

КЛАСТЕРИЗАЦИЯ И КЛАССИФИКАЦИЯ КРАСНЫХ ВИН ПО ФИЗИКО-ХИМИЧЕСКИМ СВОЙСТВАМ МЕТОДАМИ DATA MINING

© 2023 г. Н. В. Бондарев^{1,*}

¹ Харьковский национальный университет имени В. Н. Каразина, пл. Свободы 4, Харьков, 61022 Украина

*e-mail: n_bondarev@ukr.net

Поступило в редакцию 20 июля 2023 г.

После доработки 20 июля 2023 г.

Принято к печати 1 августа 2023 г.

Исследовано 178 образцов красных вин итальянских производителей, взятых из общедоступного репозитория машинного обучения UCI. Методами Data Mining выполнен компьютерный анализ влияния 13 физико-химических свойств образцов на распределение вин по трем группам. Построены классификационные модели: факторная, дискриминантная, каноническая, нейросетевые (многослойный перцептрон MLP, нейронная сеть Кохонена SOFM), прогнозные (метод опорных векторов, байесовский классификатор, метод ближайшего соседа) и деревья решений. Обучены нейросетевые кластеризатор SOFM 13-3 и классификаторы MLP 13-5-3 и SOFM 16-3. Выявлено, что пролин, флавоноиды, интенсивность цвета, белки и алкоголь определяют дискриминирующую мощность моделей.

Ключевые слова: красные вина, кластеризация, классификация, Data Mining, STATISTICA

DOI: 10.31857/S0044460X23090147, **EDN:** WUHWPB

Вино является одним из наиболее часто употребляемых в пищу алкогольных напитков со сложным многокомпонентным химическим составом. Поэтому большое социально-экономическое значение виноделия определяет актуальность подготовки специалистов, владеющих современными статистическими и кибернетическими методами компьютерного анализа физико-химических свойств вин [1–5]. Научно-обоснованные заключения таких специалистов, подтвержденные органолептической оценкой энологов и сомелье, будут способствовать выявлению фальсификации и преднамеренного искажения информации о качестве винодельческой продукции [6].

В данной статье приводятся результаты компьютерного анализа свойств образцов итальянских красных вин методами Data Mining для выявления степени их влияния на разделяющую способность классификационных моделей. С помощью классификации выявлялись сходные (близкие) свойства

вина, характеризующие группу, к которой принадлежит образец вина. Кластеризация отличается от классификации тем, что число групп заранее не задано.

Экспериментальные данные о физико-химических свойствах вин взяты из общедоступного репозитория машинного обучения UCI [7]. Файл данных включает 178 образцов трех красных итальянских вин: Бароло (Barolo – 59 образцов, Гриньолино (Grignolino – 71 образец), Барбера (Barbera – 48 образцов), приготовленных из трех разных сортов винограда: Неббиоло (Nebbiolo), Гриньолино и Барбера, собранного в 1970–1979 годах в регионе Пьемонт. Для каждого образца вина приведены значения 13 физико-химических свойств.

Авторы работы [8] выделили 8 переменных – спирт, сумма фенолов, флавоноиды, интенсивность окраски, оттенок, OD280/OD315 разбав-

Таблица 1. Описательная статистика физико-химических свойств 178 образцов итальянских вин

Свойство (v)	Среднее	Минимальное значение	Максимальное значение	Стандартное отклонение	Стандартная ошибка
Алкоголь, об% (v1)	13.0	11.0	14.8	0.8	0.06
Яблочная кислота, г/л (v2)	2.3	0.7	5.8	1.1	0.08
Зола, г/л (v3)	2.4	1.4	3.2	0.3	0.02
Щелочность золы, мг-экв NaOH/л (v4)	19.5	10.6	30.0	3.3	0.25
Магний, мг/л (v5)	99.7	70.0	162.0	14.3	1.07
Всего фенолов, г/л (v6)	2.3	1.0	3.9	0.6	0.05
Флавоноиды, г/л (v7)	2.0	0.3	5.1	1.0	0.07
Фенолы нефлавоноидные, г/л (v8)	0.4	0.1	0.7	0.1	0.01
Проантоцианидины, г/л (v9)	1.6	0.4	3.6	0.6	0.04
Интенсивность цвета, баллы (v10)	5.1	1.3	13.0	2.3	0.17
Оттенок, баллы (v11)	1.0	0.5	1.7	0.2	0.02
OD280/OD315 разбавленных вин (v12)	2.6	1.3	4.0	0.7	0.05
Пролин, мг/л (v13)	746.9	278.0	1680.0	314.9	23.60

ленных вин, OD280/OD315 флавоноидов, пролин из 28 химических и физико-химических свойств 178 образцов вин. К 8-мерному набору данных были применены методы классификации [9]: ближайшего соседа KNN, линейный дискриминантный анализ LDA, анализ главных компонент PCA и методы моделирования [9]: байесовский анализ (Bayesian analysis), формальное независимое моделирование аналогий классов SIMCA (Soft Independent Modeling of Class Analogy). Классификационная способность моделей составила 97.75% (четыре образца вина классифицированы ошибочно).

В репозитории [7] приведена дополнительная информация об использовании 13-мерного набора свойств вин (алкоголь, яблочная кислота, зола, щелочность золы, магний, всего фенолов, флавоноиды, фенолы нефлавоноидные, проантоцианидины, интенсивность цвета, оттенок, OD280/OD315 разбавленных вин, пролин) для классификации вин. Для сравнения приведена точность разделения 178 образцов вин на три класса четырьмя различными классификаторами: анализ избыточности RDA (Redundancy analysis) – 100%, квадратичный дис-

криминантный анализ QDA (Quadratic discriminant analysis) – 99.4% (один образец классифицирован неверно), линейный дискриминантный анализ LDA (Linear discriminant analysis) – 98.9%, (две ошибочные классификации образцов) метод ближайшего соседа 1NN (Nearest neighbors algorithm) – 96.1% (класс семи образцов предсказан ошибочно).

Целью данной работы является построение математических моделей кластеризации и классификации образцов итальянских вин по их физико-химическим свойствам современными методами Data Mining. Для достижения цели использованы статистические методы: дескриптивный анализ (DDA) и описание исходных данных; анализ связей (корреляционный анализ – CA, факторный анализ – FA), многомерный статистический анализ (иерархическая кластеризация – HCA hierarchical cluster analysis, алгоритм *k*-средних (k-means clustering), дискриминантный анализ – LDA, канонический корреляционный анализ – CCA, метод опорных векторов – SVM, метод ближайшего соседа – 1NN) и кибернетические методы: деревья решений (Decision Trees Algorithm – DTA), наи-

Таблица 2. Расчетные и табличные значения критериев нормальности распределения свойств вин^a

Переменная ($n = 178$)	Критерий Колмогорова–Смирнова ($D_{\text{табл}} = 0.102D_{\text{расч}}$)	Критерий Хи-квадрат, $\chi^2_{\text{расч}}$ ($\chi^2_{\text{табл}}$)
v1	0.068	13.521 (11.070)
v2	0.191	83.920 (12.592)
v3	0.058	10.064 (7.815)
v4	0.063	4.003 (9.488)
v5	0.090	14.384 (7.815)
v6	0.072	32.976 (18.307)
v7	0.085	35.900 (11.070)
v8	0.115	24.846 (5.991)
v9	0.057	9.518 (16.919)
v10	0.091	40.089 (14.067)
v11	0.062	10.245 (12.592)
v12	0.117	43.526 (19.675)
v13	0.129	48.429 (18.307)

^a n – объем выборки. Если табличное значение $D_{\text{табл}}$ больше расчетного значения $D_{\text{расч}}$, а $\chi^2_{\text{расч}} < \chi^2_{\text{табл}}$, то распределение считается соответствующим нормальному на уровне значимости $p = 0.05$.

вный баесовский классификатор (naive Bayesian classifier – NBA), искусственные нейронные сети (многослойные перцептроны – MLP, самоорганизующаяся карта признаков Кóхонена (Self-Organizing Feature Map – SOFM). Компьютерный анализ физико-химических свойств вин проведен на статистических платформах STATISTICA и SPSS в среде Windows 10.

Первичный анализ данных. В табл. 1 приведены количественные параметры описательной статистики физико-химических свойств образцов вин. Распределение можно считать симметричным, если среднее квадратическое отклонение (стандартное отклонение) данных меньше половины среднего арифметического. Проверка гипотезы нормального распределения анализируемых данных (табл. 2) выполнена по критериям Колмогорова–Смирнова и Хи квадрат ($n > 50$).

Описание свойств вин [10, 11]. По содержанию сахара и этилового спирта вина делятся на сухие (8.5–15%), полусухие (8.5–15%, сахар 4–18 г/л), полусладкие (сахар 18–45 г/л), сладкие (сахар не менее 45 г/л). Роль этилового спирта заключается в значительном усилении сладости сахаров. Кроме того, этиловый спирт может увеличивать воспринимаемую интенсивность фенольных соединений, придающих вину горький вкус, подавляя ощущение терпкости, обусловленной танина-

ми [12]. Яблочная кислота (v2) влияет на формирование вкуса вина. Повышенное содержание этой кислоты в винах вызывает резкое вкусовое ощущение, так называемую зеленую кислотность [13]. Содержание золы (v3) в винах Италии составляет 1.1–4.8 г/л. Зола и зольные элементы играют определенную роль в формировании качества и диетических свойств вин. Контролируемые показатели щелочности золы (v4), как и золы, используются для оценки подлинности и качества вин.

Магний (v5) имеет большое значение для сохранения здоровья людей: помогает росту клеток; регулирует сердечный ритм; регулирует содержание сахара в крови; является профилактическим средством против мигрени; улучшает функцию мозга; укрепляет зубную эмаль. Массовая концентрация магния в вине составляет 0.03–0.24 г/л. Фенольные вещества вина (v6) – это группа соединений, в состав которой входят красящие вещества (антоцианы) и дубильные (танины и таниды), обуславливающие полноту вкуса и оказывающие влияние на аромат, цвет, а также прозрачность вина. Флавоноиды (v7) – природные красители, пищевые антиоксиданты, дубильные вещества. Флавоноидные фенольные соединения оказывают существенное влияние на вкус и ощущения во рту как вместе, так и по отдельности. Они составляют значительную группу полимеров (танинов). В вине они вызывают ощущение горечи и терпкости, влия-

Таблица 3. Корреляционная матрица переменных^a

Свойство	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13
v1	1.00	0.09	0.21	–0.31	0.27	0.29	0.24	–0.16	0.14	0.55	–0.07	0.07	0.64
v2	0.09	1.00	0.16	0.29	–0.05	–0.34	–0.41	0.29	–0.22	0.25	–0.56	–0.37	–0.19
v3	0.21	0.16	1.00	0.44	0.29	0.13	0.12	0.19	0.01	0.26	–0.07	0.00	0.22
v4	–0.31	0.29	0.44	1.00	–0.08	–0.32	–0.35	0.36	–0.20	0.02	–0.27	–0.28	–0.44
v5	0.27	–0.05	0.29	–0.08	1.00	0.21	0.20	–0.26	0.24	0.20	0.06	0.07	0.39
v6	0.29	–0.34	0.13	–0.32	0.21	1.00	0.86	–0.45	0.61	–0.06	0.43	0.70	0.50
v7	0.24	–0.41	0.12	–0.35	0.20	0.86	1.00	–0.54	0.65	–0.17	0.54	0.79	0.49
v8	–0.16	0.29	0.19	0.36	–0.26	–0.45	–0.54	1.00	–0.37	0.14	–0.26	–0.50	–0.31
v9	0.14	–0.22	0.01	–0.20	0.24	0.61	0.65	–0.37	1.00	–0.03	0.30	0.52	0.33
v10	0.55	0.25	0.26	0.02	0.20	–0.06	–0.17	0.14	–0.03	1.00	–0.52	–0.43	0.32
v11	–0.07	–0.56	–0.07	–0.27	0.06	0.43	0.54	–0.26	0.30	–0.52	1.00	0.57	0.24
v12	0.07	–0.37	0.00	–0.28	0.07	0.70	0.79	–0.50	0.52	–0.43	0.57	1.00	0.31
v13	0.64	–0.19	0.22	–0.44	0.39	0.50	0.49	–0.31	0.33	0.32	0.24	0.31	1.00

^a Полу жирным шрифтом выделены коэффициенты парных корреляций, описывающие умеренную, заметную и высокую взаимосвязь физико-химических свойств вин.

ют на цвет, тело и букет. Флавоноидные танины – основные фенольные соединения красных вин. Фенолы нефлавоноидные (v8) являются основным фактором, влияющим на цвет белого вина. Проантоцианидины (v9) – сильнейшие антиоксиданты, которые защищают организм от окислительного стресса и опасных заболеваний.

Цвет вина (v10) говорит о его плотности, которая определяется концентрацией сахара и спирта. Чем интенсивнее цвет, тем больше в нем танинов. В обзоре [14] обобщены существующие методические подходы к оценке цвета и оттенков (v11) для различных типов вин. OD280/OD315 (OD – optical density) разбавленных вин (v12) – УФ спектроскопия для количественного определения белка и фенольных соединений [15]. Большинство фенольных соединений имеют интенсивные полосы поглощения в областях 280 и 320 нм. Белки поглощают свет с различной длиной волны за счет аминокислотных остатков с ароматическими кольцами. Пролин (v13), аргинин и треонин – основные гетероциклические аминокислоты вин, которые принимают непосредственное участие в формировании полноты вкуса, сложении аромата и биологической ценности вина.

В основу компьютерного анализа данных положены математические приемы, статистические методы и практические рекомендации, приведенные

в работах [16–19]. В работах [20–22] показаны возможности методов Data Mining для классификации и прогнозирования степени связывания (константа комплексообразования) катионов металлов краун-эфирами [20], криптандом ([222]) [21] в разных чистых растворителях, а также для прогнозирования энергии адсорбции фуллеренов лекарствами [22].

Факторный анализ. Основное назначение факторного анализа – анализ корреляций множества переменных (физико-химических свойств образцов вин) табл. 3. Для оценки силы связи в теории корреляции применяется шкала английского статистика Чеддока: слабая – от 0.1 до 0.3; умеренная – от 0.3 до 0.5; заметная – от 0.5 до 0.7; высокая – от 0.7 до 0.9; весьма высокая (сильная) – от 0.9 до 1.0.

Надежность вычислений элементов корреляционной матрицы табл. 3 и целесообразность ее описания с помощью факторного анализа подтверждены мерой адекватности выборки Кайзера–Мейера–Олкина (КМО = 0.779) и критерием сферичности Бартлетта (примерная Хи-квадрат = 1317.181 для 78 степеней свободы и уровня значимости $p = 0.000$), что свидетельствует о наличии достаточной коррелированности переменных [19]. Матрица интеркорреляций исходных данных (табл. 3) обработана с использованием анализа главных компонент. Основной принцип итераци-

Таблица 4. Объясненная совокупная дисперсия (изменчивость)

Фактор	Начальные собственные значения факторов до вращения			Извлеченные собственные значения главных факторов до вращения			Собственные значения главных факторов после варимакс вращения		
	собственное значение, λ	% объясненной дисперсии (изменчивости)	% совокупного собственного значения	собственное значение, λ	% объясненной дисперсии (изменчивости)	% совокупного собственного значения	собственное значение, λ	% объясненной дисперсии (изменчивости)	% совокупного собственного значения
1	4.71	36.20	36.20	4.71	36.20	36.20	3.56	27.37	27.37
2	2.50	19.21	55.41	2.50	19.21	55.41	2.22	17.10	44.47
3	1.45	11.12	66.53	1.45	11.12	66.53	1.59	12.25	56.72
4	0.92	7.07	73.60	0.92	7.07	73.60	1.84	14.18	70.90
5	0.85	6.56	80.16	0.85	6.56	80.16	1.20	9.26	80.16
6	0.64	4.94	85.10						
7	0.55	4.24	89.34						
8	0.35	2.68	92.02						
9	0.29	2.22	94.24						
10	0.25	1.93	96.17						
11	0.23	1.74	97.91						
12	0.17	1.30	99.20						
13	0.10	0.80	100.00						

онного выделения латентных факторов методом главных компонент – представление двух или более свойств вина одним фактором. В основе анализа главных компонент лежит математический метод нахождения собственных значений и собственных векторов корреляционной матрицы. Собственные значения λ – дисперсии (изменчивости), выделяемые факторами.

Итерационная процедура начинается с матрицы, в которой дисперсия (изменчивость) каждой переменной равна 1. Поэтому общая дисперсия равна числу переменных – 13, т. е. наибольшей изменчивости, которая может быть выделена. Изменчивость, объясненная последовательно выделяемыми факторами до и после вращения, представлена в табл. 4. Наличие информации о том, сколько дисперсии (изменчивости) выделил каждый фактор (табл. 4) позволило оставить 5 факторов (пять главных компонент) с собственными значениями $\lambda \geq 0.85$ на основе графического метода Р. Кэттелла [23] – критерия каменистой осыпи.

Качество факторного анализа тем выше, чем выше полнота факторизации. Построенная факторная модель сохраняет 80.16% (табл. 4) исходной информации, при этом число факторов сокращается с 13 до 5. Фактор 1 сохраняет 27.37% исходной информации о винах, Фактор 2 – 17.10%, Фактор 3 – 12.25%, Фактор 4 – 14.18% и Фактор 5 – 9.26% (табл. 4 и 5).

Результат варимакс-вращения главных факторов представлен в табл. 5. Строки таблицы i содержат факторные нагрузки a_{ik} переменных (13 физико-химических свойств образцов вин) по пяти факторам (столбцам k). Факторные нагрузки в этом случае, являющиеся аналогом коэффициента корреляции, изменяются от -1 до $+1$ и показывают степень взаимосвязи соответствующих переменных и факторов – чем больше абсолютная величина факторной нагрузки, тем сильнее связь переменной с фактором, тем больше данная переменная обусловлена действием соответствующего фактора [16, 17].

Таблица 5. Факторные нагрузки после варимакс-вращения^a

Переменная (свойство вина)	Факторные нагрузки (a_{ik}). Варимакс-вращение. Извлечение факторов методом главных компонент				
	Фактор 1	Фактор 2	Фактор 3	Фактор 4	Фактор 5
v1	0.150	0.873	0.004	–0.116	0.080
v2	–0.183	0.009	0.187	–0.777	–0.121
v3	0.086	0.224	0.888	–0.025	0.171
v4	–0.197	–0.434	0.719	–0.273	–0.017
v5	0.083	0.205	0.142	0.042	0.915
v6	0.849	0.257	0.027	0.196	0.053
v7	0.885	0.178	–0.011	0.292	0.078
v8	–0.556	0.012	0.452	–0.034	–0.448
v9	0.789	0.044	–0.046	–0.070	0.173
v10	–0.217	0.693	0.137	–0.459	0.129
v11	0.394	–0.091	–0.038	0.805	–0.007
v12	0.834	–0.066	–0.058	0.330	–0.026
v13	0.340	0.766	–0.002	0.227	0.244
Собственное значение, λ	3.559	2.223	1.592	1.843	1.204
Доля дисперсии	0.2737	0.1710	0.1225	0.1418	0.0926

^a Полужирным шрифтом выделены нагрузки физико-химических свойств вин на факторы, > 0.693 - высокие.

Поскольку по Фактору 1 максимальные нагрузки имеют переменные всего фенолов (v6), флавоноиды (v7), проантоцианидины (v9), OD280/OD315 разбавленных вин (v12), то Фактору 1 (новой переменной) может быть присвоено название «белки, антиоксиданты». Фактору 2 можно присвоить название «органолептический» [алкоголь (v1), интенсивность цвета (v10), пролин (v13)]. Фактор 3 можно назвать «кислотность вина» [зола (v3), щелочность золы (v4)]. Фактор 4 – «оттенок вина» [яблочная кислота (v2), оттенок (v11)]. Фактор 5 – «содержание магния».

Нетрудно заметить, что переменные, определяющие фактор, сильнее связаны друг с другом, чем с другими переменными (табл. 3). Таким образом, факторным анализом выявлено действие пяти новых латентных переменных (факторов) на качество вина, объединяющих переменные (свойства вин) в группы по степени их влияния на факторы (главные компоненты).

В табл. 6 приведены коэффициенты A_i линейных уравнений регрессии (Фактор _{k}) для расчета значения факторов для каждого образца вина по его физико-химическим свойствам. Положительное значение фактора соответствует позитивному вкладу данного образца в группирование (разделе-

ние) вин по степени сходства, а отрицательное – негативному. Анализ рассчитанных факторов (Фактор _{k}) позволяет выяснить, какие свойства вина превалируют как в положительном, так и отрицательном значении фактора каждого из 178 образцов вина.

Кластерный анализ. На дендрограмме (рис. 1) показано объединение образцов вин методом Варда [18] в три группы по степени сходства свойств вин в зависимости от расстояния объединения (выбрана евклидова метрика объединения; для свойств вин использована z-стандартизация данных [21]).

Метод k -средних. На основе решения, полученного иерархической кластеризацией методом Варда, для реализации алгоритма k -средних выбрано 3 кластера (рис. 1). В статистическом пакете STATISTICA предлагается три варианта выбора начальных центров кластеров. В работе после анализа и сопоставления результатов кластеризации предпочтение было отдано варианту – выбрать первые N (число кластеров) наблюдений (образцов вин).

Результаты дисперсионного анализа свидетельствуют (табл. 7) о том, что распределение

Таблица 6. Коэффициенты факторных моделей $\text{Фактор}_k = A_1v1 + A_2v2 + A_3v3 + \dots + A_{13}v13$, $k = 1-5$

Переменная (свойство вина)	Варимакс-вращение нестандартизированных свойств вина. Выделение факторов методом главных компонент				
	Фактор 1	Фактор 2	Фактор 3	Фактор 4	Фактор 5
$A_1, v1$	-0.007	0.434	-0.008	-0.019	-0.148
$A_2, v2$	0.188	-0.054	0.010	-0.550	-0.111
$A_3, v3$	0.033	0.078	0.601	0.122	0.042
$A_4, v4$	0.079	-0.248	0.443	-0.112	0.061
$A_5, v5$	-0.135	-0.105	0.039	0.056	0.879
$A_6, v6$	0.289	0.062	0.083	-0.066	-0.141
$A_7, v7$	0.278	0.021	0.071	-0.016	-0.099
$A_8, v8$	-0.155	0.175	0.318	0.221	-0.417
$A_9, v9$	0.347	-0.117	-0.022	-0.308	0.050
$A_{10}, v10$	-0.047	0.317	0.009	-0.190	-0.005
$A_{11}, v11$	-0.065	0.020	0.116	0.519	-0.047
$A_{12}, v12$	0.285	-0.079	0.048	-0.010	-0.138
$A_{13}, v13$	-0.042	0.363	0.036	0.187	0.018

образцов вин по трем кластерам на основе физико-химических свойств вин проведено удовлетворительно. Уровень значимости p критерия Фишера значительно меньше 0.05 для всех переменных, а наблюдаемый критерий Фишера больше критического $F_{\text{набл}} > F_{\text{кр}}$.

Табличное значение критерия Фишера $F_{\text{кр}}(2, 175, p 0.05) = 3.05$. SS_W – сумма квадратов отклонений значений каждого из предикторов (свойства образцов вин) от группового среднего значения предикторов внутри группы (кластера) – мера внутригрупповой изменчивости. SS_B – межгруп-

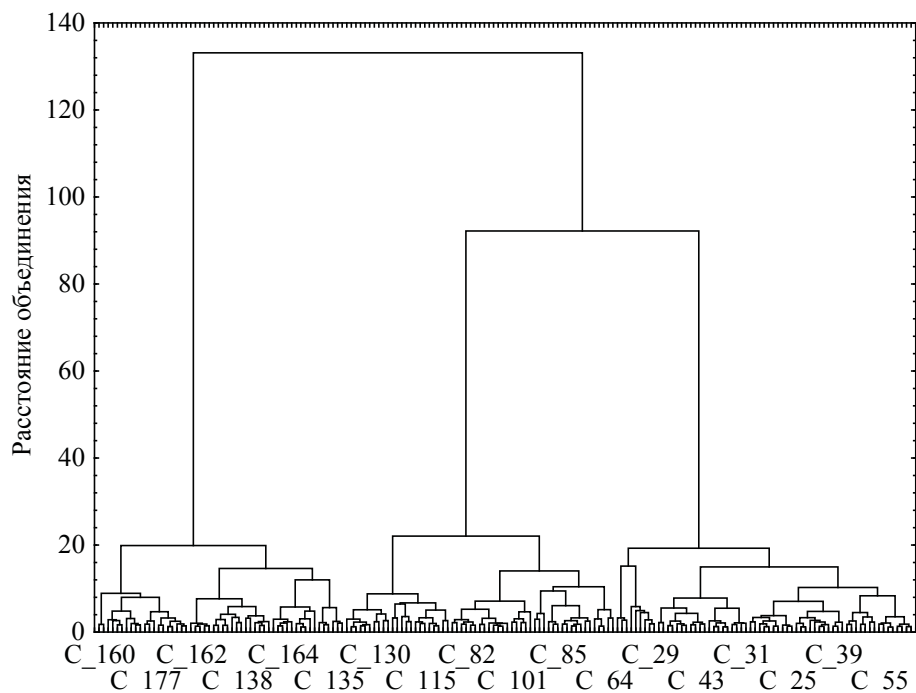
**Рис. 1.** Иерархическое дерево объединения образцов вина по степени их сходства (показана каждая десятая метка).

Таблица 7. Результаты дисперсионного анализа стандартизированных свойств вин методом k -средних

Свойство	SS_B	df_B	SS_W	df_W	$F(2, 175)$	p
v1	99.82	2	77.18	175	113.17	0.000
v2	54.25	2	122.75	175	38.67	0.000
v3	25.78	2	151.22	175	14.92	0.000
v4	38.78	2	138.22	175	24.55	0.000
v5	36.48	2	140.52	175	22.72	0.000
v6	97.32	2	79.68	175	106.88	0.000
v7	133.87	2	43.13	175	271.59	0.000
v8	46.29	2	130.71	175	30.98	0.000
v9	51.81	2	125.19	175	36.21	0.000
v10	99.34	2	77.66	175	111.92	0.000
v11	96.44	2	80.56	175	104.75	0.000
v12	126.88	2	50.12	175	221.53	0.000
v13	123.19	2	53.81	175	200.31	0.000

повая сумма квадратов отклонений средних значений предикторов в каждой из групп от суммарного среднего значения предикторов по всем группам – мера межгрупповой изменчивости [21]. df_B и df_W – межгрупповые и внутригрупповые степени свободы. Чем меньше SS_W и больше значение SS_B , тем лучше переменная характеризует принадлежность объектов к кластеру (группе) и тем «качественнее» проведена кластеризация [19].

Кластеризация вин методом k -средних выявила распределение 178 образцов вин по трем кластерам с таким набором образцов: 1 кластер – 62; 2 кластер – 65 и 3 кластер – 51.

Следует обратить внимание на то, что алгоритм k -средних минимизирует суммарное квадратичное отклонение точек (свойства вин) кластеров от центров этих кластеров:

$$V = \sum_{i=1}^k \sum_{x \in S_i} (x - \mu_i)^2, \quad (1)$$

где k – число кластеров, S_i – полученные кластеры, $i = 1, 2, \dots, k$ а μ_i – центры масс всех векторов x из кластера S_i .

При этом: (а) не гарантируется достижение глобального минимума суммарного квадратичного отклонения V , (б) результат кластеризации зависит от выбора исходных центров кластеров, (в) оптимальный выбор кластеров неизвестен, (г) число кластеров надо знать заранее. Последний

недостаток, как показано в работе, устраняется с помощью агломеративной кластеризации. Что же касается остальных трех, то результаты кластеризации, полученные алгоритмом k -средних, нуждаются в подтверждении другими методами.

Множественный дискриминантный анализ.

Для разделения образцов красных вин на группы по физико-химическим свойствам вин выполнен линейный дискриминантный анализ Фишера (LDA) на статистической платформе STATISTICA [18, 19]. Результаты, полученные при одновременном введении всех переменных для дискриминантного анализа, показаны в табл. 8. Λ -Уилкса для каждого предиктора – это отношение внутригрупповой суммы квадратов отклонений предиктора от выборочного среднего к общей сумме квадратов отклонений, иначе говоря – это отношение меры внутригрупповой изменчивости SS_W к мере общей изменчивости $SS_{Total} = SS_W + SS_B$.

Значение стандартной статистики Λ -Уилкса (0.0193) свидетельствует о высокой дискриминирующей мощности модели (1.0 – дискриминация отсутствует, 0.0 – полная дискриминация). Этот вывод также подтверждается наблюдаемым значением $F_{набл}$ -статистики, $F_{набл}(26, 326) = 77.065$, $p < 0.000$ и $F_{набл}(26, 326) > F_{кр}(26, 326)$.

Из анализа результатов табл. 8 следует, что за исключением трех свойств вин – всего фенолов (v6), фенолы нефлавоноидные (v8), проантоцианидины (v9), остальные десять демонстрируют спо-

Таблица 8. Итоги анализа дискриминантных функций. Достоверность различения трех групп образцов вин по каждому физико-химическому свойству^a

Свойство	Переменных в модели: 13; группирующая переменная: 3 группы (кластера) образцов вин; Λ-Уилкса = 0.0193; приближенное значение $F_{\text{набл}}(26, 326) = 77.065, p < 0.000; F_{\text{кр}}(26, 326) = 1.5,$ $F_{\text{кр}}(2, 163) = 3.05$					
	Λ-Уилкса	Частная Λ-Уилкса	$F_{\text{исключить}}(2, 163)$	p-Уровень	Толерантность, 1- R^2	R^2
v1	0.023	0.869	12.303	0.000	0.810	0.190
v2	0.021	0.926	6.489	0.002	0.806	0.194
v3	0.021	0.952	4.116	0.018	0.510	0.490
v4	0.021	0.951	4.169	0.017	0.544	0.456
v5	0.021	0.911	7.932	0.001	0.810	0.190
v6	0.020	0.970	2.478	0.087	0.497	0.503
v7	0.027	0.730	30.083	0.000	0.426	0.574
v8	0.020	0.990	0.811	0.446	0.746	0.254
v9	0.020	0.972	2.346	0.099	0.696	0.304
v10	0.025	0.798	20.642	0.000	0.601	0.399
v11	0.021	0.940	5.232	0.006	0.809	0.191
v12	0.024	0.804	19.848	0.000	0.750	0.250
v13	0.030	0.661	41.813	0.000	0.770	0.230

^a Курсивом выделены свойства вин, для которых p-Уровень > 0.05.

способность к дискриминации, чем больше значение Λ-Уилкса, тем более предпочтительным является это свойство в процедуре разделения образцов вин по группам. Таковыми в ряду убывания предпочтительности являются: пролин (v13) > флавоноиды (v7) > интенсивность цвета (v10) > OD280/OD315 разбавленных вин (v12) > алкоголь (v1). Частная Λ-Уилкса, характеризующая единичный вклад соответствующей переменной (свойство вина) в разделительную силу модели, подтверждает этот вывод. Она равна отношению Λ-Уилкса после добавления переменной в модель к Λ-Уилкса до добавления этой переменной. Чем меньше значение частной Λ-Уилкса, тем больший вклад этого свойства в общую дискриминацию. Наряду с этим, чем меньше значение критерия Фишера $F_{\text{искл}}$ (табл. 8), тем менее желательны свойства в модели дискриминации вин. Переменные, у которых уровень значимости $p > 0.05$, исключаются из дискриминантной модели. Толерантность (табл. 8) является мерой избыточности переменной в модели. Чем меньше ее значение, тем избыточнее переменная в модели, тем меньшую дополнительную информацию несет эта переменная [24], иначе говоря, чем ниже толерантность, тем сильнее данное свойство

связано (коррелирует) со всеми остальными (наличие мультиколлинеарности).

В табл. 9 приведены коэффициенты b_i математической модели дискриминации (линейных классификационных функций, f_k) образцов вин по их физико-химическим свойствам (2).

$$f_k = b_0 + b_1 v1 + b_2 v2 + b_3 v3 + \dots + b_{13} v13, k = 1, 2, 3. \quad (2)$$

Подставив в эти уравнения значения физико-химических свойств образцов вин, класс которых неизвестен, можно предсказать кластер (класс), к которому вина будут отнесены по наибольшему рассчитанному для каждого образца значению линейной классификационной функции.

Канонический анализ. Предсказательные модели классификации вин можно построить не только на основе линейных дискриминантных функций (LDF), но и на базе канонических дискриминантных функций (CDF), полученных с использованием алгоритмов канонического корреляционного анализа (CCA). Каноническая дискриминантная функция – это линейная комбинация дискриминантных переменных (свойств вина). Она имеет следующее математическое представление [16, 17]:

Таблица 9. Коэффициенты классифицирующих функций на основе дискриминантных нестандартизированных переменных (свойств вин)^a

Коэффициент, свойство	G_1 ($p = 0.348$)	G_2 ($p = 0.365$)	G_3 ($p = 0.287$)
$b_1, v1$	55.526	51.475	53.058
$b_2, v2$	-0.144	-0.585	1.404
$b_3, v3$	25.894	19.552	22.622
$b_4, v4$	0.979	1.550	1.693
$b_5, v5$	1.039	0.889	0.883
$b_6, v6$	-6.504	-3.680	-1.673
$b_7, v7$	16.518	10.014	2.626
$b_8, v8$	65.651	60.003	57.134
$b_9, v9$	0.242	2.149	0.586
$b_{10}, v10$	-5.339	-4.916	-2.568
$b_{11}, v11$	31.808	34.261	23.444
$b_{12}, v12$	27.752	22.911	16.132
$b_{13}, v13$	0.032	0.010	0.005
b_0	-566.269	-458.563	-467.798

^a p – апостериорная (послеопытная) вероятность, пропорциональная количеству образцов вин в каждой группе – G_1, G_2, G_3.

Таблица 10. Статистические показатели извлекаемых (ортогональных) дискриминантных функций (корней)^a

Извлечено корней	Критерий Хи-квадрат с последовательно извлеченными корнями					
	λ	R	Λ	χ^2	ν	p
0	11.011	0.957	0.020	664.708	26	0.000
1	3.252	0.875	0.235	244.605	12	0.000

^a λ – собственное значение дискриминантной функции D_{km} , R – коэффициент канонической корреляции, Λ – значение статистики Λ Уилкса, χ^2 – значение статистики Хи-квадрат, ν – число степеней свободы, p – уровень значимости соответствующего канонического корня.

$$D_{km} = u_0 + u_1 X_{1km} + u_2 X_{2km} + \dots + u_p X_{pkm}, \quad (3)$$

где D_{km} – значение канонической дискриминантной функции для m -го свойства вина в группе k ; X_{ikm} – значение дискриминантной переменной (свойства вина) X_i для m -го объекта (свойства вина) в группе k ; u_i – коэффициенты дискриминантных функций.

В табл. 10 представлены результаты канонического анализа с пошаговым критерием χ^2 для канонических корней (Root) – канонических линейных дискриминантных функций D_{km} . Первая строка содержит критерий значимости χ^2 для всех дискриминантных функций (корней). Во второй строке представлена значимость дискриминантных функций, оставшихся после удаления первой функции

и т. д. Как следует из табл. 10, статистически значимыми являются две дискриминантные функции.

В табл. 11 приведены коэффициенты u_i канонических линейных дискриминантных функций для исходных (нестандартизированных) свойств образцов вин. Первая дискриминантная функция Root 1 наиболее значимая, так как отвечает за 77.2% объясненной дисперсии свойств вин. Вторая Root 2 отвечает за 22.8% объясненной дисперсии. Подставив в дискриминантные уравнения значения свойств вин, класс которых неизвестен, можно рассчитать значения дискриминантных функций – D_{km} для каждого образца вина. Вина, для которых рассчитаны Root 1 и Root 2, будут отнесены к одной из трех групп по минимальному расстоянию до соответствующего центроида группы (класте-

Таблица 11. Коэффициенты канонических линейных классификационных функций D_{km} (Root)

Коэффициент, свойство	$D_{km} = u_0 + u_1 X_{1km} + u_2 X_{2km} + \dots + u_p X_{pkm}$	
	Root 1	Root 2
$u_1, v1$	0.300	0.788
$u_2, v2$	–0.186	0.309
$u_3, v3$	0.398	1.306
$u_4, v4$	–0.086	–0.066
$u_5, v5$	0.019	0.021
$u_6, v6$	–0.583	–0.167
$u_7, v7$	1.674	0.046
$u_8, v8$	1.028	0.475
$u_9, v9$	–0.043	–0.471
$u_{10}, v10$	–0.333	0.226
$u_{11}, v11$	1.003	–1.687
$u_{12}, v12$	1.400	–0.123
$u_{13}, v13$	0.003	0.002
u_0	–12.330	–15.076
Собственное значение, λ	11.011	3.252
% объясненной дисперсии	77.2	1.0

Таблица 12. Координаты центроидов трех групп образцов итальянских вин

Группа	Средние канонических переменных	
	Root 1	Root 2
G_1	3.759	1.345
G_2	–0.025	–2.358
G_3	–4.539	1.370

ра). Координаты центроидов кластеров (средних значений канонических переменных) приведены в табл. 12.

На рис. 2 приведена диаграмма рассеяния канонических значений образцов вин дискриминантными функциями – Root (Корень). Первая дискриминантная функция, определяющая 77.2% дискриминирующей мощности, в координатах Root 1–Root 2 четко дискриминирует образцы вин между группами G_1 и G_3. Дискриминирующей мощности (22.8%) второй дискриминантной функции недостаточно для полного разделения образцов вина между группами G_1, G_3 и группой G_2.

В табл. 13 приведены стандартизированные коэффициенты для канонических переменных. Стандартизированные коэффициенты применяются

для выявления тех переменных (свойств), которые вносят наибольший вклад в значение дискриминантной функции. Чем больше абсолютная величина коэффициента, тем больше вклад этой переменной в значение дискриминантной функции.

Для первой дискриминантной функции вклад свойства вина флавоноиды (v7) в разделяющую мощность максимален (0.830), затем разделяющая способность снижается в ряду: (0.569) пролин (v13) > (0.532) OD280/OD315 разбавленных вин (v12) > (–0.515) интенсивность цвета (v10). Причем вклад интенсивность цвета (v10) отрицательный (–0.515). Остальные девять свойств вина второстепенны. Для второй функции только два свойства вина – (0.425) алкоголь (v1) и (0.433) пролин (v13) вносят относительно заметный вклад в

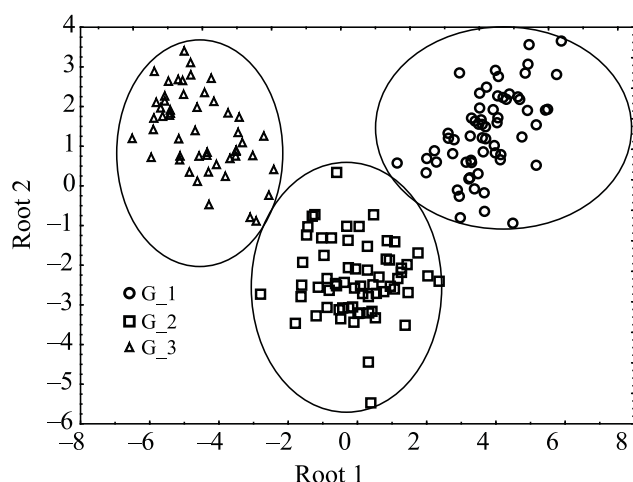


Рис. 2. Диаграмма рассеяния канонических значений свойств вин для пар значений дискриминантных функций Root 1 – Root 2.

разделяющую мощность второй дискриминантной функции.

В табл. 14 приведена классификационная матрица, которая позволяет говорить о точности дис-

криминантной процедуры, количестве правильно классифицированных образцов вин и тем самым косвенно подтвердить степень разделения классов. Тринадцать переменных (свойств вин) правильно предсказывают распределение по группам всех образцов вин. Точность предсказания в этом случае – 100%.

Таким образом, дискриминантная и каноническая модели классификации образцов итальянских красных вин на основе физико-химических свойств вин на 100% подтвердили результаты кластеризации вин методом k -средних.

Деревья классификации. Вся выборка свойств 178 образцов вин использовалась как обучающая. Для ветвления дерева выбрано дискриминантное ветвление по линейным комбинациям порядковых предикторов (физико-химические свойства вин). Остановка ветвления осуществлялась по ошибке классификации [21] (табл. 15).

В табл. 16 приведена матрица ошибок классификации глобальной кросс-проверки [24]. Из данной таблицы следует, что при глобальной кросс-проверке два образца вина Класса 2 неверно классифицированы как Класс 3, один образец вина

Таблица 13. Стандартизированные коэффициенты канонических линейных классификационных функций D_{km} (Root)^a

Коэффициент, свойство	$D_{km} = X_{1km} + u_2 X_{2km} + \dots + u_p X_{pkm}$	
	Root 1	Root 2
$u_1, v1$	0.162	0.425
$u_2, v2$	–0.174	0.289
$u_3, v3$	0.102	0.333
$u_4, v4$	–0.256	–0.196
$u_5, v5$	0.241	0.272
$u_6, v6$	–0.246	–0.071
$u_7, v7$	0.830	0.023
$u_8, v8$	0.111	0.051
$u_9, v9$	–0.021	–0.228
$u_{10}, v10$	–0.515	0.350
$u_{11}, v11$	0.156	–0.262
$u_{12}, v12$	0.532	–0.047
$u_{13}, v13$	0.569	0.433
Собственное значение, λ	11.011	3.252
% объясненной дисперсии	77.2	1.0

^a Полужирным шрифтом выделены свойства вин, определяющие разделяющую мощность дискриминантных функций.

Таблица 14. Классификационная матрица дискриминантного анализа^a

Группы	Точность предсказания, %	G_1 ($p = 0.331$)	G_2 ($p = 0.399$)	G_3 ($p = 0.270$)
G_1	100	62	0	0
G_2	100	0	65	0
G_3	100	0	0	51
Всего	100	62	65	51

^a Строки – наблюдаемая классификация, колонки – предсказанная классификация.

Таблица 15. Структура дерева классификации

Вершина	Дочерние вершины, наблюдаемые, предсказанные классы, условия ветвления						
	Левая ветвь	Правая ветвь	Класс 1	Класс 2	Класс 3	Предсказанный класс	Значение постоянной ветвления
1	2	3	62	65	51	2	0.2133
2	4	5	62	64	1	2	0.6829
3			0	1	50	3	
4			62	0	0	1	
5			0	64	1	2	

Таблица 16. Матрица ошибок классификации глобальной кросс-проверки

Класс	Класс 1	Класс 2	Класс 3
1		0	0
2	0		2
3	0	1	

Класса 3 ошибочно классифицирован как Класс 2, все остальные образцы вин классифицированы верно. При этом цена глобальной кросс-проверки (Global CV cost) составила 0.01685, стандартное отклонение (s.d. Global CV cost) цены – 0.00965.

Классификации образцов вин по физико-химическим свойствам методом Деревя решений на 98.3% (175/178) подтвердили результаты кластерного, дискриминантного и канонического анализов.

Нейросетевой анализ. Нейронные сети обучались по физико-химическим свойствам вина на выборке (train) из 70% наблюдений (образцов вин), процесс обучения контролировался (control) на контрольной выборке (15% образцов), обученная сеть независимо тестировалась на проверочной (validation) выборке (15% наблюдений), которая в процесс обучения не вовлекалась.

Многослойный персептрон. Обучение проведено итерационным методом численной оптимизации BFGS N (где N – количество итераций) – алгоритм Бройдена–Флетчера–Гольдфарба–Шанно [25, 26]. В табл. 17 приведены основные характеристики двух лучших (из 1000 обученных) нейросетевых классификаторов – многослойного персептрона MLP 13-5-3, в табл. 18 – итоги классификации. Архитектура MLP 13-5-3 означает: многослойный персептрон с 13 входными (физико-химические свойства вин) и 3 выходными (три группы вин) переменными. Персептрон состоит из трех слоев: входной – 13 нейронов, промежуточный – 5 нейронов и выходной – 3 нейрона. Обучение проведено итерационным методом численной оптимизации BFGS.

Точность кластеризации составила 100% в сравнении с результатами алгоритма k -средних.

Таблица 17. Итоги обучения нейросетевого классификатора MLP 13-5-3^a

Архитектура	Производительность обучения	Контрольная производительность	Тестовая производительность	Алгоритм обучения	Функция ошибки	Функция активации скрытых нейронов	Функция активации выходных нейронов
MLP 13-5-3	100	100	100	BFGS 23	SOS	Logistic	Exponent
MLP 13-5-3	100	100	100	BFGS 25	SOS	Logistic	Identity

^a SOS – сумма квадратов; Identity (тождественная) – $\varphi(x) = x$; Logistic (логистическая) – $\varphi(x) = 1/[1 + \exp(-tx)]$; Exponent (экспоненциальная) – $\varphi(x) = e^x$.

Таблица 18. Результаты классификации образцов вин персептроном MLP 13-5-3

		G_1	G_2	G_3	G
MLP 13-5-3	Все	62	65	51	178
	Правильно	62	65	51	178
	Правильно, %	100	100	100	100
MLP 13-5-3	Все	62	65	51	178
	Правильно	62	65	51	178
	Правильно, %	100	100	100	100

Таблица 19. Итоги кластеризации образцов вин по физико-химическим свойствам вин самоорганизующейся сетью Кохонена SOFM 13-3 (1)

Сеть	Ошибка			Алгоритм обучения	Количественный состав кластеров		
	обучающая выборка 70%	контрольная выборка 15 %	тестовая выборка 15%		Класс 1	Класс 2	Класс 3
SOFM 13-3 (1)	0.308	0.260	0.284	Kohonen 1000	61	66	51

По степени влияния на классификацию (группирование) образцов вин свойства вин располагаются в следующий ряд: $v_{12}(2.2) > v_{13}(2.0) > v_7(1.9) > v_1(1.4) > v_{10}(1.3)$, $v_{11}(1.3) > v_5(1.2)$, $v_6(1.2)$, $v_2(1.2)$, $v_9(1.2) > v_4(1.1)$, $v_8(1.1)$, $v_3(1.1)$. Техника проведения анализа чувствительности нейронной сети состоит в изменении выбранного параметра (свойства вина) в определенных пределах, при условии, что остальные параметры остаются неизменными. Таким образом, такие свойства вин:

OD280/OD315 разбавленных вин (v_{12}), пролин (v_{13}), флавоноиды (v_7), оттенок (v_{11}) – это наиболее критические переменные, которые в наибольшей степени влияют на эффективность разделения образцов вин на три класса нейронной сетью MLP 13-5-3.

Самоорганизующаяся нейронная сеть Кохонена (SOFM). В табл. 19 и 20 приведены результаты кластеризации и классификации образцов вин нейронной сетью SOFM. Один образец вина Класса 1

Таблица 20. Итоги кластеризации образцов вин по физико-химическим свойствам и трем группам самоорганизующейся сети Кохонена SOFM 16-3 (2)

Сеть	Ошибка			Алгоритм обучения	Количественный состав кластеров		
	обучающая выборка 70%	контрольная выборка 15 %	тестовая выборка 15%		Класс 1	Класс 2	Класс 3
SOFM 16-3	0.308	0.259	0.290	Kohonen 1000	62	65	51

Таблица 21. Сопоставление точности кластеризации и классификации вин методами Data Mining

Метод	<i>k</i> -means clustering	LDA	CCA	SVM	1NN	DTA	NBA	MLP	SOFM (1)/(2)
G ₁ – 62, G ₂ – 65, G ₃ – 51									
Точность, %	100	100	100	99.4	100	98.3	98.3	100	99.4/100
G ₁ – 59, G ₂ – 71, G ₃ – 48									
Точность, %	100	100	100	99.4	100	96.1	97.2	99.2	94.4/100

ошибочно отнесен к Классу 2. Точность кластеризации сетью SOFM 13-3 (1) составляет 99.4%. Сеть SOFM 16-3 (2) на 100% подтвердила результаты кластеризации методом *k*-средних.

Прогнозирующие методы Data Mining – метод опорных векторов (SVM), наивный байесовский классификатор (NBA), метод ближайших соседей (1NN) подтвердили классификацию алгоритмом *k*-средних на 99.4%, 98.3% и 100% соответственно.

По аналогии с вышеизложенным в работе проведен компьютерный анализ влияния физико-химических свойств вин на распределение вин по трем классам (G₁ – 59, G₂ – 71, G₃ – 48), приведенное в репозитории [2]. В табл. 21 показана точность разделения 178 образцов итальянских красных вин на три класса методами Data Mining в сравнении с кластеризацией алгоритмом *k*-means (верхние строки) и приведенной в репозитории [2] (нижние две строки).

Таким образом, выполненный анализ демонстрирует высокие прогнозные возможности методов Data Mining при кластеризации и классификации красных итальянских вин по физико-химическим свойствам образцов. Сочетание органолептических методов контроля качества

вина с современными методами физико-химического анализа, математического моделирования и математической статистики [20] будет способствовать выявлению грубо сфальсифицированных вин, а также надежному подтверждению качества и подлинности продуктов виноделия.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Бондарев Николай Васильевич, ORCID: <https://orcid.org/0000-0003-2443-1633>

КОНФЛИКТ ИНТЕРЕСОВ

Автор заявляет об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. Ye Ch., Li K., Jia G. // J. Phys. Conf. Ser. 2020. Vol. 1684. N 1. 012067. doi 10.1088/1742-6596/1684/1/012067
2. Kumar S., Agrawal K., Mandan N. // 2020 Int. Conf. on Computer Communication and Informatics (ICCCI). Coimbatore, India. 2020. P. 1. doi 10.1109/ICCCI48352.2020.9104095
3. Gupta Y. // Procedia Comput. Sci. 2018. Vol. 125. P. 305. doi 10.1016/j.procs.2017.12.041

4. *Er Y., Atasoy A.* // Int. J. Intelligent Syst. Appl. Eng. 2016. Vol. 4. P. 23. doi 10.18201/ijisae.265954
5. *Baykal H., Yildirim H.K.* // Crit. Rev. Food. Sci. Nutr. 2013. Vol. 53. N 5. P. 415. doi 10.1080/10408398.2010.540359
6. *Левин А.Д., Нагаев А.И., Садагов А.Ю., Карахотин С.Н.* // Аналитика и контроль. 2018. Т. 22. № 2. С. 147. doi 10.15826/analitika.2018.22.2.001
7. Wine – UCI Machine Learning Repository files. <https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data>
8. *Forina M., Armanino C., Castiglioni M., Ubigli M.* // 1986. Vitis. Vol. 25. P. 189.
9. *Forina M., Leardi R., Armanino C., Lanteri S.* // J. Chemometrics. 1990. Vol. 4. N 2. P. 191. doi 10.1002/cem.1180040210
10. *Bai X., Wang L., Li H.* // 5th Int. Conf. on Education Technology, Management and Humanities Science (ETMHS 2019). Xi'an, China, 2019. P. 1443. doi 10.25236/etmhs.2019.309
11. *Mor N.S., Asras T., Gal E., Demasia T., Tarab E., Ezekiel N., Nikapros O., Semimufar O., Gladky E., Karpenko M., Sason D., Maslov D., Mor O.* // AgriRxiv. 2022. doi 10.31220/agriRxiv.2022.00126
12. *Amerine M.A., Roessler E.B.* Wines: Their Sensory Evaluation. San Francisco: W.H. Freeman & Co, 1983. 432 p.
13. *Кишковский З.Н., Скурихин И.М.* Химия вина. М.: 1996. 462 с.
14. *Аникина Н.С., Червяк С.Н., Гниломедова Н.В.* // Аналитика и контроль. 2019. Т. 23. № 2. С. 158. doi 10.15826/analitika.2019.23.2.003
15. *Alexandre-Tudo J.L., Du Toit W.* In: Frontiers and New Trends in the Science of Fermented Food and Beverages. London: IntechOpen, 2018. P. 1. doi 10.5772/intechopen.79550
16. *Ким Дж.-О., Мьюллер Ч.У., Клекка У.Р.* Факторный, дискриминантный и кластерный анализ. М.: Финансы и статистика. 1989. 216 с.
17. *Малхорта Н.К.* Маркетинговые исследования. Практическое руководство. М.: Издательский дом «Вильямс», 2002. 960 с.
18. *Боровиков В.П.* STATISTICA. Искусство анализа данных на компьютере. СПб: Питер, 2003. 686 с.
19. *Наследов А.* IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных. СПб: Питер, 2013. 416 с.
20. *Бондарев Н.В.* // ЖОХ. 2020. Т. 90. Вып. 10. С. 1583. doi 10.31857/S0044460X20100145; *Bondarev N.V.* // Russ. J. Gen. Chem. 2020. Vol. 90. N 10. P. 1906. doi 10.1134/S107036322010014X
21. *Бондарев Н.В.* // ЖОХ. 2021. Т. 91. Вып. 3. С. 449. doi 10.31857/S0044460X21030112; *Bondarev N.V.* // Russ. J. Gen. Chem. 2021. Vol. 91. N 3. P. 409. doi 10.1134/S1070363221030117
22. *Kalika E., Bondarev N., Katin K., Kochaev A., Grekova A., Kaya S., Bauetdinov Y., Maslov M.* // J. Mol. Liq. 2023. Vol. 377. 121559. doi 10.1016/j.molliq.2023.121559
23. *Cattell R.B.* // Multivariate Behav. Res. 1966. Vol. 1. N 2. P. 245. doi 10.1207/s15327906mbr0102_10
24. *Халафян А.А.* Современные статистические методы медицинских исследований. М.: ЛКИ, 2008. 320 с.
25. *Nocedal J., Wright S.J.* Numerical Optimization. Springer, 2006. 683 p.
26. *Al-Baali M., Spedicato E., Maggioni F.* // Optimization Methods and Software. 2013. Vol. 29. N 5. P.937. doi 10.1080/10556788.2013.856909
27. *Халафян А.А., Темердашев З.А., Т.И. Гугучкина Т.И., Якуба Ю.Ф.* // Аналитика и контроль. 2017. Т. 21. N 2. С. 161. doi 10.15826/analitika.2017.21.2.010

Clustering and Classification of Red Wines by Physical-Chemical Properties Using Data Mining Methods

N. V. Bondarev^{a,*}

^a V.N. Karazin Kharkiv National University, Kharkiv, 61022 Ukraine
e-mail: n_bondarev@ukr.net

Received July 20, 2023; revised July 20, 2023; accepted August 1, 2023

A study was performed of 178 samples of red wines of Italian manufacturers, taken from the public repository for machine learning UCI. The methods of Data Mining made a computer analysis of the influence of 13 physico-chemical properties of the samples on the distribution of wines in three groups. The next classification models were built: factor, discriminant, canonical, MLP multilayer perceptron, Kohonen neural network SOFM, predictive (support vector method, Bayesian classifier, nearest neighbor method) and decision trees. The neural network classifiers SOFM 13-3, MLP 13-5-3 and clusterer SOFM 16-3 were trained. It was shown that proline, flavonoids, color intensity, proteins and alcohol determine the discriminatory power of the models.

Keywords: red wines, clustering, classification, Data Mining, STATISTICA