ГУАП

КАФЕДРА № 42

ОТЧЕТ
ЗАЩИЩЕН С ОЦЕНКОЙ _____

ПРЕПОДАВАТЕЛЬ

профессор, д-р.т.н., профессор				В. В. Фомин
должность, уч. степень, звание		подпись, дата		инициалы, фамилия

ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ №5

МЕТОД НАИВНОГО БАЙЕСОВСКОГО ПОДХОДА

Вариант 5

по курсу: МЕТОДЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

РАБОТУ ВЫПОЛНИЛ

СТУДЕНТ ГР. №	4128				Воробьев В. А.
			подпись, дата		инициалы, фамилия

Санкт-Петербург 2024

СОДЕРЖАНИЕ

1 Введение

1.1 Цель лабораторной работы

Изучение основ организации работы с технологической платформой для создания законченных аналитических решений использованием наивного Байесовского подхода.

1.2 Задание

1. Для набора данных выполнить классификацию с помощью Байесовского подхода.
2. Выполнить оценку качества классификации.

2 Выполнение работы

2.1 Набор данных

Набор данных взят с Kaggle (URI - `https://www.kaggle.com/datasets/sudhanshu2198/wheat-variety-classification`).

Набор данных включает зерна пшеницы, принадлежащие к трем различным сортам пшеницы: **Кама, Роза и Канадская**, по 70 элементов каждый.

Для построения данных были измерены семь геометрических параметров зерен пшеницы:

1) Область — размер поверхности зерна пшеницы.
2) Периметр — общая длина внешней границы зерна.
3) Компактность — насколько форма зерна близка к идеальной круговой.
4) Длина ядра — измерение самой длинной оси внутренней части зерна пшеницы.
5) Ширина ядра — поперечное измерение внутренней части зерна.
6) Коэффициент асимметрии — отклонение формы зерна от симметричной.
7) Длина бороздки ядра — протяженность центральной линии или углубления в зерне.

Для каждого этого параметра был сопоставлен сорт пшеницы:

- **Кама** — сорт пшеницы, известный своей устойчивостью к болезням и приспособленностью к различным климатическим условиям.
- **Роза** — сорт пшеницы, который ценится за качество зерна и применяется для муки высшего сорта.
- **Канадская** — сорт пшеницы с высоким содержанием белка, используемый для производства высококачественной муки.

2.2 Рабочий процесс

Целью создания данной системы является проверка гипотезы, что вышеуказанных 7 параметров достаточно для определения сорта пшеницы. Гипотезу будем считать доказанной, если точность составит `95%`.

Для создания модели в программе KNIME создаём следующие узлы:

- `Excel Reader` для считывания файла;

- `Number to String` для преобразования номера сорта пшеницы в строку.
- `String Manipulation` для сопоставления номера сорта с его названием.
- `Color Manager` для цветового разделения на графике;
- `Partitioning` для разделения данных на обучающие и тестовые(50/50). Дополнительно выбран `Linear Sampling`, так как набор данных отсортирован по сорту пшеницы;
- `Naive Bayes Learner` для обучения модели;
- `Naive Bayes Predictor` непосредственно для предсказания;
- `Scorer` для вычисления статистики;

На рисунке 2.1 представлена схема рабочего процесса.

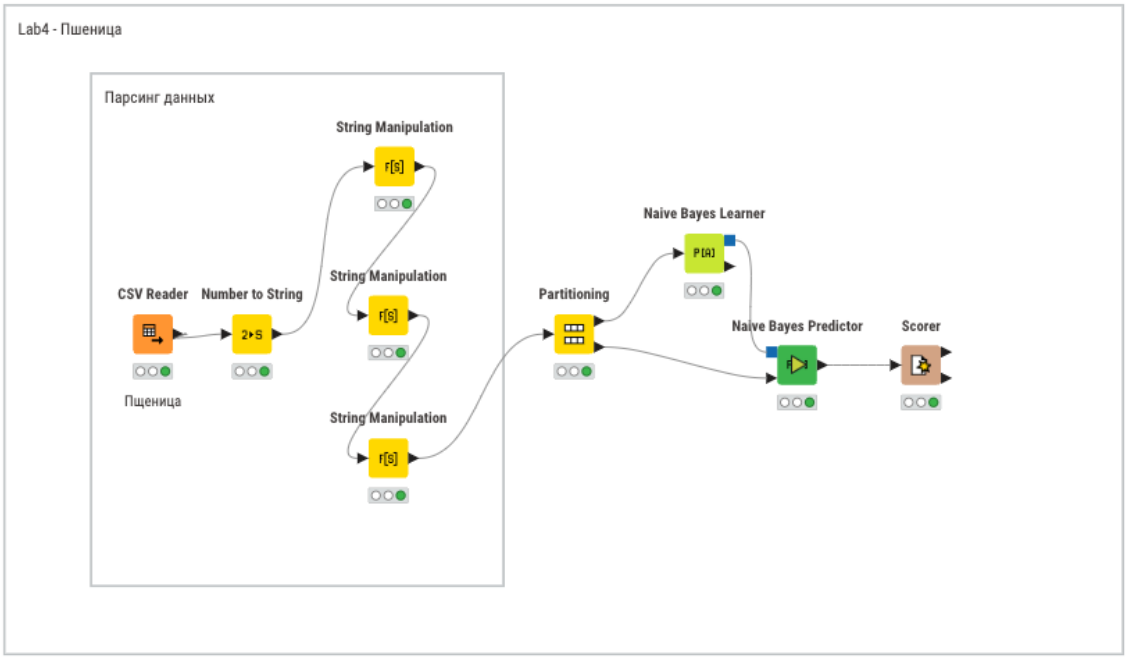

Рисунок 2.1 - Схема в KNIME

В результате из 98 тестовых записей 91 предсказаны верно, а 7 нет. Точность попадания равняется `93.333%`. На рисунке 2 представлена матрица сопряженности. На рисунке 3 – метрики оценки качества.

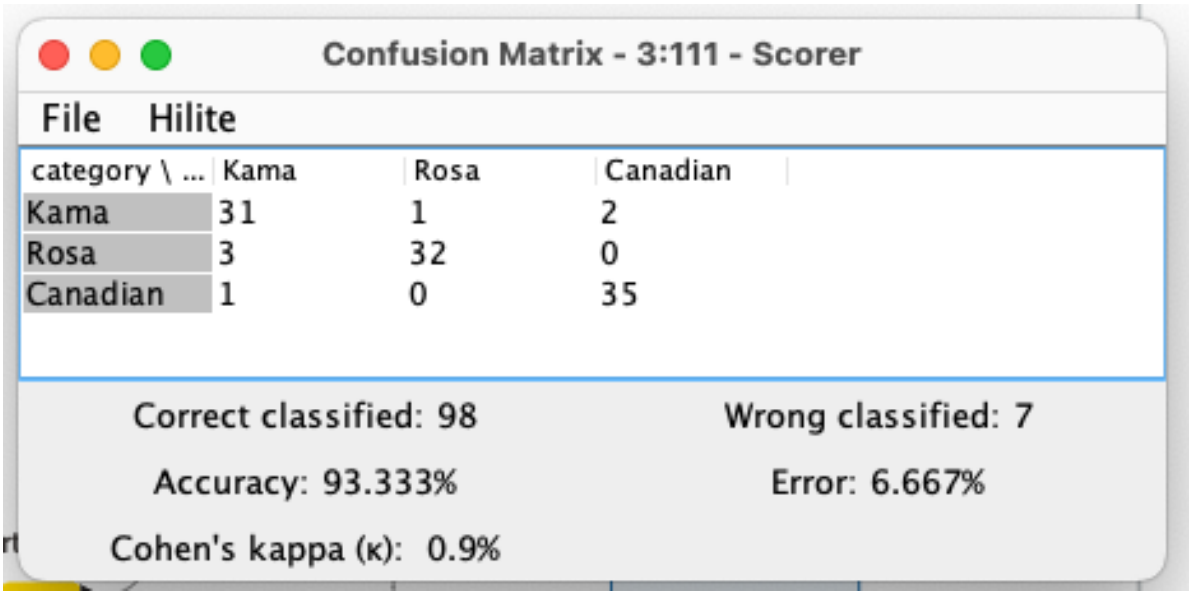

Рисунок 2.2 - Матрица смежности

#	RowID	TruePositives	FalsePositives	TrueNegatives	FalseNegatives	Recall	Precision	Sensitivity	Specificity	F-measure	Accuracy	Cohen's kappa
1	Kama	31	4	67	3	0.912	0.886	0.912	0.944	0.899		
2	Rosa	32	1	69	3	0.914	0.97	0.914	0.986	0.941		
3	Cana...	35	2	67	1	0.972	0.946	0.972	0.971	0.959		
4	Overall										0.933	0.9

Рисунок 2.3 - Метрики оценки качества

Из метрик оценки качества следует, что как и для метода дерево решений, лучше всего определяется сорт канадский. У этого сорта самая высокая точность и полнота. Сорт Кама определяется хуже всего, если полнота примерно равна сорту Роза, то точность - ниже. Сорт Кама чаще всего путается с сортом Роза.

На рисунке 4 представлена таблица результатов обучения модели, демонстрирующая средние значения атрибутов по классам и их стандартное отклонение. Можно сделать вывод, что сорт Кама чаще всего путается с сортом Роза по таким параметрам: ширина и компактность.

Rows: 24 | Columns: 7

Table | Statistics

#	RowID	Attribute (String)	Value (String)	Class (String)	Count (Number (integer))	Missing value count (Number (integer))	Mean (Number (double))	Standard deviation (Number (double))
1	Row0	area		Canadian	34	0	11.848	0.764
2	Row1	area		Kama	36	0	14.28	1.324
3	Row2	area		Rosa	35	0	18.313	1.308
4	Row3	asymmetry coefficient		Canadian	34	0	4.736	1.434
5	Row4	asymmetry coefficient		Kama	36	0	2.708	1.259
6	Row5	asymmetry coefficient		Rosa	35	0	3.507	1.233
7	Row6	category		Canadian	34	0		
8	Row7	category		Kama	36	0		
9	Row8	category		Rosa	35	0		
10	Row9	compactness		Canadian	34	0	0.851	0.025
11	Row10	compactness		Kama	36	0	0.88	0.016
12	Row11	compactness		Rosa	35	0	0.884	0.015
13	Row12	groove length		Canadian	34	0	5.094	0.163
14	Row13	groove length		Kama	36	0	5.081	0.245
15	Row14	groove length		Rosa	35	0	6.014	0.232
16	Row15	length		Canadian	34	0	5.216	0.141
17	Row16	length		Kama	36	0	5.492	0.228
18	Row17	length		Rosa	35	0	6.127	0.243
19	Row18	perimeter		Canadian	34	0	13.217	0.336
20	Row19	perimeter		Kama	36	0	14.266	0.603
21	Row20	perimeter		Rosa	35	0	16.122	0.565
22	Row21	width		Canadian	34	0	2.859	0.164
23	Row22	width		Kama	36	0	3.241	0.189
24	Row23	width		Rosa	35	0	3.682	0.172

Рисунок 2.4 - Таблица результатов обучения

Naive Bayes Learner View - 3:110 - Naive Bayes Learner

File

Class counts for category

Class:	Canadian	Kama	Rosa
Count:	34	36	35

Total count: 105

Threshold to used for zero probabilities: 1.0E-4

Gaussian distribution for area per class value

	Canadian	Kama	Rosa
Count:	34	36	35
Mean:	11.84824	14.28028	18.31314
Std. Deviation:	0.76431	1.32443	1.30837
Rate:	32%	34%	33%

Gaussian distribution for asymmetry coefficient per class value

	Canadian	Kama	Rosa
Count:	34	36	35
Mean:	4.73647	2.70803	3.50714
Std. Deviation:	1.43375	1.25893	1.2329
Rate:	32%	34%	33%

Gaussian distribution for compactness per class value

	Canadian	Kama	Rosa
Count:	34	36	35
Mean:	0.8514	0.87974	0.88431
Std. Deviation:	0.02516	0.01606	0.01525
Rate:	32%	34%	33%

Gaussian distribution for groove length per class value

	Canadian	Kama	Rosa
Count:	34	36	35
Mean:	5.09365	5.08053	6.01397
Std. Deviation:	0.16297	0.24528	0.23169
Rate:	32%	34%	33%

Gaussian distribution for length per class value

	Canadian	Kama	Rosa
Count:	34	36	35
Mean:	5.21609	5.49242	6.12749
Std. Deviation:	0.14053	0.22847	0.24312
Rate:	32%	34%	33%

Gaussian distribution for perimeter per class value

	Canadian	Kama	Rosa
Count:	34	36	35
Mean:	13.21706	14.26611	16.12229
Std. Deviation:	0.33632	0.60257	0.56491
Rate:	32%	34%	33%

Gaussian distribution for width per class value

	Canadian	Kama	Rosa
Count:	34	36	35
Mean:	2.85874	3.24083	3.68183
Std. Deviation:	0.16447	0.18874	0.17193
Rate:	32%	34%	33%

Рисунок 2.5 - Модель наивного Байеса

3 Вывод

Полученная точность 93.333% при наивном байесовском подходе, что больше 90.476% при методе K ближайших соседей, но всё равно остается недостаточным для подтверждения гипотезы. К тому же точность предсказания сорта Кама равняется ~ 85%, что делает модель худшей для определения этого сорта.