеться вказати параметри тестування. Наприклад, при виборі "cross-validation" треба вказати, на скільки блоків (фолдів) розбивати вибірку.

Для обробки береться датасет bridges-version2.arff

Viewer

Relation: bridges-version2

No.	1: IDENTIF Nominal	2: RIVER Nominal	3: LOCATION Nominal	4: ERCTED Nominal	5: PURPOSE Nominal	6: LENGTH Nominal	7: LANES Nominal	8: CLEAR-G Nominal	9: T-OR-D Nominal	10: MATERIAL Nominal	11: SPAN Nominal	12: REL-L Nominal	13: TYPE Nominal
1	E1	M	3	CRAFTS	HIGHWAY		2	N	THROUGH	WOOD	SHORT	S	WOOD
2	E2	A	25	CRAFTS	HIGHWAY	MEDIUM	2	N	THROUGH	WOOD	SHORT	S	WOOD
3	E3	A	39	CRAFTS	AQUEDUCT		1	N	THROUGH	WOOD		S	WOOD
4	E5	A	29	CRAFTS	HIGHWAY	MEDIUM	2	N	THROUGH	WOOD	SHORT	S	WOOD
5	E6	M	23	CRAFTS	HIGHWAY		2	N	THROUGH	WOOD		S	WOOD
6	E7	A	27	CRAFTS	HIGHWAY	SHORT	2	N	THROUGH	WOOD	MEDIUM	S	WOOD
7	E8	A	28	CRAFTS	AQUEDUCT	MEDIUM	1	N	THROUGH	IRON	SHORT	S	SUSPEN
8	E9	M	3	CRAFTS	HIGHWAY	MEDIUM	2	N	THROUGH	IRON	SHORT	S	SUSPEN
9	E10	A	39	CRAFTS	AQUEDUCT		1	N	DECK	WOOD		S	WOOD
10	E11	A	29	CRAFTS	HIGHWAY	MEDIUM	2	N	THROUGH	WOOD	MEDIUM	S	WOOD
11	E12	A	39	CRAFTS	RR		2	N	DECK	WOOD		S	WOOD
12	E14	M	6	CRAFTS	HIGHWAY	MEDIUM	2	N	THROUGH	WOOD	MEDIUM	S	WOOD
13	E13	A	33	CRAFTS	HIGHWAY		2	N	THROUGH	WOOD		S	WOOD
14	E15	A	28	CRAFTS	RR		2	N	THROUGH	WOOD		S	WOOD
15	E16	A	25	CRAFTS	HIGHWAY	MEDIUM	2	N	THROUGH	IRON	MEDIUM	S-F	SUSPEN
16	E17	M	4	CRAFTS	RR	MEDIUM	2	N	THROUGH	IRON	MEDIUM		SIMPLE-T
17	E18	A	28	CRAFTS	RR	MEDIUM	2	N	THROUGH	IRON	SHORT	S	SIMPLE-T
18	E19	A	29	CRAFTS	HIGHWAY	MEDIUM	2	N	THROUGH	WOOD	MEDIUM	S	WOOD
19	E20	A	32	EMERGING	HIGHWAY	MEDIUM	2	N	THROUGH	WOOD	MEDIUM	S	WOOD
20	E21	M	16	EMERGING	RR		2		THROUGH	IRON			SIMPLE-T
21	E23	M	1	EMERGING	HIGHWAY	MEDIUM			THROUGH	STEEL	LONG	F	SUSPEN
22	E22	A	24	EMERGING	HIGHWAY	MEDIUM	4	G	THROUGH	WOOD	SHORT	S	WOOD
23	E24	O	45	EMERGING	RR		2	G		STEEL			SIMPLE-T
24	E25	M	10	EMERGING	RR		2	G		STEEL			SIMPLE-T
25	E27	A	39	EMERGING	RR		2	G	THROUGH	STEEL		F	SIMPLE-T
26	E26	M	12	EMERGING	RR	MEDIUM	2	G	THROUGH	STEEL	MEDIUM	S	SIMPLE-T
27	E30	A	31	EMERGING	RR		2	G	THROUGH	STEEL	MEDIUM	F	SIMPLE-T
28	E29	A	26	EMERGING	HIGHWAY	MEDIUM	2	G	THROUGH	STEEL	MEDIUM		SUSPEN
29	E28	M	3	EMERGING	HIGHWAY	MEDIUM	2	G	THROUGH	STEEL	MEDIUM	S	ARCH
30	E32	A	30	EMERGING	HIGHWAY		2	G	THROUGH	IRON	MEDIUM	F	SIMPLE-T

Рисунок 2 - bridges-version2.arff

Як видно із змісту даного датасету і його метаданих, то він вже має опис після фільтра Discretizion.

1. IDENTIF / -- / -- / identifier of the examples
2. RIVER / n / A, M, O / --
3. LOCATION / n / 1 to 52 / --
4. ERECTED / c,n / 1818-1986 ; CRAFTS, EMERGING, MATURE, MODERN / --
5. PURPOSE / n / WALK, AQUEDUCT, RR, HIGHWAY / --
6. LENGTH / c,n / 804-4558 ; SHORT, MEDIUM, LONG / --
7. LANES / c,n / 1, 2, 4, 6 ; 1, 2, 4, 6 / --
8. CLEAR-G / n / N, G / --
9. T-OR-D / n / THROUGH, DECK / --
10. MATERIAL / n / WOOD, IRON, STEEL / --
11. SPAN / n / SHORT, MEDIUM, LONG / --
12. REL-L / n / S, S-F, F / --
13. TYPE / n / WOOD, SUSPEN, SIMPLE-T, ARCH, CANTILEV, CONT-T / --

Рисунок 3 - інформація атрибутів

1.2.1 Classification with Decision Tree

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier: Choose **J48 -C 0.25 -M 2**

Test options:

- ☒ Use training set
- ☐ Supplied test set (Set...)
- ☐ Cross-validation (Folds: 10)
- ☐ Percentage split (%: 66)
- More options...

(Nom) TYPE: **Start** **Stop**

Result list (right-click for options):

- 05:50:12 - trees.J48
- 06:16:04 - trees.J48**

Classifier output:

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	59	56.1905 %
Incorrectly Classified Instances	46	43.8095 %
Kappa statistic	0.392	
Mean absolute error	0.1643	
Root mean squared error	0.3124	
Relative absolute error	64.836 %	
Root relative squared error	87.9681 %	
Total Number of Instances	105	
Ignored Class Unknown Instances	2	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	WOOD
	0,000	0,043	0,000	0,000	0,000	-0,068	0,694	0,176	SUSPEN
	0,727	0,377	0,582	0,727	0,646	0,346	0,777	0,654	SIMPLE-T
	0,385	0,152	0,263	0,385	0,313	0,199	0,747	0,321	ARCH
	0,182	0,021	0,500	0,182	0,267	0,257	0,696	0,221	CANTILEV
	0,400	0,032	0,571	0,400	0,471	0,434	0,707	0,429	CONT-T
Weighted Avg.	0,562	0,187	0,536	0,562	0,535	0,383	0,783	0,549	

=== Confusion Matrix ===

	a	b	c	d	e	f	<-- classified as
16	0	0	0	0	0	0	a = WOOD
0	0	7	4	0	0	0	b = SUSPEN
0	2	32	8	0	2	0	c = SIMPLE-T
0	0	7	5	1	0	0	d = ARCH
0	2	5	1	2	1	0	e = CANTILEV
0	0	4	1	1	4	0	f = CONT-T

Рисунок 4 - використання J48 на датасеті

За результатами цього алгоритму, отримуємо точність класифікації близько 56%, і отримані наступні класи:

Wood	Cantilev
Suspen	Cont-T
Simple-T	
Arch	

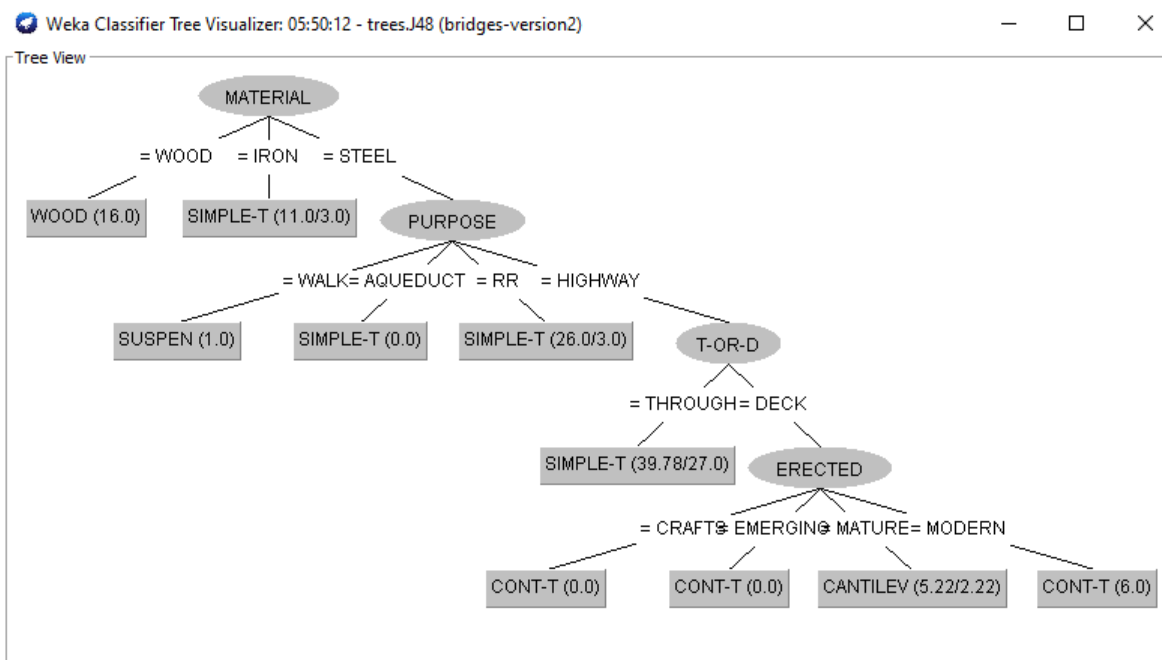
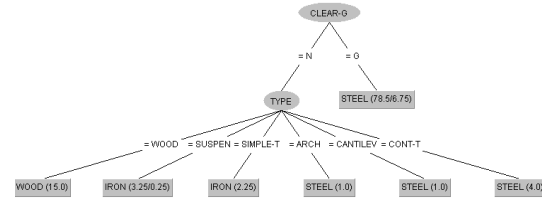
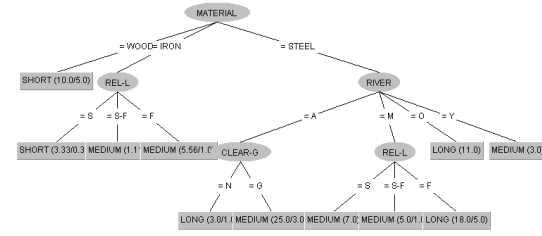
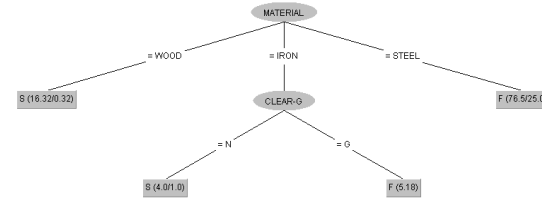
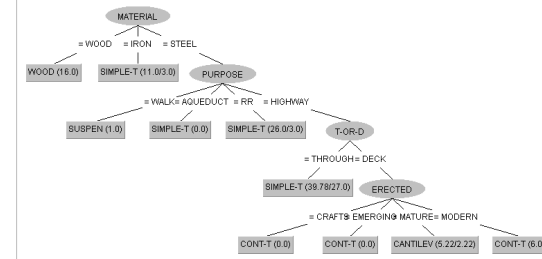


Рисунок 5 - побудоване Decision Tree

Та використаємо по різним атрибутам і оберемо з найбільшою точністю:

Атрибут	Точність (%)	Дерево
IDENTIF	55,1	
RIVER	99	
LOCATION	48,1	

ERECTED	75,7	<pre> graph TD CLEAR-G --> N MATERIAL CLEAR-G --> 0 TYPE MATERIAL --> WOOD MATERIAL --> IRON MATERIAL --> STEEL TYPE --> WOOD TYPE --> SUSPENSION TYPE --> SIMPLE-T TYPE --> ARCH TYPE --> CANTILEV TYPE --> CONT-T WOOD --> CRAFTS["CRAFTS (1.5)"] WOOD --> EMERGING["EMERGING (1.1)"] WOOD --> MATURE["MATURE (6.2)"] SUSPENSION --> MATURE2["MATURE (2.41)"] SUSPENSION --> MATURE3["MATURE (4.92)"] SUSPENSION --> MODERN["MODERN (4.92)"] SIMPLE-T --> MATURE4["MATURE (1.1)"] SIMPLE-T --> MATURE5["MATURE (2.41)"] SIMPLE-T --> MATURE6["MATURE (4.92)"] SIMPLE-T --> MODERN2["MODERN (4.92)"] ARCH --> MATURE7["MATURE (1.1)"] ARCH --> MATURE8["MATURE (2.41)"] ARCH --> MATURE9["MATURE (4.92)"] ARCH --> MODERN3["MODERN (4.92)"] CANTILEV --> MATURE10["MATURE (1.1)"] CANTILEV --> MATURE11["MATURE (2.41)"] CANTILEV --> MATURE12["MATURE (4.92)"] CANTILEV --> MODERN4["MODERN (4.92)"] CONT-T --> MATURE13["MATURE (1.1)"] CONT-T --> MATURE14["MATURE (2.41)"] CONT-T --> MATURE15["MATURE (4.92)"] CONT-T --> MODERN5["MODERN (4.92)"] </pre>
PURPOSE	66,3	<p>HIGHWAY (107.0/36.0)</p>
LENGTH	59,2	<p>MEDIUM (81.0/33.0)</p>
LANES	89,1	<pre> graph TD PURPOSE --> VALUED PURPOSE --> UNVALUED PURPOSE --> HIGHWAY VALUED --> ERECTED["ERECTED (2.0/0.0)"] VALUED --> 1["1 (4.0)"] VALUED --> 2["2 (28.0/1.0)"] ERECTED --> CRAFT ERECTED --> EMERGING ERECTED --> MATURE ERECTED --> MODERN CRAFT --> CRAFT2["CRAFT (2.11.0)"] CRAFT --> CRAFT3["CRAFT (5.01.0)"] EMERGING --> CLEAR-G EMERGING --> TYPE CLEAR-G --> N CLEAR-G --> G N --> TYPE2 TYPE2 --> WOOD TYPE2 --> SUSPENSION TYPE2 --> SIMPLE-T TYPE2 --> ARCH TYPE2 --> CANTILEV TYPE2 --> CONT-T WOOD --> CRAFTS2["CRAFTS (2.0/0.0)"] WOOD --> EMERGING2["EMERGING (4.5.01.0)"] WOOD --> MATURE2["MATURE (2.12.01.0)"] WOOD --> MODERN2["MODERN (4.5.01.0)"] SUSPENSION --> MATURE3["MATURE (2.6.02.0)"] SUSPENSION --> MODERN3["MODERN (2.0/0.0)"] SIMPLE-T --> MATURE4["MATURE (4.0/0.0)"] SIMPLE-T --> MODERN4["MODERN (4.0/0.0)"] ARCH --> MATURE5["MATURE (2.1.0)"] ARCH --> MODERN5["MODERN (6.4.0)"] CANTILEV --> MATURE6["MATURE (2.2.01.0)"] CANTILEV --> MODERN6["MODERN (4.6.02.0)"] CONT-T --> MATURE7["MATURE (4.0/0.0)"] CONT-T --> MODERN7["MODERN (4.0/0.0)"] </pre>
CLEAR-G	93,3	<pre> graph TD MATERIAL --> WOOD MATERIAL --> IRON MATERIAL --> STEEL WOOD --> ERECTED["ERECTED (N (16.31/1.31), G (78.56.0))"] IRON --> ERECTED STEEL --> ERECTED ERECTED --> CRAFTS ERECTED --> EMERGING ERECTED --> MATURE ERECTED --> MODERN CRAFTS --> CRAFTS2["CRAFTS (N (5.0))"] EMERGING --> EMERGING2["EMERGING (G (2.0))"] MATURE --> MATURE2["MATURE (G (3.0))"] MODERN --> MODERN2["MODERN (G (0.19))"] </pre>
T-OR-D	93,1	<pre> graph TD TYPE --> WOOD TYPE --> SUSPENSION TYPE --> SIMPLE-T TYPE --> ARCH TYPE --> CANTILEV TYPE --> CONT-T WOOD --> THROUGH1["THROUGH (16.02.0)"] SUSPENSION --> THROUGH2["THROUGH (11.0)"] SIMPLE-T --> THROUGH3["THROUGH (41.01.0)"] ARCH --> THROUGH4["THROUGH (13.01.0)"] CANTILEV --> SPAN CONT-T --> DECK["DECK (10.02.0)"] SPAN --> SHORT SPAN --> MEDIUM SPAN --> LONG SHORT --> THROUGH5["THROUGH (0.0)"] MEDIUM --> DECK2["DECK (4.01.0)"] LONG --> THROUGH6["THROUGH (7.0)"] </pre>

MATERIAL	93,3	
SPAN	82,6	
REL-L	74,5	
TYPE	65,7	

Таблиця 1 - результати J48 за атрибутами

Серед всіх отриманих за атрибутами моделей, найточніша RIVER, є найбільш розбитою на класи, що робить її незручною, тому беремо наступну за точністю CLEAR-G.

Там отримали наступні класи: N і G, а також DECISION TREE:

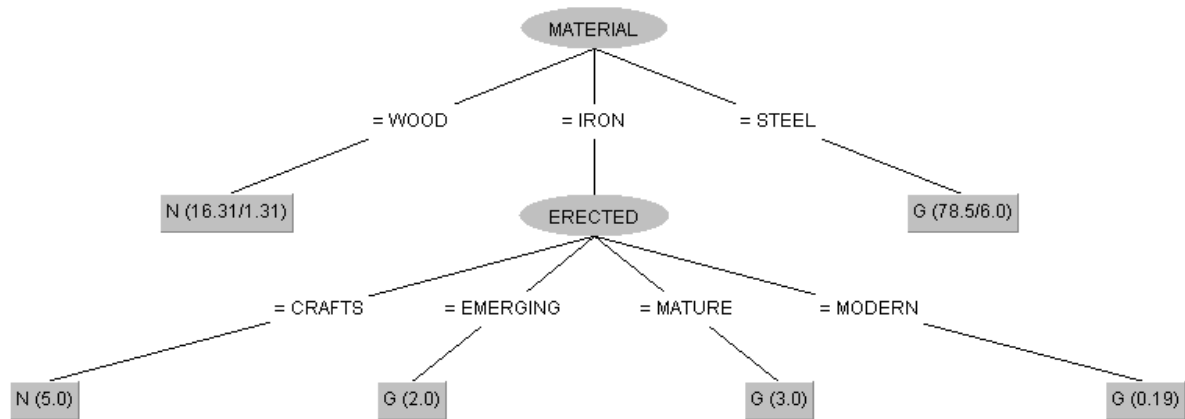


Рисунок 6 - побудоване Decision Tree

1.2.2 Classification with Naive Bayers

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose NaiveBayes

Test options

☒ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 10

☐ Percentage split % 66

More options...

(Nom) TYPE

Start Stop

Result list (right-click for options)

05:50:12 - trees.J48

06:16:04 - trees.J48

06:24:16 - bayes.NaiveBayes

Classifier output

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	92	87.619 %
Incorrectly Classified Instances	13	12.381 %
Kappa statistic	0.8362	
Mean absolute error	0.0661	
Root mean squared error	0.1758	
Relative absolute error	26.1066 %	
Root relative squared error	49.5472 %	
Total Number of Instances	105	
Ignored Class Unknown Instances	2	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	1,000	0,034	0,842	1,000	0,914	0,902	0,999	0,993	WOOD
	0,545	0,000	1,000	0,545	0,706	0,720	0,979	0,879	SUSPEN
	0,932	0,033	0,953	0,932	0,943	0,902	0,979	0,979	SIMPLE-T
	0,923	0,043	0,750	0,923	0,828	0,806	0,984	0,918	ARCH
	0,818	0,032	0,750	0,818	0,783	0,757	0,991	0,931	CANTILEV
	0,800	0,011	0,889	0,800	0,842	0,828	0,995	0,961	CONT-T
Weighted Avg.	0,876	0,029	0,889	0,876	0,873	0,849	0,986	0,957	

=== Confusion Matrix ===

```

a b c d e f <-- classified as
16 0 0 0 0 0 | a = WOOD
2 6 0 1 2 0 | b = SUSPEN
1 0 41 2 0 0 | c = SIMPLE-T
0 0 1 12 0 0 | d = ARCH
0 0 1 0 9 1 | e = CANTILEV
0 0 0 1 1 8 | f = CONT-T
  
```

Рисунок 7- використання NaiveBayers

По результатам цього алгоритму, маємо вищу точність ніж при J48, яка складає близько 87,6%. І отримані тіж самі класи:

Wood

Cantilev

Suspen
Simple-T
Arch

Cont-T

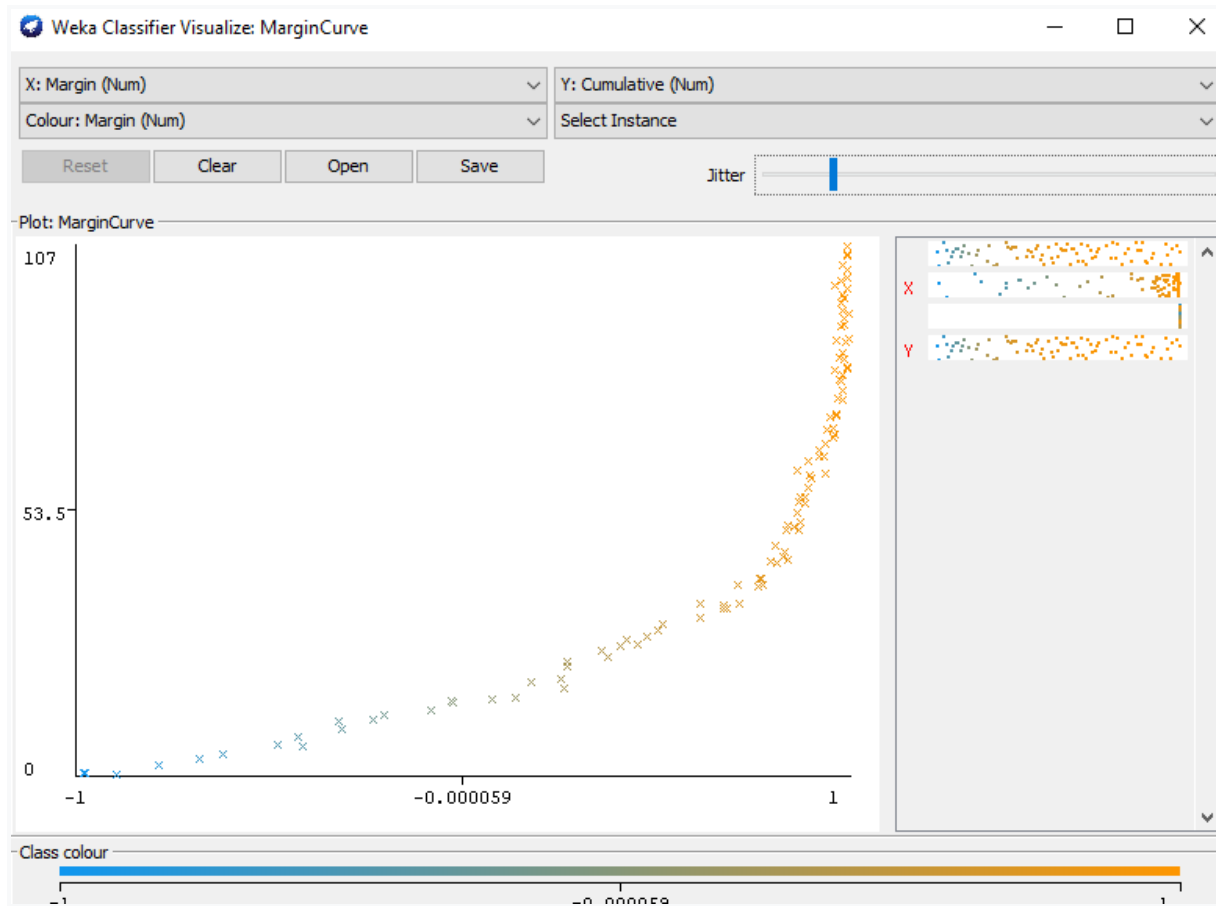


Рисунок 8 - візуалізація прогнозування від отриманої моделі

Тепер використаємо по різним атрибутам і оберемо з найбільшою точністю:

Атрибут	Точність (%)	WEKA результати
IDENTIF	99	<pre> === Summary === Correctly Classified Instances 106 99.0654 % Incorrectly Classified Instances 1 0.9346 % Kappa statistic 0.9906 Mean absolute error 0.0133 Root mean squared error 0.0715 Relative absolute error 72.7103 % Root relative squared error 74.626 % Total Number of Instances 107 </pre>

RIVER	86,9	<p>=== Summary ===</p> <table> <tr> <td>Correctly Classified Instances</td><td>93</td><td>86.9159 %</td></tr> <tr> <td>Incorrectly Classified Instances</td><td>14</td><td>13.0841 %</td></tr> <tr> <td>Kappa statistic</td><td>0.7933</td><td></td></tr> <tr> <td>Mean absolute error</td><td>0.1155</td><td></td></tr> <tr> <td>Root mean squared error</td><td>0.2211</td><td></td></tr> <tr> <td>Relative absolute error</td><td>36.3466 %</td><td></td></tr> <tr> <td>Root relative squared error</td><td>55.6399 %</td><td></td></tr> <tr> <td>Total Number of Instances</td><td>107</td><td></td></tr> </table>	Correctly Classified Instances	93	86.9159 %	Incorrectly Classified Instances	14	13.0841 %	Kappa statistic	0.7933		Mean absolute error	0.1155		Root mean squared error	0.2211		Relative absolute error	36.3466 %		Root relative squared error	55.6399 %		Total Number of Instances	107				
Correctly Classified Instances	93	86.9159 %																											
Incorrectly Classified Instances	14	13.0841 %																											
Kappa statistic	0.7933																												
Mean absolute error	0.1155																												
Root mean squared error	0.2211																												
Relative absolute error	36.3466 %																												
Root relative squared error	55.6399 %																												
Total Number of Instances	107																												
LOCATION	78,3	<p>=== Summary ===</p> <table> <tr> <td>Correctly Classified Instances</td><td>83</td><td>78.3019 %</td></tr> <tr> <td>Incorrectly Classified Instances</td><td>23</td><td>21.6981 %</td></tr> <tr> <td>Kappa statistic</td><td>0.7778</td><td></td></tr> <tr> <td>Mean absolute error</td><td>0.0219</td><td></td></tr> <tr> <td>Root mean squared error</td><td>0.093</td><td></td></tr> <tr> <td>Relative absolute error</td><td>60.5256 %</td><td></td></tr> <tr> <td>Root relative squared error</td><td>69.127 %</td><td></td></tr> <tr> <td>Total Number of Instances</td><td>106</td><td></td></tr> <tr> <td>Ignored Class Unknown Instances</td><td>1</td><td></td></tr> </table>	Correctly Classified Instances	83	78.3019 %	Incorrectly Classified Instances	23	21.6981 %	Kappa statistic	0.7778		Mean absolute error	0.0219		Root mean squared error	0.093		Relative absolute error	60.5256 %		Root relative squared error	69.127 %		Total Number of Instances	106		Ignored Class Unknown Instances	1	
Correctly Classified Instances	83	78.3019 %																											
Incorrectly Classified Instances	23	21.6981 %																											
Kappa statistic	0.7778																												
Mean absolute error	0.0219																												
Root mean squared error	0.093																												
Relative absolute error	60.5256 %																												
Root relative squared error	69.127 %																												
Total Number of Instances	106																												
Ignored Class Unknown Instances	1																												
ERECTED	87,8	<p>=== Summary ===</p> <table> <tr> <td>Correctly Classified Instances</td><td>94</td><td>87.8505 %</td></tr> <tr> <td>Incorrectly Classified Instances</td><td>13</td><td>12.1495 %</td></tr> <tr> <td>Kappa statistic</td><td>0.8173</td><td></td></tr> <tr> <td>Mean absolute error</td><td>0.1052</td><td></td></tr> <tr> <td>Root mean squared error</td><td>0.2165</td><td></td></tr> <tr> <td>Relative absolute error</td><td>31.3393 %</td><td></td></tr> <tr> <td>Root relative squared error</td><td>52.9671 %</td><td></td></tr> <tr> <td>Total Number of Instances</td><td>107</td><td></td></tr> </table>	Correctly Classified Instances	94	87.8505 %	Incorrectly Classified Instances	13	12.1495 %	Kappa statistic	0.8173		Mean absolute error	0.1052		Root mean squared error	0.2165		Relative absolute error	31.3393 %		Root relative squared error	52.9671 %		Total Number of Instances	107				
Correctly Classified Instances	94	87.8505 %																											
Incorrectly Classified Instances	13	12.1495 %																											
Kappa statistic	0.8173																												
Mean absolute error	0.1052																												
Root mean squared error	0.2165																												
Relative absolute error	31.3393 %																												
Root relative squared error	52.9671 %																												
Total Number of Instances	107																												
PURPOSE	81,3	<p>=== Summary ===</p> <table> <tr> <td>Correctly Classified Instances</td><td>87</td><td>81.3084 %</td></tr> <tr> <td>Incorrectly Classified Instances</td><td>20</td><td>18.6916 %</td></tr> <tr> <td>Kappa statistic</td><td>0.6572</td><td></td></tr> <tr> <td>Mean absolute error</td><td>0.103</td><td></td></tr> <tr> <td>Root mean squared error</td><td>0.233</td><td></td></tr> <tr> <td>Relative absolute error</td><td>42.5381 %</td><td></td></tr> <tr> <td>Root relative squared error</td><td>67.6495 %</td><td></td></tr> <tr> <td>Total Number of Instances</td><td>107</td><td></td></tr> </table>	Correctly Classified Instances	87	81.3084 %	Incorrectly Classified Instances	20	18.6916 %	Kappa statistic	0.6572		Mean absolute error	0.103		Root mean squared error	0.233		Relative absolute error	42.5381 %		Root relative squared error	67.6495 %		Total Number of Instances	107				
Correctly Classified Instances	87	81.3084 %																											
Incorrectly Classified Instances	20	18.6916 %																											
Kappa statistic	0.6572																												
Mean absolute error	0.103																												
Root mean squared error	0.233																												
Relative absolute error	42.5381 %																												
Root relative squared error	67.6495 %																												
Total Number of Instances	107																												
LENGTH	90,1	<p>=== Summary ===</p> <table> <tr> <td>Correctly Classified Instances</td><td>73</td><td>90.1235 %</td></tr> <tr> <td>Incorrectly Classified Instances</td><td>8</td><td>9.8765 %</td></tr> <tr> <td>Kappa statistic</td><td>0.8314</td><td></td></tr> <tr> <td>Mean absolute error</td><td>0.1429</td><td></td></tr> <tr> <td>Root mean squared error</td><td>0.2351</td><td></td></tr> <tr> <td>Relative absolute error</td><td>38.0471 %</td><td></td></tr> <tr> <td>Root relative squared error</td><td>54.4194 %</td><td></td></tr> <tr> <td>Total Number of Instances</td><td>81</td><td></td></tr> <tr> <td>Ignored Class Unknown Instances</td><td>26</td><td></td></tr> </table>	Correctly Classified Instances	73	90.1235 %	Incorrectly Classified Instances	8	9.8765 %	Kappa statistic	0.8314		Mean absolute error	0.1429		Root mean squared error	0.2351		Relative absolute error	38.0471 %		Root relative squared error	54.4194 %		Total Number of Instances	81		Ignored Class Unknown Instances	26	
Correctly Classified Instances	73	90.1235 %																											
Incorrectly Classified Instances	8	9.8765 %																											
Kappa statistic	0.8314																												
Mean absolute error	0.1429																												
Root mean squared error	0.2351																												
Relative absolute error	38.0471 %																												
Root relative squared error	54.4194 %																												
Total Number of Instances	81																												
Ignored Class Unknown Instances	26																												
LANES	83,6	<p>=== Summary ===</p> <table> <tr> <td>Correctly Classified Instances</td><td>77</td><td>83.6957 %</td></tr> <tr> <td>Incorrectly Classified Instances</td><td>15</td><td>16.3043 %</td></tr> <tr> <td>Kappa statistic</td><td>0.7037</td><td></td></tr> <tr> <td>Mean absolute error</td><td>0.0946</td><td></td></tr> <tr> <td>Root mean squared error</td><td>0.2204</td><td></td></tr> <tr> <td>Relative absolute error</td><td>37.4639 %</td><td></td></tr> <tr> <td>Root relative squared error</td><td>62.6744 %</td><td></td></tr> <tr> <td>Total Number of Instances</td><td>92</td><td></td></tr> <tr> <td>Ignored Class Unknown Instances</td><td>15</td><td></td></tr> </table>	Correctly Classified Instances	77	83.6957 %	Incorrectly Classified Instances	15	16.3043 %	Kappa statistic	0.7037		Mean absolute error	0.0946		Root mean squared error	0.2204		Relative absolute error	37.4639 %		Root relative squared error	62.6744 %		Total Number of Instances	92		Ignored Class Unknown Instances	15	
Correctly Classified Instances	77	83.6957 %																											
Incorrectly Classified Instances	15	16.3043 %																											
Kappa statistic	0.7037																												
Mean absolute error	0.0946																												
Root mean squared error	0.2204																												
Relative absolute error	37.4639 %																												
Root relative squared error	62.6744 %																												
Total Number of Instances	92																												
Ignored Class Unknown Instances	15																												

CLEAR-G	94,2	<p>=== Summary ===</p> <table> <tr> <td>Correctly Classified Instances</td><td>99</td><td>94.2857 %</td></tr> <tr> <td>Incorrectly Classified Instances</td><td>6</td><td>5.7143 %</td></tr> <tr> <td>Kappa statistic</td><td>0.8383</td><td></td></tr> <tr> <td>Mean absolute error</td><td>0.061</td><td></td></tr> <tr> <td>Root mean squared error</td><td>0.2256</td><td></td></tr> <tr> <td>Relative absolute error</td><td>16.273 %</td><td></td></tr> <tr> <td>Root relative squared error</td><td>52.2577 %</td><td></td></tr> <tr> <td>Total Number of Instances</td><td>105</td><td></td></tr> <tr> <td>Ignored Class Unknown Instances</td><td></td><td>2</td></tr> </table>	Correctly Classified Instances	99	94.2857 %	Incorrectly Classified Instances	6	5.7143 %	Kappa statistic	0.8383		Mean absolute error	0.061		Root mean squared error	0.2256		Relative absolute error	16.273 %		Root relative squared error	52.2577 %		Total Number of Instances	105		Ignored Class Unknown Instances		2
Correctly Classified Instances	99	94.2857 %																											
Incorrectly Classified Instances	6	5.7143 %																											
Kappa statistic	0.8383																												
Mean absolute error	0.061																												
Root mean squared error	0.2256																												
Relative absolute error	16.273 %																												
Root relative squared error	52.2577 %																												
Total Number of Instances	105																												
Ignored Class Unknown Instances		2																											
T-OR-D	96	<p>=== Summary ===</p> <table> <tr> <td>Correctly Classified Instances</td><td>98</td><td>96.0784 %</td></tr> <tr> <td>Incorrectly Classified Instances</td><td>4</td><td>3.9216 %</td></tr> <tr> <td>Kappa statistic</td><td>0.8437</td><td></td></tr> <tr> <td>Mean absolute error</td><td>0.0635</td><td></td></tr> <tr> <td>Root mean squared error</td><td>0.1489</td><td></td></tr> <tr> <td>Relative absolute error</td><td>24.8341 %</td><td></td></tr> <tr> <td>Root relative squared error</td><td>42.0284 %</td><td></td></tr> <tr> <td>Total Number of Instances</td><td>102</td><td></td></tr> <tr> <td>Ignored Class Unknown Instances</td><td></td><td>5</td></tr> </table>	Correctly Classified Instances	98	96.0784 %	Incorrectly Classified Instances	4	3.9216 %	Kappa statistic	0.8437		Mean absolute error	0.0635		Root mean squared error	0.1489		Relative absolute error	24.8341 %		Root relative squared error	42.0284 %		Total Number of Instances	102		Ignored Class Unknown Instances		5
Correctly Classified Instances	98	96.0784 %																											
Incorrectly Classified Instances	4	3.9216 %																											
Kappa statistic	0.8437																												
Mean absolute error	0.0635																												
Root mean squared error	0.1489																												
Relative absolute error	24.8341 %																												
Root relative squared error	42.0284 %																												
Total Number of Instances	102																												
Ignored Class Unknown Instances		5																											
MATERIAL	96,1	<p>=== Summary ===</p> <table> <tr> <td>Correctly Classified Instances</td><td>101</td><td>96.1905 %</td></tr> <tr> <td>Incorrectly Classified Instances</td><td>4</td><td>3.8095 %</td></tr> <tr> <td>Kappa statistic</td><td>0.9036</td><td></td></tr> <tr> <td>Mean absolute error</td><td>0.0485</td><td></td></tr> <tr> <td>Root mean squared error</td><td>0.1439</td><td></td></tr> <tr> <td>Relative absolute error</td><td>17.2744 %</td><td></td></tr> <tr> <td>Root relative squared error</td><td>38.7161 %</td><td></td></tr> <tr> <td>Total Number of Instances</td><td>105</td><td></td></tr> <tr> <td>Ignored Class Unknown Instances</td><td></td><td>2</td></tr> </table>	Correctly Classified Instances	101	96.1905 %	Incorrectly Classified Instances	4	3.8095 %	Kappa statistic	0.9036		Mean absolute error	0.0485		Root mean squared error	0.1439		Relative absolute error	17.2744 %		Root relative squared error	38.7161 %		Total Number of Instances	105		Ignored Class Unknown Instances		2
Correctly Classified Instances	101	96.1905 %																											
Incorrectly Classified Instances	4	3.8095 %																											
Kappa statistic	0.9036																												
Mean absolute error	0.0485																												
Root mean squared error	0.1439																												
Relative absolute error	17.2744 %																												
Root relative squared error	38.7161 %																												
Total Number of Instances	105																												
Ignored Class Unknown Instances		2																											
SPAN	92,3	<p>=== Summary ===</p> <table> <tr> <td>Correctly Classified Instances</td><td>85</td><td>92.3913 %</td></tr> <tr> <td>Incorrectly Classified Instances</td><td>7</td><td>7.6087 %</td></tr> <tr> <td>Kappa statistic</td><td>0.869</td><td></td></tr> <tr> <td>Mean absolute error</td><td>0.1031</td><td></td></tr> <tr> <td>Root mean squared error</td><td>0.2411</td><td></td></tr> <tr> <td>Relative absolute error</td><td>27.8134 %</td><td></td></tr> <tr> <td>Root relative squared error</td><td>56.1841 %</td><td></td></tr> <tr> <td>Total Number of Instances</td><td>92</td><td></td></tr> <tr> <td>Ignored Class Unknown Instances</td><td></td><td>15</td></tr> </table>	Correctly Classified Instances	85	92.3913 %	Incorrectly Classified Instances	7	7.6087 %	Kappa statistic	0.869		Mean absolute error	0.1031		Root mean squared error	0.2411		Relative absolute error	27.8134 %		Root relative squared error	56.1841 %		Total Number of Instances	92		Ignored Class Unknown Instances		15
Correctly Classified Instances	85	92.3913 %																											
Incorrectly Classified Instances	7	7.6087 %																											
Kappa statistic	0.869																												
Mean absolute error	0.1031																												
Root mean squared error	0.2411																												
Relative absolute error	27.8134 %																												
Root relative squared error	56.1841 %																												
Total Number of Instances	92																												
Ignored Class Unknown Instances		15																											
REL-L	88,2	<p>=== Summary ===</p> <table> <tr> <td>Correctly Classified Instances</td><td>90</td><td>88.2353 %</td></tr> <tr> <td>Incorrectly Classified Instances</td><td>12</td><td>11.7647 %</td></tr> <tr> <td>Kappa statistic</td><td>0.7842</td><td></td></tr> <tr> <td>Mean absolute error</td><td>0.1099</td><td></td></tr> <tr> <td>Root mean squared error</td><td>0.2475</td><td></td></tr> <tr> <td>Relative absolute error</td><td>28.3183 %</td><td></td></tr> <tr> <td>Root relative squared error</td><td>56.3014 %</td><td></td></tr> <tr> <td>Total Number of Instances</td><td>102</td><td></td></tr> <tr> <td>Ignored Class Unknown Instances</td><td></td><td>5</td></tr> </table>	Correctly Classified Instances	90	88.2353 %	Incorrectly Classified Instances	12	11.7647 %	Kappa statistic	0.7842		Mean absolute error	0.1099		Root mean squared error	0.2475		Relative absolute error	28.3183 %		Root relative squared error	56.3014 %		Total Number of Instances	102		Ignored Class Unknown Instances		5
Correctly Classified Instances	90	88.2353 %																											
Incorrectly Classified Instances	12	11.7647 %																											
Kappa statistic	0.7842																												
Mean absolute error	0.1099																												
Root mean squared error	0.2475																												
Relative absolute error	28.3183 %																												
Root relative squared error	56.3014 %																												
Total Number of Instances	102																												
Ignored Class Unknown Instances		5																											
TYPE	87,6	<p>=== Summary ===</p> <table> <tr> <td>Correctly Classified Instances</td><td>92</td><td>87.619 %</td></tr> <tr> <td>Incorrectly Classified Instances</td><td>13</td><td>12.381 %</td></tr> <tr> <td>Kappa statistic</td><td>0.8362</td><td></td></tr> <tr> <td>Mean absolute error</td><td>0.0661</td><td></td></tr> <tr> <td>Root mean squared error</td><td>0.1758</td><td></td></tr> <tr> <td>Relative absolute error</td><td>26.1066 %</td><td></td></tr> <tr> <td>Root relative squared error</td><td>49.5472 %</td><td></td></tr> <tr> <td>Total Number of Instances</td><td>105</td><td></td></tr> <tr> <td>Ignored Class Unknown Instances</td><td></td><td>2</td></tr> </table>	Correctly Classified Instances	92	87.619 %	Incorrectly Classified Instances	13	12.381 %	Kappa statistic	0.8362		Mean absolute error	0.0661		Root mean squared error	0.1758		Relative absolute error	26.1066 %		Root relative squared error	49.5472 %		Total Number of Instances	105		Ignored Class Unknown Instances		2
Correctly Classified Instances	92	87.619 %																											
Incorrectly Classified Instances	13	12.381 %																											
Kappa statistic	0.8362																												
Mean absolute error	0.0661																												
Root mean squared error	0.1758																												
Relative absolute error	26.1066 %																												
Root relative squared error	49.5472 %																												
Total Number of Instances	105																												
Ignored Class Unknown Instances		2																											

Таблиця 2 - результати NaiveBayers за атрибутами

Маємо знов схожий момент, коли найточніша модель IDENTIF занадто розбиває на класи, тому обираємо наступне за точністю MATERIAL з класами WOOD, IRON, STEAL.

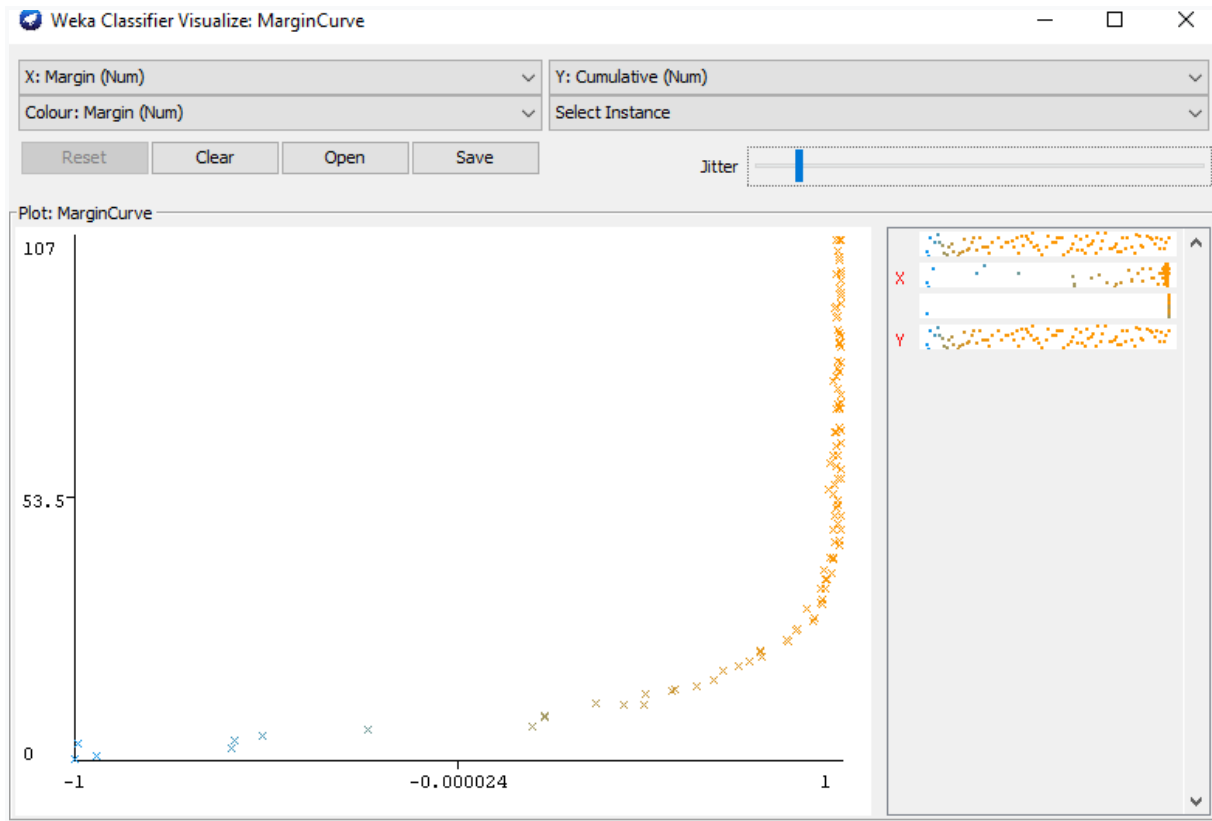


Рисунок 9 - візуалізація прогнозування від отриманої моделі

Висновки

Закріплено знання з класифікації і її використання у WEKA.

Були виконані відповідні завдання, найточнішими моделями були за атрибутом IDENTIF, але при врахуванні доцільності таких моделей, були взяті другі за точністю, при J48 за атрибутом CLEAR-G з класами N і G, NaiveBayes за атрибутом MATERIAL з класами WOOD, IRON, STEAL.

У порівнянні J48 і NaiveBayes ці два алгоритми мають різні характеристики. У результаті два алгоритми демонструють добру точність ($> 70\%$). Однак NaiveBayes має краще значення точності (на основі F-вимірювання, повторного виклику та точності) і правильно знаходить значення Classified Instances порівняно з іншими алгоритмами

Класифікація — це проблема ідентифікації, до якої з набору категорій належить нове спостереження, на основі навчального набору даних, що містить спостереження та приналежність до категорій які відомі. Інтелектуальний аналіз даних необхідний для виконання прогностичного аналізу наборів даних. Знання, які можна отримати за допомогою методів аналізу даних, є прогнозом включання в певну категорію, який потім буде використовуватися для прийняття рішень.

СПИСОК ДЖЕРЕЛ ІНФОРМАЦІЇ

- 1 Data Mining: Practical Machine Learning Tools and Technique \ Ian Witten, Eibe Frank, Mark Hall
- 2 WEKA. Руководство по использованию \ Хабр online ресурс
<https://habr.com/ru/post/590565/>
- 3 Інтелектуальний аналіз даних: Навчальний посібник \ А. О. Олійник, С. О. Субботін, О. О. Олійник