УДК 004.855.5.663.253.4

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ КЛАССИФИКАЦИИ ДЛЯ ОПРЕДЕЛЕНИЯ КАЧЕСТВА ВИНА ПО ЕГО ХИМИЧЕСКОМУ СОСТАВУ

Репкин Владимир Сергеевич,
repkin_vova@mail.ru

Ли Артемий Витальевич,
79131048719@yandex.ru

Семенов Григорий Юрьевич,
semenov.g.749-1@e.tusur.ru

Сермавкин Никита Игоревич,
iis.vseverske@mail.ru

Коваленко Александр Сергеевич,
a.s.kovalenko@mail.ru

Егошин Николай Сергеевич,
ens@fb.tusur.ru

Томский государственный университет систем управления и радиоэлектроники,
Россия, 634050, г. Томск, пр. Ленина, 40.

Актуальность *исследования обусловлена необходимостью решения проблем, связанных с контрафактной и некачественной продукцией в винодельческой отрасли. Несмотря на комплексную систему регулирования оборота алкогольной продукции, потребители всё же подвергаются риску неосознанного приобретения низкокачественных винных изделий. Недобросовестные производители приоритизируют снижение затрат перед обеспечением должного качества продукции, что негативно сказывается на общем потребительском опыте и подрывает репутацию более добросовестных производителей винных изделий. Последнее не способствует развитию конструктивной конкуренции на рынке и негативно сказывается как на рынке в целом, так и в части обеспечения интересов потребителя. В этом контексте исследовательская деятельность, направленная на автоматизированную объективную оценку качества вина по его химическому составу с помощью методов машинного обучения, представляется актуальной. Создание инструментов, которые обеспечивают надежный и объективный способ отличить подлинные высококачественные вина от поддельных или низкокачественных аналогов, важно для обеспечения интересов потребителей и развития конструктивной конкуренции на рынке винных продуктов.* ***Цель:*** *создание системы для автоматизированной оценки качества вина по его химическому составу на основе модели классификации, обеспечивающей лучшее соответствие эталонному набору данных.* ***Объекты:*** *модели классификации, в числе которых метод опорных векторов, дерево решений, алгоритм случайного леса, нейронная сеть, множественная регрессия и их применение для автоматизированной оценки качества вина.* ***Методы:*** *методы машинного обучения для формирования моделей классификации; статистические методы для оценки качества классификации и сравнения классификаторов.* ***Результаты.*** *С применением эталонного набора данных «Wine_Quality_Data» сформированы пять альтернативных решений на основе распространённых моделей многоуровневой классификации. С использованием статистических критериев проведено их комплексное сравнение. Наилучшим решением, положенным в основу системы автоматизированной оценки, показало себя решение на основе модели случайного леса.*

Ключевые слова: *модель классификации, качество вина, множественная регрессия, корреляционный анализ, машинное обучение.*

Введение

В современном мире люди всё чаще встречаются с поддельной продукцией. Поддельная продукция – это товары, подвергнутые производственным или модификационным процессам с целью создания внешнего сходства с оригинальными продуктами, но не обладающие качеством, безопасностью и стандартами, которые гарантируются оригинальными производителями. Эта проблема является все более распространенной и охватывает множество различных товарных групп, от одежды до электроники и медицинского оборудования [1–3]. Другой проблемой является тенденция снижения качества продукции. Согласно сведениям Федеральной службы государственной статистики, в 2021 г. было изъято 3,2 % (от количества отобранных образцов) вин, что на 2,5 процентных пункта больше чем в 2020 г. [4].

Как поддельная, так и низкокачественная продукция могут представлять угрозу для здоровья и безопасности потребителей. Известны случаи, когда поддельное вино было изготовлено с использованием неправильных ингредиентов или неправильных дозировок, что может привести к разочарованию потребителя или причинить вред здоровью. По свидетельствам в региональных средствах массовой информации, фиксировались и экстремальные случаи, когда недобросовестные производители научились выдавали смесь воды, сахара, анилинового красителя и лимонной кислоты за настоящее виноградное вино [5].

Определение подделки или товара низкого качества представляет собой сложную задачу, требующую наличия специального оборудования, а также соответствующих знаний и компетенций. Внешний вид вина не всегда может однозначно указывать на его качество или подлинность. Подделки могут быть профессионально оформлены и имитировать оригинальные бутылки, этикетки и упаковку.

Достоверное определение качества вина требует проведения химических исследований, для того чтобы измерить содержание различных химических субстанций, таких как алкоголь, кислоты, сахара и других компонентов. Проведение соответствующих исследований требует особой экспертизы и специального оборудования. Однако понимание химического состава продукта не является достаточным условием для вынесения решения о его подлинности или качестве. Формализованное сопоставление концентрации тех или иных субстанций с некоторым допустимым уровнем позволяет выявить очевидно контрафактный продукт, не отвечающий требованиям безопасности и представляющий непосредственную угрозу для жизни и здоровья потребителя, однако такого

подхода недостаточно для выявления низкокачественных или низкопробных продуктов. Решение, основанное на сравнение химического состава с некоторым эталоном, соответствующим торговой марки или производителю, также не является оптимальным, поскольку состав вина может изменяться в зависимости от многих сезонных, географических и логистических факторов [6–8].

В этой связи автоматизированная и объективная оценка качества вина по его химическому составу с помощью методов машинного обучения представляется актуальной. Создание инструментов, которые обеспечивают надежный и объективный способ отличить подлинные высококачественные вина от поддельных или низкокачественных аналогов, важно для обеспечения интересов потребителей и развития конструктивной конкуренции на рынке.

Среди методов машинного обучения известны и находят широкое применение бинарная и множественная, одноклассовая и мультиклассовая классификации. Например, данные типы классификации используются в таких задачах, как определение спама в электронной почте, распознавание изображений, определение тональности текста, распознавание рукописных цифр, определение жанра музыки и многих других [9]. В системе для автоматизированной оценки качества вина по его химическому составу тоже можно использовать модель классификации. При этом химический состав можно определять в лаборатории, аккредитованной Федеральной службой по аккредитации.

В данной работе решается задача множественной одноклассовой классификации посредством нескольких моделей и определяется наиболее эффективная для набора данных «*Wine_Quality_Data*» [10]. Набор данных описывает образцы португальского вина «*Vinho Verde*» и включает в себя концентрацию основных химических субстанций, присутствующих в продукте, а также оценки потребительских качеств вина, присвоенных экспертами в строгом соответствии с процедурой, принятой в профессиональном винодельческом сообществе. Задача классификатора, соответственно, заключается в том, чтобы по химическому составу вина определить его качество и присвоить ему класс, от 0 (самое низкопробное вино) до 10 (вино высочайшего качества).

Обзор исследований

Задаче формализованной оценки потребительских качеств гастрономических изделий с помощью математических моделей посвящено большое количество научных работ [11–13]. Одной из важных и популярных задач в этой области является оценка качества вина по его химическому составу, сырьевому происхождению или технологическим аспектам производства, хранения или транспортировки [14–19]. Многокритериальный характер задачи, а также её неявное описание в части данных, имеющихся на входе алгоритма оценки качества, сделали методы машинного обучения очевидным выбором для её решения.

Первые широко известные работы, посвящённые применения статистических методов и машинного обучения для оценки качества винных изделий, представлены в 2009 г. [20]. Цель исследований – построение модели для поддержки дегустационных оценок экспертами и улучшения производства вина. Было рассмотрено три подхода: метод опорных векторов, множественная регрессия и нейронная сеть. В связи с особенностями вычислительной мощности ЭВМ того времени одно тестирование занимало порядка 26 минут. В результате показатели модели с использованием метода опорных векторов оказались наилучшими. В последующем аналогичный формализованный подход применялся к различным сортам и категориям винных изделий [19, 21], а также преследовал различные цели – от оптимизации технологических процессов производства вина [22] до повышения качества подготовки профессиональных дегустаторов [23].

Взаимосвязь между химическим составом и качеством вина всегда являлась основным объектом исследования. В частности, работы [14, 24] преследовали цель посредством простых статистических методов, таких как корреляционный анализ, определить наиболее значимые факторы в химическом составе вина с точки зрения обеспечения его вкусовых качеств. Более комплексные, с точки зрения математического описания модели, системы классификации винных изделий обсуждались в [15–19]. В частности, описанные в работе [17] исследования показали, что математический аппарат множественной регрессии способен с достаточной степенью точности оценивать качество вин на основании оценок ряда физико-химических показателей исходного виноматериала. При этом оценка качества производилась как в соответствии с числовой шкалой [15, 16, 18], так и в соответствии с некоторым набором классов, отношения между которыми не были определены так явно [19]. В целом опубликованные исследования наглядно и убедительно демонстрируют, что качественно построенный классификатор на базе методов машинного обучения позволяет производить оценку винных изделий на основании их химического состава, которая по качеству не уступает полноценной экспертизе, производимой несколькими профессионалами [19].

Далее рассматриваются работы, наиболее значимые с точки зрения представленного исследования. В работе [25] для решения задачи оценки качества вина по его химическому составу используют многослойные полносвязные нейронные сети. Использовался набор данных белого португальского вина «*Vinho Verde*» (3298 экземпляров). Методом обучения нейронной сети является алгоритм обратного распространения ошибки, который использует стохастический градиентный спуск в качестве способа оптимизации. Значение метрики *accuracy* по результатам тестирования равно 45,35 %.

В исследовании [22] рассматривался подход интеллектуального анализа данных для прогнозирования вкусовых предпочтений человека в вине, основанный на легкодоступных аналитических тестах на этапе сертификации. Использован большой набор данных с образцами вина *Vinho Verde* (6497 экземпляров). Были применены два метода классификации. Алгоритм случайного леса достиг многообещающих результатов, превзойдя метод k ближайших соседей. Стоит отметить, что для оценки ка-

чества модели использовалась среднеквадратичная ошибка. Задавалось пороговое значение, и если ошибка меньше этого значения, то результат классификации принимался за истину.

В научной статье [21] применялся метод *k* ближайших соседей (метод-*k*) в сочетании с методом главных компонент (МГК). Исследование показывает, что в сочетании метод-*k* дает гораздо более простой и интерпретируемый классификатор, который обладает сопоставимой производительностью с методом-*k* на основе всех переменных. Исследования проводились на наборе данных из вин сортов Неббиоло, Гриньолино и Барбера (178 экземпляров). Для оценки показателей модели использовалась Евклидова метрика.

В работе [26] для прогнозирования качества белого вина было использовано пять алгоритмов машинного обучения. В поставленной задаче классификации наилучшими оказались *J48* и алгоритм случайного леса, превзойдя наивный байесовский алгоритм, многослойную нейронную сеть и метод опорных векторов. Набор исследуемых данных включает в себя 4898 экземпляров португальского белого вина. Сравнение моделей производилось с использованием метрик и статистики Каппы Коэна.

Применяют и другие методы классификации, например дерево решений или *M5P*. Для изменения набора данных используют технику пересэмплирования синтетического меньшинства и метод многомерного шкалирования [27–29].

В рассмотренных работах преимущественно исследовались три набора данных: из португальского вина *Vinho Verde* (6497 экземпляров), из натуральных сухих красных и белых виноградных вин российского производства (330 экземпляров) и из вин сортов Неббиоло, Гриньолино и Барбера, выращенных в регионе Пьемонт на северо-западе Италии (178 экземпляров). В рамках данной работы в качестве основного набора данных был выбран *Vinho Verde* как наиболее представительный с точки зрения номенклатуры винных изделий.

В качестве сравниваемых моделей, на базе которых будет строиться классификатор, выбраны следующие: множественная регрессия, нейронная сеть, алгоритм случайного леса, дерево решений и метод опорных векторов. Все перечисленные модели находили применение в тех или иных рассмотренных работах, хотя их системное сравнение применительно к сопоставимым задачам не производилось. Несмотря на то, что в исследовательских работах качество классификации характеризуется преимущественно только параметром точности (*accuracy*), в данной работе для обеспечения объективности сравнения будут использоваться все основные метрики, в частности точность и F1-мера, принятые для оценки моделей машинного обучения.

Методы и данные

Для проведения исследований был выбран набор данных «*Wine_Quality_Data*», который содержит сведения об экземплярах португальского вина «*Vinho Verde*» из провинции Минью, расположенной на севере Португалии. Данные были собираны с мая 2004 по февраль 2007 гг., при этом образцы вин были протестированы в официальном государственном регуляторном органе (*Comissao Vinicola Regional Vinhos Verdes*) [30].

Каждый экземпляр вина оценивался минимум тремя экспертами (с использованием слепых дегустаций) по шкале от 0 (самое низкопробное вино) до 10 (вино высочайшего качества). Это один из крупнейших наборов данных, который не имеет выбросов и пропущенных значений и является общедоступным [31]. Набор состоит из одиннадцати факторов (концентрации химических субстанций) и одной зависимой величины (вкусовые качества вина). Используются следующие факторы: *fixed acidity* (фиксированная кислотность), *volatile acidity* (летучая кислотность), *citric acid* (лимонная кислота), *residual sugar* (остаточный сахар), *chlorides* (хлориды), *free sulfur dioxide* и *total sulfur dioxide* (свободный диоксид серы и общий диоксид серы), *density* (плотность), *pH* (показатель кислотности вина), *sulphates* (сульфаты), *alcohol* (содержание алкоголя в вине). В качестве зависимой величины выступает *Quality* – качество вина. Выборка набора данных – 6497 записи.

В работе проводится предварительный анализ данных для проверки наличия зависимости между химическим составом вина и его качеством, а также для отбора наиболее значимых факторов, которые будут использованы при построении моделей классификации. Для этого используется корреляционный анализ – парный коэффициент корреляции (ПКК) и множественный коэффициент корреляции (МКК), а также регрессионный анализ – коэффициент эластичности, бета и дельта коэффициенты.

Для отбора факторов был использован ПКК, который показывает связь между факторами и *quality* (для удобства зависимая величина *quality* в корреляционном анализе будет рассматриваться как фактор). В матрице парной корреляции для всех факторов определяются пары рядов, для которых коэффициенты не значимы с помощью *t*-критерия Стьюдента. На основе теоретического анализа в сочетании с использованием статистических приемов определяются факторы, которые будут применены при обучении моделей.

Для проверки, имеет ли рациональный смысл решать задачу классификации в исследуемом наборе данных, вычисляется МКК. МКК помогает определить тесноту связи качества вина с его химическим составом. Для оценки значимости МКК использован критерий Фишера [32].

Регрессионный анализ включает в себя анализ вклада факторов. При анализе влияния факторов важную роль играют коэффициенты регрессионной модели. Но с их помощью нельзя сопоставить факторы по степени их влияния на зависимую переменную из-за различия единиц измерения и разной степени колеблемости, поэтому необходимо использовать коэффициенты эластичности, бета и дельта коэффициенты [33].

Одним из классификаторов в данном исследовании является линейная модель множественной регрессии, которая описывает зависимость результирующей величины от отобранных факторов. В рассматриваемом наборе данные факторы могут иметь мультиколлинеарность. Мультиколлинеарность устраняется стратегией шагового

отбора, а именно методом исключения. Метод пошаговой регрессии основан на последовательном исключении факторов с помощью *t*-критерия [34].

Все сравниваемые классификаторы были программно реализованы на языке *Python* (модуль *sklearn*). Эвристическим методом для каждого из классификаторов были определены субоптимальные параметры, представленные в табл. 1 [35–38]. Не упомянутые параметры имеют значения по умолчанию.

Таблица 1. *Значения параметров моделей классификации*
Table 1. *Parameter values of classification models*

Модель классификации Classification model	Параметры Options [39]
Нейронная сеть Neural network	Функция активации для слоя ввода: *Activation*=«*tanh*». Функция активации для скрытого слоя: *Activation*=«*relu*». Функция активации для слоя выхода: *Activation*=«*sigmoid*». Функция потерь: *LossFun*=«*mse*». Оптимизатор: *Optimizer*=«*Adam*». Размер пакета обучения: *Batch*=32. Число входных нейронов: *Neurons_input*=6497. Число скрытых нейронов: *Neurons_hidden*=12993. Количество эпох: *Epochs*=10.
Метод опорных векторов Support vector machine	Параметр регуляризации: *C*=1. Ядро, определяющее вид функции: *Kernel*=«*rbf*». Значение коэффициента для ядра: *Gamma*=«*scale*».
Дерево решений Decision tree	Критерий оценки качества разбиения: *Criterion*=«*gini*». Максимальная глубина дерева: *Max_depth*=«None». Минимальное кол-во образцов разделения: *Min_samples_split*=2. Минимальное кол-во образцов для листа: *Min_samples_leaf*=1. Кол-во признаков для поиска разбиения: *Max_features*=«*None*».
Алгоритм случайного леса Random forest algorithm	Кол-во деревьев*: N_estimators*=180. Максимальная глубина деревьев: *Max_depth*=20.

Для сравнения моделей осуществлена оценка показателей моделей с помощью метрик: точность и F1-мера. Точность часто используется для оценки показателей, в том числе в рассмотренных исследованиях. В данном контексте точность – это естественная характеристика, которая показывает количество правильно классифицированных объектов.

Точность как единственная метрика может быть валидна как индикатор того, применим ли классификатор на практике, но она не валидна для сравнения методов классификации, потому что может быть недостаточно информативной. Точность не учитывает положительную прогностическую ценность и полноту предсказаний классификатора, а также некорректно оценивает модели, обучаемые на несбалансированных данных. В данной работе в наборе данных два класса составляют большую часть выборки. В этом случае модель может правильно классифицировать большинство примеров в большем классе, но ошибаться в меньших классах, соответственно, точность может дать завышенную оценку производительности модели.

Предпочтительнее будет использовать метрику F1-мера. F1-мера – это гармоническое среднее между положительной прогностической ценностью (*precision*) и полнотой (*recall*) в задачах классификации. Ее значение находится в диапазоне от 0 до 1, где 1 означает идеальную точность и полноту, а 0 – неудовлетворительные результаты. F1-мера вычисляется по формуле (1). При вычислении данной метрики существуют различные подходы к усреднению значений по классам. Для выбранного набора корректно будет применить подход *weighted*, чтобы учитывался размер каждого класса путем присвоения различных весов классам в зависимости от их представленности в наборе данных [40]. Соответственно, данная метрика валидна для сравнения моделей, так как учтены положительная прогностическая ценность и полнота, а также несбалансированность данных.

$$F_1 = \frac{2\cdot(presicion\cdot recall)}{(presicion+recall)}, \quad (1)$$

где *presicion* – отношение количества правильно отнесенных к классу объектов ко всем объектам класса в выборке; *recall* – отношение количества правильно отнесенных к классу объектов ко всем объектам, отнесенным классификатором к данному классу.

Результаты и обсуждение

Набор данных состоит из 6497 строк. Классы упорядочены и сильно не сбалансированы, что можно увидеть на рис. 1. Исходный набор был разделен на обучающую (70 % – 4549 записей) и валидационную (30 % – 1949 записи) выборки.

Отбор наиболее значимых факторов способствует повышению эффективности и точности классификации и позволяет уменьшить возможность переобучения модели на шумовых или незначимых переменных. В этой связи были вычислены коэффициенты парной корреляции между качеством вина и другими признаками. Сила связи оценивалась по шкале Чеддока. Самая сильная связь (заметная, по шкале Чеддока) – у *alcohol* и *density*. Факторы *volatile acidity* и *pH* имеют незначительную связь. Коэффициенты данных факторов очень малы, поэтому они могут быть исключены при работе с машинным обучением. Также был рассчитан МКК, который значим и равен 0,54. Это умеренная связь, соответственно, в исследуемом наборе данных химический состав влияет на качество вина. Таким образом, проведение задачи классификации имеет рациональный смысл, поскольку алгоритмы не будут делать предсказания случайным образом. Они будут основаны на обучении и структурированных методах, что позволит получить результаты близкие к достоверным.

Результаты регрессионного анализа, а именно значения коэффициента эластичности и дельта, бета коэффициентов можно увидеть в табл. 2. При анализе полученных значений был выявлен фактор, который имеет наибольший вклад, – *alcohol*. У данного фактора все три коэффициента имеют высокие значения. Это означает, что увеличение или уменьшение уровня алкоголя может привести к заметным изменениям качества вина. Данный факт учитывается при машинном обучении путем добавления веса для факторов и выделения главных признаков.

Quality	Count
0	0
1	0
2	0
3	30
4	216
5	2138
6	2836
7	1079
8	193
9	5
10	0

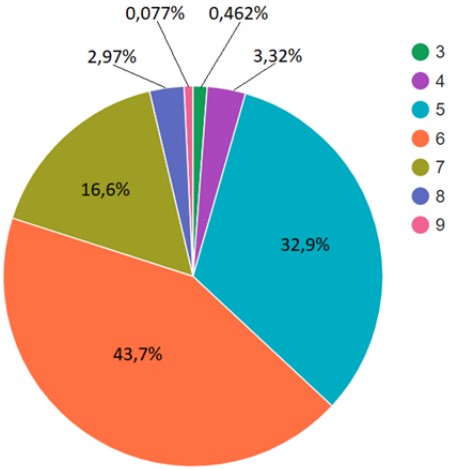

Рис. 1. *Диаграмма распределения вин по качеству*
Fig. 1. *Diagram of the distribution of wines by quality*

Таблица 2. *Значения коэффициентов для оценки влияния факторов на quality*
Table 2. *Values of the coefficients for assessing the influence of factors on quality*

Коэффициенты Odds	Факторы/Factors								
	Volatile acidity	Residual sugar	Chlorides	Free sulfur dioxide	Total sulfur dioxide	Density	PH	Sulphates	Alcohol
Эластичность Elasticity	–0,08	0,01	–0,012	0,03	–0,04	0,309	0,09	0,06	0,61
Бета Beta	–0,27	0,07	–0,049	0,15	–0,14	0,005	0,03	0,12	0,43
Дельта Delta	0,07	–0,004	0,009	0,01	0,003	–0,001	0,001	0,01	0,19

Для каждой модели классификации были вычислены значения метрик F1-мера и точности по валидационной выборке (табл. 3). Наилучшей моделью классификации для набора данных «*Wine_Quality_Data*» является алгоритм случайного леса со значением F1-меры, равным 66 %. Худшей моделью оказалась нейронная сеть, F1-мера которой равна 46 %. Стоит отметить, что даже 46 % является неплохим результатом, так как классификатору необходимо отнести объект к одному из одиннадцати классов. Также можно было бы сравнивать методы по времени выполнения алгоритмов, но каждый метод работает быстро для исследуемого набора данных (до 5 минут).

Таблица 3. *Значения F1-меры и точности для сравнения моделей*
Table 3. *F1-measure and accuracy values for comparing models*

Модель классификации Classification model	F1-мера F1-measure	Точность Accuracy
	%	
Алгоритм случайного леса Random forest algorithm	66	69
Дерево решений/Decision tree	60	63
Метод опорных векторов Support vector machine	55	59
Линейная модель множественной регрессии Linear multiple regression model	49	53
Нейронная сеть/Neural network	46	54

Если сравнивать модели по значениям точности, то выводы идентичны. Однако стоит помнить, что точность отражает лишь долю верных ответов классификатора, а при оценке моделей важно одновременно учитывать как положительную прогностическую ценность, так и полноту [40]. В данном случае при оценивании с помощью точности получается завышенная оценка производительности у каждой модели.

Если сравнить полученные результаты с результатами предыдущих исследований, можно заметить сходство. Другие исследования [22, 26] также указывали на эффективность алгоритма случайного леса в задаче классификации качества вина. Это свидетельствует о стабильности и повторяемости результатов, подтверждая применимость данного алгоритма.

Тем не менее следует отметить, что возможны расхождения в результатах исследований. В работе [23] наилучшим оказался алгоритм *M5 Prime*, который превзошел алгоритм случайного леса. Эти расхождения могут быть обусловлены разными факторами, такими как различные выборки данных, применение разных методов анализа, использование других предобработок данных. Важно учитывать эти различия при интерпретации результатов и проведении последующих исследований.

Стоит отметить, что в дальнейших исследованиях можно увеличить объем данных и вариативность его содержания. Включение в набор данных вин разных стран может способствовать более обобщенным и надежным выводам, а также расширить применимость разработанных моделей и систем в различных контекстах.

Для проверки и оценки производительности модели на реальных данных, с целью убеждения в ее эффективности и применимости в реальных ситуациях, произво-

дится апробация модели. В рамках данного исследования для апробации было взято вино «Chateau Tamagne Cabernet», которое является лидером по соотношению цены и качества в России в 2022 г. [41]. Химический состав данного вина*: fixed acidity*=8,2, *volatile acidity*=0,28, *citric acid*=0,4, *residual sugar*=2,4, *chlorides*=0,05, *free sulfur dioxide*=4, *total sulfur dioxide*=10, *density*=0,99, *pH*=3,32, *sulphate*=0,71, *alcohol*=13. Алгоритм случайного леса предсказал качество вина равное 7, что говорит о соответствии модели потребительскому выбору.

Заключение

Результат этой работы может быть применен в винодельческой промышленности. На этапе сертификации проверка качества вина должна проводиться дегустаторами-людьми. Тем не менее оценки основаны на опыте и знаниях экспертов, которые подвержены субъективным факторам. Предлагаемый подход, основанный на объективных тестах, может быть интегрирован в систему поддержки принятия решений, способствуя скорости и качеству работы энолога. Например, эксперт может повторить дегустацию только в том случае, если его оценка далека от той, которая предсказывается моделью машинного обучения. Кроме того, модель может быть использована для обучения студентов-энологов и повышения качества подготовки профессиональных дегустаторов. Также ритейлерам вина можно с помощью данной модели классификации определить качество вина лишь по его химическому составу, что может способствовать уменьшению количества недобросовестных поставщиков и производителей вина.

Однако, несмотря на полученные результаты, следует отметить, что проблема контрафактной и низкокачественной продукции в винодельческой отрасли не может быть полностью решена только с помощью автоматизированной оценки качества по химическому составу. Дальнейшие исследования и разработки должны основываться на комплексном подход, который включает не только химические параметры, но и другие факторы, такие как процессы производства, маркировка и лабораторные анализы.

Таким образом, хотя проведенное исследование является важным шагом в решении проблемы контрафакции и низкого качества винных изделий, требуется дальнейшая работа и совместные усилия от участников отрасли. Через совместный труд и применение инновационных подходов можно достичь более надежной защиты интересов потребителей и развития конструктивной конкуренции на рынке виноделия.

СПИСОК ЛИТЕРАТУРЫ

1. Аналитики оценили в 10 % рост продаж поддельных товаров в России в 2021 году. URL: https://www.forbes.ru/biznes/448813-analitiki-ocenili-v-10-rost-prodaz-poddel-nyh-tovarov-v-rossii-v-2021-godu (дата обращения: 04.04.2023).
2. «Количество контрафакта растет»: эксперт об алкогольном рынке России. URL: https://radiokp-ru.turbopages.org/radiokp.ru/s/obschestvo/kolichestvo-kontrafakta-rastet-ekspert-ob-alkogolnom-rynke-rossii_nid559979_au51752au (дата обращения: 04.04.2023).
3. Фальсификат заполонил алкогольный рынок России. URL: https://www.osnmedia.ru/video/falsifikat-zapolonil-alkogolnyj-rynok-rossii/ (дата обращения: 04.04.2023).
4. Российский статистический ежегодник. 2022. – М.: Росстат, 2022. – 691 с. URL: https://rosstat.gov.ru/storage/mediabank/Ejegodnik_2022.pdf (дата обращения: 04.04.2023)
5. Белое и красное, фальсифицированное. URL: https://www.rosbalt.ru/piter/2016/05/08/1512604.html (дата обращения: 04.04.2023)
6. Червова Н.В., Ивашкин М.В. Российский бизнес и проблема фальсификации товаров: современное состояние и способы решения проблемы // Гуманитарные, социально-экономические и общественные науки. – 2021. – № 3. – С. 241–244. DOI: 10.23672/d6487-2749-7537-v EDN: TSXOJF.
7. Иванова А.М. Проблема фальсификации винодельческой продукции и перспективные направления ее решения // Научное обеспечение агропромышленного комплекса: Сборник статей по материалам 74-й научно-практической конференции студентов по итогам НИР за 2018. – Краснодар, 26 апреля 2019. – Краснодар: Кубанский государственный аграрный университет имени И.Т. Трубилина, 2019. – С. 792–795. EDN: EXADIT
8. Агеева Н.М., Гугучкина Т.И., Оселедцева И.В. Обеспечение качества и безопасности винодельческой продукции – важнейшая государственная задача // Пищевая промышленность. – 2010. – № 12. – С. 50–52. EDN: NCOENX
9. Duda R.O., Hart P.E., Stork D.G. Pattern Classification. – NY: Wiley-Interscience, 2012. – 688 p.
10. Wine_Quality_Data. URL: https://www.kaggle.com/datasets/ghassenkhaled/wine-quality-data?select=Wine_Quality_Data.csv (дата обращения: 04.04.2023).
11. Омарова М.М. Применение методов математического моделирования объектов и процессов в производстве пищевой продукции // StudNet. – 2021. – № 2. URL: https://cyberleninka.ru/article/n/primenenie-metodov-matematicheskogo-modelirovaniya-obektov-i-protsessov-v-proizvodstve-pischevoy-produktsii/viewer (дата обращения: 04.04.2023).
12. Modelling, responses and applications of time-temperature indicators (TTIs) in monitoring fresh food quality / T. Gao, Y. Tian, Z. Zhu, D.W. Sun // Trends in Food Science & Technology. – 2020. – V. 99. – P. 311–322. DOI: 10.1016/j.tifs.2020.02.019
13. Improving the quality of cupcakes by optimizing the recipe using a mathematical modeling method / A. Tkachenko, O. Olkhovska, O. Chernenko, T. Chilikina, Y. Basova // Eastern-European Journal of Enterprise Technologies. – 2022. – V. 6. – № 11. – P. 99–108. DOI: 10.15587/1729-4061.2022.268973
14. Виноградные вина, проблемы оценки их качества и региональной принадлежности / Ю.Ф. Якуба, А.А. Каунова, З.А. Темердашев, В.О. Титаренко, А.А Халафян. // Аналитика и контроль. – 2014. – Т. 18. – № 4. – С. 344–372. EDN: TAFQXV
15. Аналитический контроль качества вин и виноматериалов / Н.Т. Сиюхова, З.Т. Тазова, Л.В. Лунина, З.Н. Блягоз // Новые технологии. – 2022. – Т. 18. – № 4. – С. 78–94. DOI: https://doi.org/10.47370/2072-0920-2022-18-4-78-94
16. Гугучкина Т.И., Лопатина Л.М. Математическая модель прогноза качества виноградных вин // Виноделие и виноградарство. – 2003. – № 4. – С. 23–24.
17. Прогнозирование качества игристого вина на основе определения дополнительных показателей физико-химического состава исходного виноматериала / Е.В. Дубинина, Л.А. Оганесянц, В.А. Песчанская, В.К. Семипятный, А.А. Чистова // Пиво и напитки. – 2020. – № 1. – С. 9–13. DOI: 10.24411/2072-9650-2020-10010 EDN: UMEONQ.
18. Дубинина Е.В., Песчанская В.А., Семипятный В.К. Прогнозирование качества красных игристых вин // Контроль качества продукции. – 2021. – № 12. – С. 43–47. DOI: 10.35400/2541-9900-2021-12-43-47 EDN: CCHHZD.
19. Якуба Ю.Ф., Темердашев З.А., Халафян А.А. Применение классификационного анализа для оценки качества вин в номинальной шкале // Журнал аналитической химии. – 2016. – Т. 71. – № 2. – С. 212–222. DOI: 10.7868/S004445021602016X EDN: TFDLFT.
20. Modeling wine preferences by data mining from physicochemical properties / P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis // Decision Support Systems. – 2009. – V. 47. – № 4. – P. 547–553. DOI: https://doi.org/10.1016/j.dss.2009.05.016

21. Classification of wines using principal component analysis / J. Barth, D. Katumullage, C. Yang, J. Cao // Journal of Wine Economics – 2021. – V. 16. – № 1. – P. 56–67. DOI: 10.1017/jwe.2020.35
22. Wine quality analysis using machine learning algorithms / U. Mahima Gupta, Y. Patidar, A. Agarwal, K.P. Singh // Micro-Electronics and Telecommunication Engineering / Eds. D.K. Sharma, V.E. Balas, Le Hoang Son, R. Sharma, K. Cengiz. – Singapore: Springer, 2020. – P. 11–18.
23. Mohit G., Vanmathi C. A study and analysis of machine learning techniques in predicting wine quality // International Journal of Recent Technology and Engineering – 2021. – V. 10. – № 1. – P. 314–321. DOI: 10.35940/ijrte.A5854.0510121
24. Титова Е.М. Анализ оценки качества вина на основе данных о его химическом составе // Инновационные аспекты развития науки и техники: сборник статей V Международной научно-практической конференции. – Саратов: НОО «Цифровая наука», 2021. – С. 62–70.
25. Сидорина С.А., Воронова Л.И. Применение методов интеллектуального анализа данных в задаче классификации экспертных оценок качества винных изделий // Научное обозрение. Педагогические науки. – 2019. – № 4-3. – С. 76–78. EDN: XDSNPZ
26. Analysis of white wine using machine learning algorithms / M. Koranga, R. Pandey, M. Joshi, M. Kumar // Materials Today: Proceedings – 2021. – V. 46. – Pt. 20. – P. 11087–11093. DOI: https://doi.org/10.1016/j.matpr.2021.02.229
27. Anurag S., Atul K. Wine quality and taste classification using machine learning model // International Journal of Innovative Research in Applied Sciences and Engineering. – 2020. – V. 4. – Iss. 4. – P. 715–721. DOI: 10.29027/IJIRASE.v4.i4.2020.715-721
28. Evaluation and analysis model of wine quality based on mathematical model / Z. Yunhui, L. Yingxia, W. Lubin, D. Hanjiang, Z. Yuanbiao, G. Hongfei, G. Zisheng, W. Shuyang, L. Yao // Studies in Engineering and Technology. – 2019. – V. 6. – № 1. – P. 6–15. DOI: https://doi.org/10.11114/set.v6i1.3626
29. Shaw B., Suman A.K., Chakraborty B. Wine quality analysis using machine learning // Emerging Technology in Modelling and Graphics. – 2019. – V. 937. – P. 239–247. DOI: 10.1007/978-981-13-7403-6_23
30. Portuguese Wine – Vinho Verde. Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV). URL: http://www.vinhoverde.pt (дата обращения: 04.04.2023).
31. Overview and importance of data quality for machine learning tasks / A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. Mittal, V. Munigala // KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. – 2020. – P. 3561–3562. DOI: https://doi.org/10.1145/3394486.3406477.
32. Asuero A.G., Sayago A., González A.G. The correlation coefficient: an overview // Critical Reviews in Analytical Chemistry. – 2006. – V. 36. – № 1. – P. 41–59. DOI: https://doi.org/10.1080/10408340500526766
33. Maulud D.H., Abdulazeez A.M. A review on linear regression comprehensive in machine learning // Journal of Applied Science and Technology Trends. – 2020. – V. 1. – № 4. – P. 140–147. DOI: https://doi.org/10.38094/jastt1457
34. Орлова И.В., Филонова Е.С. Выбор экзогенных факторов в модель регрессии при мультиколлинеарности данных // Международный журнал прикладных и фундаментальных исследований. – 2015. – № 5-1. – С. 108–116. EDN: UAUQCV
35. Fan J., Han F., Liu H. Challenges of Big Data analysis // National Science Review. – 2014. – V. 1. – Iss. 2. – P. 293–314. DOI: https://doi.org/10.1093/nsr/nwt032
36. Raschka S., Patterson J., Nolet C. Machine learning in Python: main developments and technology trends in data science, machine learning, and artificial intelligence // Information. – 2020. – V. 11. – № 4. DOI: https://doi.org/10.3390/info11040193. URL: https://www.mdpi.com/2078-2489/11/4/193 (дата обращения: 04.04.2023).
37. A comprehensive survey on support vector machine classification: applications, challenges and trends / J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez // Neurocomputing. – 2020. – V. 408. – P. 189–215. DOI: 10.1016/j.neucom.2019.10.118
38. Boateng E., Otoo J., Abaye D. Basic tenets of classification algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: a review // Journal of Data Analysis and Information Processing – 2020. – V. 8. – P. 341–357. DOI: 10.4236/jdaip.2020.84020
39. Scikit-learn machine learning in Python. URL: https://scikit-learn.org/stable/index.html (дата обращения: 04.04.2023).
40. Hossin M., Sulaiman M.N. A review on evaluation metrics for data classification evaluations // International Journal of Data Mining & Knowledge Management Process – 2015. – V. 5. – № 2. – P. 01–11. DOI: 10.5121/ijdkp.2015.5201
41. «Винный гид России-2022»: названы лучшие красные и белые вина страны. URL: https://roskachestvo.gov.ru/news/vinnyy-gid-rossii-2022-nazvany-luchshie-krasnye-i-belye-vina-strany/ (дата обращения: 04.04.2023).

Дата поступления: 10.04.2023 г.
Дата принятия: 17.06.2023 г.

Информация об авторах

Репкин В.С., техник научно-инжинирингового центра «Доверенные системы с использованием квантовых технологий и криптографии» Томского государственного университета систем управления и радиоэлектроники.

Ли А.В., техник научно-инжинирингового центра «Доверенные системы с использованием квантовых технологий и криптографии» Томского государственного университета систем управления и радиоэлектроники.

Семенов Г.Ю., техник научно-инжинирингового центра «Доверенные системы с использованием квантовых технологий и криптографии» Томского государственного университета систем управления и радиоэлектроники.

Сермавкин Н.И., техник научно-инжинирингового центра «Доверенные системы с использованием квантовых технологий и криптографии» Томского государственного университета систем управления и радиоэлектроники.

Коваленко А.С., техник научно-инжинирингового центра «Доверенные системы с использованием квантовых технологий и криптографии» Томского государственного университета систем управления и радиоэлектроники.

Егошин Н.С., кандидат технических наук, доцент кафедры комплексной информационной безопасности электронно-вычислительных систем Томского государственного университета систем управления и радиоэлектроники.

UDC 004.855.5.663.253.4

COMPARATIVE ANALYSIS OF CLASSIFICATION MODELS FOR DETERMINING THE QUALITY OF WINE BY ITS CHEMICAL COMPOSITION

Vladimir S. Repkin,
repkin_vova@mail.ru

Artemy V. Li,
79131048719@yandex.ru

Grigory Yu. Semenov,
semenov.g.749-1@e.tusur.ru

Nikita I. Sermavkin,
iis.vseverske@mail.ru

Alexander S. Kovalenko,
a.s.kovalenko@mail.ru

Nikolai S. Egoshin,
ens@fb.tusur.ru

Tomsk State University of Control Systems and Radioelectronics,
40, Lenin avenue, Tomsk, 634050, Russia.

The relevance *of the research is caused by the need to solve problems associated with counterfeit and low-quality products in the wine industry. Despite the comprehensive system of regulation of the turnover of alcoholic products, consumers are still at risk of unconscious purchase of low-quality wine products. Unscrupulous producers prioritize cost reduction over product quality, which negatively affects the overall consumer experience and undermines the reputation of more conscientious wine producers. The latter does not contribute to the development of constructive competition in the market and has a negative impact both on the market as a whole and in terms of ensuring the interests of the consumer. In this context, research activities aimed at an automated objective assessment of the quality of wine in terms of its chemical composition using machine learning methods seem to be relevant. Creating tools that provide a reliable and objective way to distinguish genuine high-quality wines from counterfeit or low-quality counterparts is important to safeguard the interests of consumers and promote constructive competition in the wine market.* ***The purpose*** *of the research is to create a system for automated assessment of wine quality by its chemical composition based on a classification model that provides better compliance with the reference data set.* ***Objects:*** *classification models, including the support vector machine, decision tree, random forest algorithm, neural network, multiple regression and their application for automated wine quality assessment.* ***Methods:*** *machine learning methods for the formation of classification models; statistical methods for assessing the quality of classification and comparing classifiers.* ***Results.*** *Using the reference dataset «Wine_Quality_Data», five alternative solutions were formed based on common multilevel classification models. Using statistical criteria, their complex comparison was carried out. The best solution underlying the automated evaluation system proved to be the solution based on the random forest model.*

Key words: *classification model, wine quality, multiple regression, correlation analysis, machine learning.*

REFERENCES

1. *Analitiki otsenili v 10 % rost prodazh poddelnykh tovarov v Rossii v 2021 godu* [Analysts have estimated a 10 % increase in counterfeit goods sales in Russia in 2021]. Available at: https://www.forbes.ru/biznes/448813-analitiki-ocenili-v-10-rost-prodaz-poddel-nyh-tovarov-v-rossii-v-2021-godu (accessed: 4 April 2023).
2. *«Kolichestvo kontrafakta rastet»: ekspert ob alkogolnom rynke Rossii* [«The number of counterfeit products is growing»: an expert on the alcohol market in Russia]. Available at: https://radiokp-ru.turbopages.org/radiokp.ru/s/obschestvo/kolichestvo-kontrafakta-rastet-ekspert-ob-alkogolnom-rynke-rossii_nid559979_au51752au (accessed: 4 April 2023).
3. *Falsifikat zapolonil alkogolnyy rynok Rossii* [Counterfeit flooded the Russian alcohol market]. Available at: https://www.osnmedia.ru/video/falsifikat-zapolonil-alkogolnyj-rynok-rossii/ (accessed: 4 April 2023).
4. *Rossiyskiy statisticheskiy yezhegodnik. 2022* [Russian Statistical Yearbook. 2022]. Moscow, Rosstat, 2022. 691 p. Available at: https://rosstat.gov.ru/storage/mediabank/Ejegodnik_2022.pdf (accessed: 4 April 2023).
5. *Beloe i krasnoe, falsifitsirovannoe* [White and red, falsified]. Available at: https://www.rosbalt.ru/piter/2016/05/08/1512604.html (accessed: 4 April 2023).
6. Chervova N.V., Ivashkin M.V. Russian business and the problem of falsification of goods: the current state and ways to solve the problem. *Humanitarian, socio-economic and social sciences*. 2021, no. 3, pp. 241–244. In Rus. DOI: 10.23672/d6487-2749-7537-v. EDN: TSXOJF.
7. Ivanova A.M. Problema falsifikatsii vinodelcheskoy produktsii i perspektivnye napravleniya ee resheniya [The problem of falsification of wine products and promising directions for its solution]. *Nauchnoe obespechenie agropromyshlennogo kompleksa. Sbornik statey po materialam 74-y nauchno-prakticheskoy konferentsii studentov po itogam NIR za 2018 god* [Scientific support of the agro-industrial complex. Collection of articles based on the materials of the 74th scientific and practical conference of students following the results of research for 2018]. Krasnodar, Kuban State Agrarian University named after I.T. Trubilina, 2019. pp. 792–795. EDN: EXADIT.
8. Ageeva N.M., Guguchkina T.I., Oseledtseva I.V. Maintenance of quality and safety of wine-making production – the major state problem. *Food Industry*, 2010, no. 12, pp. 50–52. In Rus. EDN: NCOENX.
9. Duda R.O., Hart P.E., Stork D.G. *Pattern Classification*. NY, Wiley-Interscience, 2012. 688 p.
10. *Wine_Quality_Data*. Available at: https://www.kaggle.com/datasets/ghassenkhaled/wine-quality-data?select=Wine_Quality_Data.csv (accessed 4 April 2023).
11. Omarova M.M. Application of methods of mathematical modeling of objects and processes in food production. *StudNet*. 2021, no. 2. Available at: https://cyberleninka.ru/article/n/primenenie-metodov-matematicheskogo-modelirovaniya-obektov-i-protsessov-v-proizvodstve-pischevoy-produktsii/viewer (accessed: 4 April 2023).
12. Gao T., Tian Y., Zhu Z., Sun D.W. Modelling, responses and applications of time-temperature indicators (TTIs) in monitoring fresh food quality. *Trends in Food Science & Technology*, 2020, vol. 99, pp. 311–322. DOI: 10.1016/j.tifs.2020.02.019
13. Tkachenko A., Olkhovska O., Chernenko O., Chilikina T., Basova Y. Improving the quality of cupcakes by optimizing the recipe using a mathematical modeling method. *Eastern-European Journal of Enterprise Technologies*, 2022, vol. 6, no. 11, pp. 99–108. DOI: 10.15587/1729-4061.2022.268973
14. Yakuba Yu.F., Kaunova A.A., Temerdashev Z.A., Titarenko V.O., Halafyan A.A. Grape wines, problems of their quality and regional origin evaluation. *Analytics and control*, 2014, vol. 18, no. 4, pp. 344–372. In Rus. EDN: TAFQXV.
15. Siyukhova N.T., Tazova Z.T., Lunina L.V., Blyagoz Z.N. Analytical quality control of wines and wine materials. *New technologies*, 2022, vol. 18, no. 4, pp. 78–94. In Rus. DOI: https://doi.org/10.47370/2072-0920-2022-18-4-78-94.
16. Guguchkina T.I., Lopatina L.M. Matematicheskaya model prognoza kachestva vinogradnykh vin [Mathematical model for predicting the

quality of grape wines]. *Vinodelie i vinogradarstvo*, 2003, no. 4, pp. 23–24.
17. Dubinina E.V., Oganesyants L.A., Peschanskaya V.A., Semipyatniy V.K., Chistova A.A. Prediction of sparkling wine quality based on original wine material determination of additional indicators of physicochemical composition. *Beer and beverages*, 2020, no. 1, pp. 9–13. In Rus. DOI: 10.24411/2072-9650-2020-10010. EDN: UMEONQ.
18. Dubinina E.V., Peschansky V.A., Sempatny V.K. Forecasting the quality of red sparkling wines. *Production quality control*, 2021, no. 12, pp. 43–47. In Rus. DOI: 10.35400/2541-9900-2021-12-43-47 EDN: CCHHZD.
19. Yakuba Y.F., Temerdashev Z.A., Khalafyan A.A. Application of ranging analysis to the quality assessment of wines on a nominal scale. *Journal of Analytical Chemistry*, 2016, vol. 71, no. 2, pp. 205–214. DOI: 10.1134/S1061934816020155 EDN: WPMIXJ.
20. Cortez P., Cerdeira A., Almeida F., Matos T., Reis J. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 2009, vol. 47, no. 4, pp. 547–553. DOI: https://doi.org/10.1016/j.dss.2009.05.016
21. Barth J., Katumullage D., Yang C., Cao J. Classification of wines using principal component analysis. *Journal of Wine Economics*, 2021, vol. 16, Iss. 1, pp. 56–67. DOI: 10.1017/jwe.2020.35
22. Mahima Gupta U., Patidar Y., Agarwal A., Singh K.P. Wine quality analysis using machine learning algorithms. *Micro-Electronics and Telecommunication Engineering*. Eds. D.K. Sharma, V.E. Balas, Le Hoang Son, R. Sharma, K. Cengiz. Singapore, Springer, 2020. pp. 11–18.
23. Mohit G., Vanmathi C. A study and analysis of machine learning techniques in predicting wine quality. *International Journal of Recent Technology and Engineering*, 2021, vol. 10, no. 1, pp. 314–321. DOI: 10.35940/ijrte.A5854.0510121
24. Titova E.M. Analiz otsenki kachestva vina na osnove dannykh o ego khimicheskom sostave [Analysis of wine quality assessment based on data on its chemical composition]. *Innovatsionnye aspekty razvitiya nauki i tekhniki. Sbornik statey V Mezhdunarodnoy nauchno-prakticheskoy konferentsii* [Innovative aspects of the development of science and technology. Collection of articles of the V International Scientific and Practical Conference]. Saratov, Tsifrovaya nauka Publ., 2021. pp. 62–70.
25. Sidorina S.A., Voronova L.I. Application of data mining methods in the task of classifying expert quality assessments of wine products. *Nauchnoe obozrenie. Pedagogicheskie nauki*, 2019, no. 4-3, pp. 76–78. In Rus. EDN: XDSNPZ.
26. Koranga M., Pandey R., Joshi M., Kumar M. Analysis of white wine using machine learning algorithms. *Materials Today: Proceedings*, 2021, vol. 46, P. 20, pp. 11087–11093. DOI: https://doi.org/10.1016/j.matpr.2021.02.229
27. Anurag S., Atul K. Wine quality and taste classification using machine learning model. *International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)*, 2020, vol. 4, Iss. 4, pp. 715–721. DOI: 10.29027/IJIRASE.v4.i4.2020.715-721
28. Yunhui Z., Yingxia L., Lubin W., Hanjiang D., Yuanbiao Z., Hongfei G., Zisheng G., Shuyang W., Yao L. Evaluation and analysis model of wine quality based on mathematical model. *Studies in Engineering and Technology*, 2019, vol. 6, no. 1, pp. 6–15. DOI: https://doi.org/10.11114/set.v6i1.3626
29. Shaw B., Suman A.K., Chakraborty B. Wine quality analysis using machine learning. *Emerging Technology in Modelling and Graphics*, 2019, vol. 937, pp. 239–247. DOI: 10.1007/978-981-13-7403-6_23
30. *Portuguese Wine – Vinho Verde. Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV)* [Portuguese Wine – Vinho Verde. Commission for Viticulture of the Vinho Verde Region (CVRVV)]. Available at: http://www.vinhoverde.pt (accessed: 4 April 2023).
31. Jain A., Patel H., Nagalapatti L., Gupta N., Mehta S., Guttula S., Mujumdar S., Afzal S., Mittal R., Munigala V. Overview and importance of data quality for machine learning tasks. *KDD '20: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 3561–3562. DOI: https://doi.org/10.1145/3394486.3406477.
32. Asuero A.G., Sayago A., González A.G. The correlation coefficient: an overview. *Critical Reviews in Analytical Chemistry*, 2006, vol. 36, no. 1, pp. 41–59. DOI: https://doi.org/10.1080/10408340500526766
33. Maulud D.H., Abdulazeez A.M. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 2020, vol. 1, no. 4, pp. 140–147. DOI: https://doi.org/10.38094/jastt1457
34. Orlova I.V., Filonova E.S. The choice of exogenous factors in the regression model with multicollinearity in the data. *Mezhdunarodny zhurnal prikladnykh i fundamentalnykh issledovaniy*, 2015, no. 5-1, pp. 108–116. In Rus. EDN: UAUQCV
35. Fan J., Han F., Liu H. Challenges of Big Data analysis. *National Science Review*, 2014, vol. 1, Iss. 2, pp. 293–314. DOI: https://doi.org/10.1093/nsr/nwt032
36. Raschka S., Patterson J., Nolet C. Machine learning in Python: main developments and technology trends in data science, machine learning, and artificial intelligence. *Information*, 2020, vol. 11, no. 4. DOI: https://doi.org/10.3390/info11040193. Available at: https://www.mdpi.com/2078-2489/11/4/193 (accessed: 4 April 2023).
37. Cervantes J., Garcia-Lamont F., Rodríguez-Mazahua L., Lopez A. A comprehensive survey on support vector machine classification: applications, challenges and trends. *Neurocomputing*, 2020, vol. 408, pp. 189–215. DOI: 10.1016/j.neucom.2019.10.118
38. Boateng E., Otoo J., Abaye D. Basic tenets of classification algorithms K-Nearest-Neighbor, Support Vector Machine, Random Forest and Neural Network: A Review. *Journal of Data Analysis and Information Processing*, 2020, vol. 8, pp. 341–357. DOI: 10.4236/jdaip.2020.84020
39. *Scikit-learn machine learning in Python*. Available at: https://scikit-learn.org/stable/index.html (accessed: 4 April 2023).
40. Hossin M., Sulaiman M.N. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 2015, vol. 5, no. 2, pp. 01–11. DOI: 10.5121/ijdkp.2015.5201
41. *«Vinnyy gid Rossii-2022»: nazvany luchshie krasnye i belye vina strany* [«Wine Guide of Russia-2022»: the best red and white wines of the country are named]. Available at: https://roskachestvo.gov.ru/news/vinnyy-gid-rossii-2022-nazvany-luchshie-krasnye-i-belye-vina-strany/ (accessed: 4 April 2023).

Received: 10 April 2023.
Reviewed: 17 June 2023.

Information about the authors

Vladimir S. Repkin, technician, Tomsk State University of Control Systems and Radioelectronics.

Artemy V. Li, technician, Tomsk State University of Control Systems and Radioelectronics.

Grigory Yu. Semenov, technician, Tomsk State University of Control Systems and Radioelectronics.

Nikita I. Sermavkin, technician, Tomsk State University of Control Systems and Radioelectronics.

Alexander S. Kovalenko, technician, Tomsk State University of Control Systems and Radioelectronics.

Nikolai S. Egoshin, Cand. Sc., associate professor, Tomsk State University of Control Systems and Radioelectronics.