УДК: 634.8 DOI 10.30679/2587-9847-2019-24-106-116

ПРОГРАММНЫЙ ПАКЕТ «ORANGE» И ПРЕИМУЩЕСТВА ЕГО ПРИМЕНЕНИЯ В БИОИНФОРМАТИКЕ (КЕЙС ПО ДАННЫМ О СОРТАХ ВИНА И МИКРОГРАФИЧЕСКИМ СНИМКАМ ЛОКАЛИЗАЦИИ ДРОЖЖЕВОГО БЕЛКА)

Малышенко К.А., *канд. экон. наук,* **Малышенко В.А.,** *канд. экон. наук*

Федеральное государственное бюджетное учреждение науки «Всероссийский национальный научно-исследовательский институт виноградарства и виноделия «Магарач» РАН» (Ялта)

Анашкина М.В., *аспирант*

Гуманитарно-педагогическая академия (филиал) ФГАОУ ВО «КФУ им. В.И. Вернадского» (Ялта)

Реферат. В представленной работе рассмотрены возможности применения программного продукта с открытым кодом «Orang» для реализации комплекса операций исследовательского и классификационного характера на основе данных о химическом составе вина (условный пример по данным из библиотеки баз программного продукта). Преимуществом выступает системный функционал программы, обеспечивающий достаточно широкий спектр статистических исследований по анализу графических объектов (например, типичных материалов биологического исследования – фотографий дрожжей), и работа со специальными файлами баз данных геномов биологических объектов. Методы исследований – «Дерево решений», «Логистическая регрессия» и другие, используемые для классификации объектов.

Ключевые слова: виджет, Data mining, Big Data, биоинформатика, обобщенное программирование, визуальное программирование

Summary. In the present work, the progressive possibilities of using the open source software “Orang” for the implementation of a complex of research and classification operations based on data on the chemical composition of wine (a conditional example from library data of software databases) are considered. The advantage is the system functionality of the program, which provides a fairly wide range of statistical studies on the analysis of graphical objects (for example, typical materials of biological research – photos of yeast) and work with special files of databases of genomes of biological.

Key words: widget, Data mining, Big Data, bioinformatics, generalized programming, visual programming.

Введение. Развитие информационного общества порождает значительные объёмы информации, влияние которых на различные социально-экономические институты ещё предстоит оценить исследователям. Определённая работа ведётся по многим направлениям, одним из которых является дата-майнинг как совокупность инструментов по анализу неструктурированных данных различного содержания.

Целью данного исследования является определение возможностей нового программного продукта «Orange» для реализации целей биоинформатики как базиса для развития целого спектра информационно-аналитических методов междисциплинарных исследований.

Обсуждение. Разработчиком данной программы является старейший и крупнейший в Словении Университет Любляны. Первая версия программы была представлена в 1996 году. В настоящий момент продукт имеет версию 3.22.0 (от 26 июня 2019 года), в котором

значительно расширен функционал по импорту файлов, добавлены новые инструменты (так называемые виджеты) и т.п. «Orange» реализуется как система всесторонней визуализации данных, которая помогает выявлять скрытые шаблоны, обеспечивает интуитивное понимание процедур анализа информации или поддерживает связь между подсистемами в предметной области. Визуализации включают в себя такие методы, как график рассеяния, блочную диаграмму и гистограмму, а также визуализации для конкретной модели, такие как дендрограмма, силуэтная графика и древовидная визуализация, и это лишь некоторые из них. Многие другие визуализации доступны в виде дополнений и включают визуализации сетей, облаков слов, географических карт и многое другое. Визуализация Orange является интерактивной – можно выбирать точки данных на точечной диаграмме, узлы в дереве, ветви в дендрограмме [1].

Прежде, чем перейти непосредственно к описанию возможностей данной программы, необходимо в целом определиться с понятием такой новой науки, находящейся на пересечении двух областей знаний, как биоинформатика.

Биоинформатика — совокупность методов и подходов, включающих в себя [2]:

– математические методы компьютерного анализа в сравнительной геномике (геномная биоинформатика);

– разработку алгоритмов и программ для предсказания пространственной структуры биополимеров (структурная биоинформатика);

– исследование стратегий соответствующих вычислительных методологий, а также общее управление информационной сложностью биологических систем [3].

В данной работе сделан акцент на третьем направлении.

В биоинформатике используются методы прикладной математики, статистики и информатики. Ее методы используется в биохимии, биофизике и в других областях [4].

Российские ученые Н.Н. Назипова, Е.А. Исаев, и др. определяют современную биоинформатику как науку, которая занимается «… развитием и использованием компьютерных методов для анализа разнообразных геномных данных. Огромную роль в развитии биоинформатики сыграло стремительное развитие компьютерной техники и вычислительных методов обработки данных, появление современных телекоммуникационных технологий. Биоинформатика является одной из тех областей науки, которые в большей степени зависимы от Интернета и успешно могут развиваться благодаря Интернету…» [5].

Объекты и методы исследований. Orange – это набор инструментов для визуализации данных, машинного обучения и интеллектуального анализа данных с открытым исходным кодом. Он имеет интерфейс визуального программирования для исследовательского анализа данных и их интерактивной визуализации, а также может использоваться в качестве библиотеки Python [6]. Orange – это основанный на компонентах пакет программного обеспечения для визуального программирования для визуализации данных, машинного обучения, интеллектуального анализа данных и анализа данных [7].

Orange-компоненты называются виджетами – это элементы графического интерфейса пользователя, имеющего стандартный внешний вид и выполняющего стандартные действия [8]. Такие действия варьируются от простой визуализации данных, выбора подмножеств и предварительной обработки до эмпирической оценки алгоритмов обучения и прогнозного моделирования.

Визуальное программирование реализуется через интерфейс, в котором рабочие процессы создаются путём связывания виджетов (функционального и графического – плавными линиями на мониторе внутри рабочего стола программы), предварительно определённых или разработанных их пользователем (баз данных и систем их обработки), опытные пользователи могут использовать Orange в качестве библиотеки Python для манипулирования данными и изменения виджетов (коррекции их функционала) [7].

Установка по умолчанию включает в себя несколько алгоритмов машинного обучения, предварительной обработки и визуализации данных в 6 наборах виджетов (данные, визуализация, классификация, регрессия, оценка и контроль). Дополнительные функции доступны в виде дополнений (биоинформатика, объединение данных и анализ текста) [6].

Orange поддерживается в операционных системах macOS, Windows и Linux, а также может быть установлен из репозитория Python Package Index (pip install Orange3). По состоянию на май 2018 года стабильная версия 3.13 работает на Python 3, а устаревшая версия 2.7 – на Python 2.7 (все еще доступна) [9].

Структурно Orange состоит из интерфейса Canvas, на который пользователь помещает виджеты и создает рабочий процесс анализа данных. Виджеты предлагают базовые функции, такие как чтение данных, отображение таблицы данных, выбор функций, предикторы обучения, сравнение алгоритмов обучения, визуализация элементов данных и т. д. Пользователь может интерактивно исследовать визуализации или передавать выбранное подмножество в другие виджеты [6].

Виджеты по типам подразделяются на:

- «Данные»: виджеты для ввода данных, фильтрации данных, выборки, вменения, управления функциями и выбора функций;
- «Визуализация»: виджеты для общей визуализации (блочная диаграмма, гистограммы, точечная диаграмма) и многомерной визуализации (мозаичное отображение, ситовая диаграмма) ;
- «Классификация»: набор контролируемых алгоритмов машинного обучения для классификации;
- «Регрессия»: набор контролируемых алгоритмов машинного обучения для регрессии;
- «Оценка»: перекрестная проверка, процедуры на основе выборки, оценка надежности и оценка методов прогнозирования;
- «Без надзора»: неконтролируемые алгоритмы обучения для кластеризации (k-средних, иерархическая кластеризация) и методы проецирования данных (многомерное масштабирование, анализ главных компонентов, анализ соответствия) ;
- «Ассоциированные»: виджеты для горнодобывающей промышленности часто встречающихся наборов и обучения ассоциативных правил;
- «Биоинформатика»: виджеты для анализа и доступа к библиотекам путей;
- «Слияние данных»: виджеты для объединения различных наборов данных, факторизации коллективной матрицы и исследования скрытых факторов;
- «Образовательные»: виджеты для обучения понятиям машинного обучения, таким как кластеризация k-средних, полиномиальная регрессия, стохастический градиентный спуск;
- «Geo»: виджеты для работы с геопространственными данными;
- «Аналитика изображений»: виджеты для работы с изображениями;
- «Сеть»: виджеты для графа и анализа сети;
- «Текстовый майнинг»: виджеты для обработки естественного языка и текстового майнинга [10];
- «Временные ряды»: виджеты для анализа и моделирования временных рядов;
- «Спектроскопия»: виджеты для анализа и визуализации (гипер) спектральных наборов данных [11];
- «Виджет рисования» данных в сочетании с иерархической кластеризацией и k-средними.

Программа обеспечивает платформу для выбора экспериментов, систем рекомендаций и прогнозного моделирования и используется в биомедицине, биоинформатике, геномных исследованиях и обучении. Чаще всего в науке она используется в качестве плат-

формы для тестирования новых алгоритмов машинного обучения и для внедрения новых методов в генетике и биоинформатике. В образовании ее использовали для освоения методов машинного обучения и интеллектуального анализа данных студентам-биологам, биомедицине и информатике [6].

Краткий обзор программы не позволяет оценить работу данной программы, поэтому перейдем к практической части нашего исследования. Для этого, после того как программа установлена (весь процесс занимает относительно небольшое время, сначала необходимо установить базовую программу «Python» для первоначальной версии и загрузить дополнительные утилиты, например, «Биоинформатика»), запускается программа и создаётся файл. Необходимо отметить наличие большого количества шаблонов, предлагаемых данной программой, а также поддержку в виде организованной на «Youtube» канале подборки видеороликов с подробной презентацией работы различных модулей программы (на английском языке).

Задача тестирования: определить по химическим показателям сорт вина. База исследования взята из библиотеки Orange: база «Wine» (1992), от UCI ML Repository. Это данные о винах, выращенных в Италии, но полученных из трех разных сортов. Вина профилированы с помощью химического анализа, который сообщает о наличии ограниченного количества (тринадцати) составляющих-параметров, включая алкоголь, яблочную кислоту и флаваноиды. (Источник: владельцы: Форина, М. и др., Институт фармацевтического и пищевого анализа и технологий, Via Brigata Salerno, 16147 Генуя, Италия). Информация о наборе данных: эти данные являются результатами химического анализа вин, выращенных в Италии, но полученных из трех разных сортов. Анализ определил количество 13 компонентов, найденных в каждом из трех типов вин.

Атрибуты (предоставлены Риккардо Лирди, riclea@anchem.unige.it):

1) алкоголь; 2) яблочная кислота; 3) зола; 4) щелочность золы; 5) магнезия; 6) всего фенолов; 7) флаваноиды; 8) нефлаваноидные фенолы; 9) проантоцианины; 10) интенсивность цвета; 11) оттенок; 12) OD280 / OD315 разбавленных вин; 13) пролин.

Информация об атрибутах: все атрибуты являются непрерывными. Рабочий стол исследований с потоками и виджетами для данных условий может иметь вид представленный на рисунке 1.

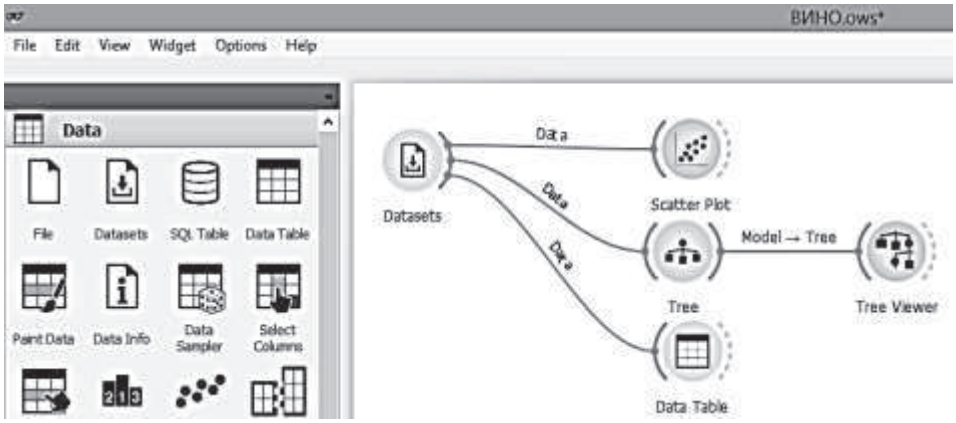


Рис. 1. Рабочий стол исследований с потоками и виджетами в «Оранж» (задача)

Как видно из рисунка 1, из набора данных был выбран файл базы исследования «Wine», который был связан нами с несколькими виджетами: Scatter Plot – визуализация «графика рассеяния» с расширенным анализом и интеллектуальными улучшениями визуализации данных (рис. 2), что позволило сделать визуальный анализ по магнезии и флава-

ноидам, вместе с тем график можно видоизменять выбором других переменных из нашего набора. Из данного графика можно определить 3 группы объектов со схожими признаками (выделены цветом).

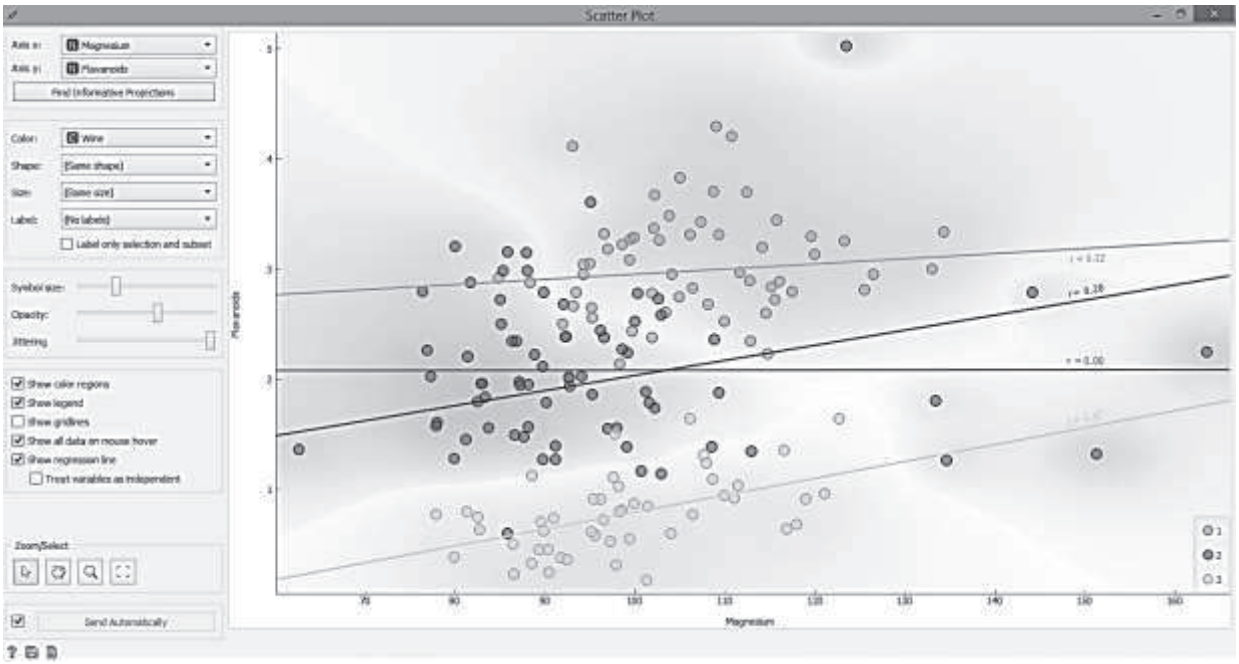


Рис. 2. График (диаграмма) рассеяния по двум признакам (переменным) – магнезии и флаваноидам (по базе «Wine»)

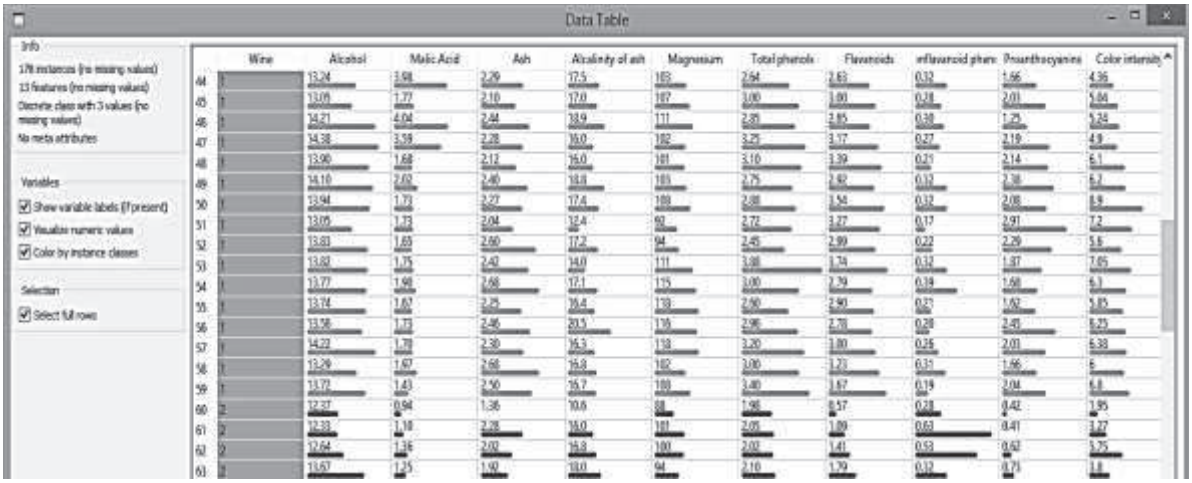


Рис. 3. Таблица данных 3 различных вин с 13 атрибутами

1. Таблица данных (Data Table) – виджет принимает один или несколько наборов данных из его входных данных (например, файлов других приложений) и представляет их в виде таблицы. Экземпляры данных могут быть отсортированы по значениям атрибутов. Виджет также поддерживает ручной выбор экземпляров данных (рис. 3).

2. «Дерево» (виджет «Tree») – это простой алгоритм, который разбивает данные на узлы по «чистоте» класса. Это предшественник «случайного леса». Алгоритм «Дерево» в

Оранж разработан собственными силами и может работать как с дискретными, так и с непрерывными типами данных. Визуализируется его работа с помощью следующего виджета (виджет «Tree Viever»).

3. Визуализация «деревьев» (виджет «Tree Viever») классификации и регрессии («дерево решений») – это универсальный виджет с двумерной визуализацией «деревьев» классификации и регрессии. Пользователь может выбрать узел (связь между разными компонентами, например, выбор данных из набора), инструктируя виджет выводить данные, связанные с этим узлом, что позволяет проводить исследовательский анализ данных (рис. 4). Как видно из рисунка 4, для целей прогнозирования этих данных будет недостаточно – атрибуты не привязаны к конкретному сорту вина.

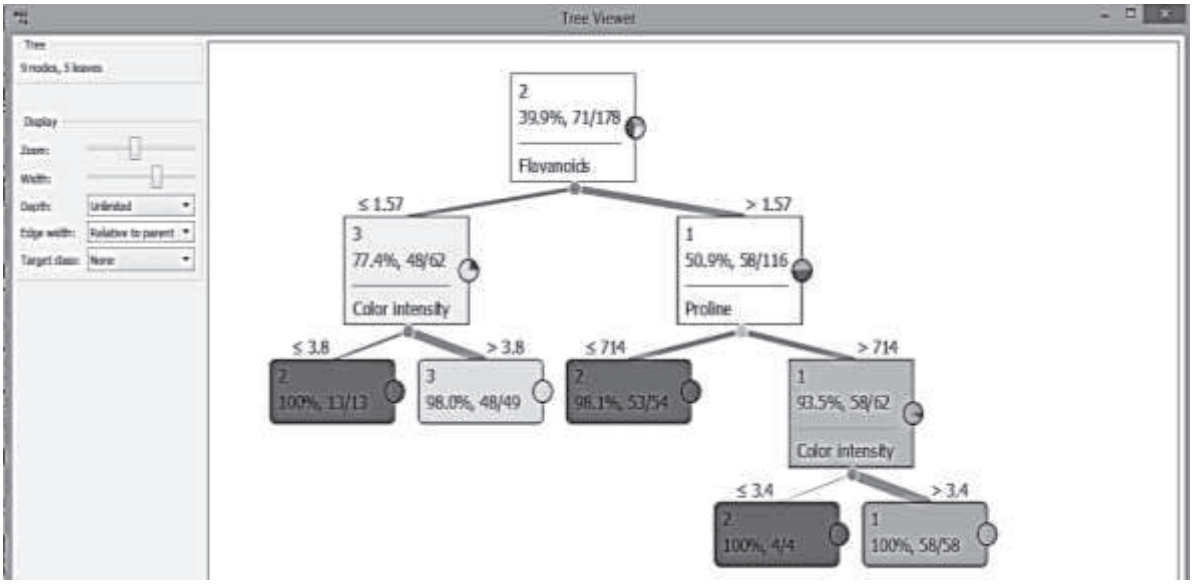
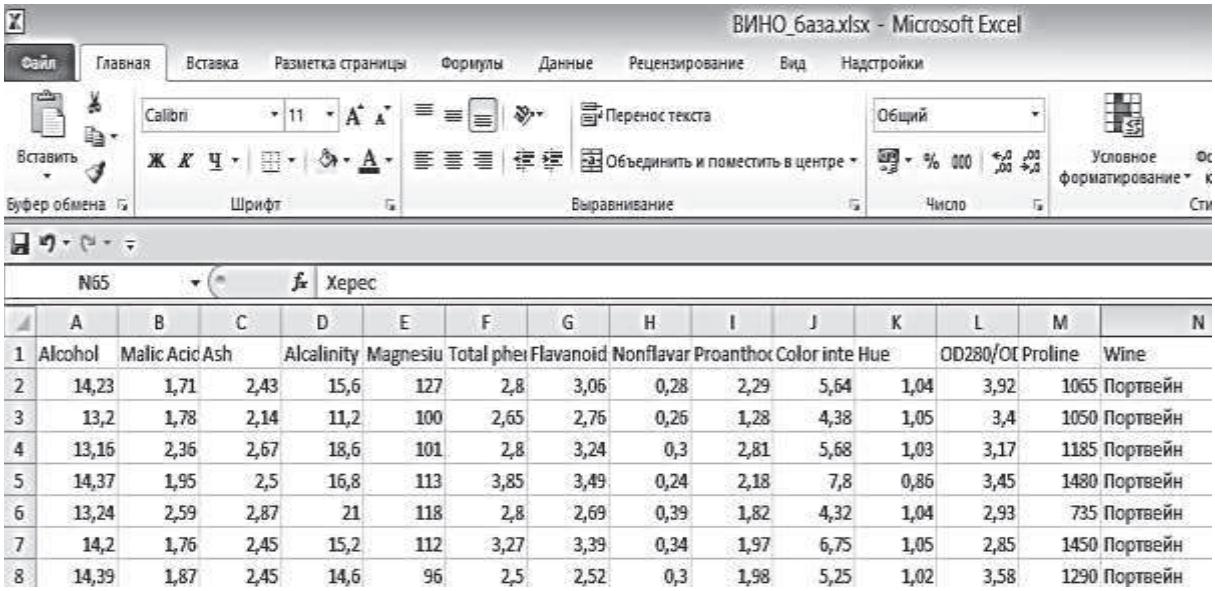


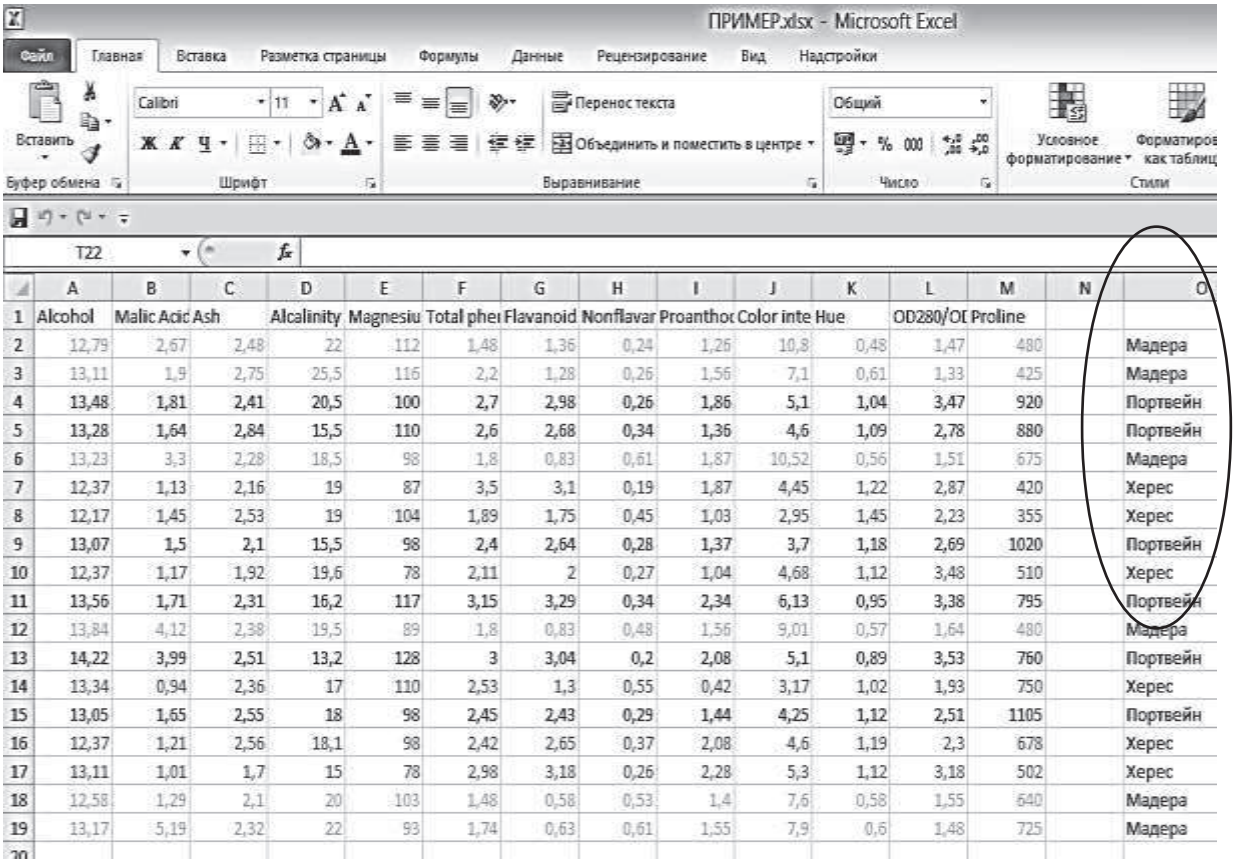
Рис. 4. Визуализация «деревьев» (виджет “Tree Viever”)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Alcohol	Malic Acid	Ash	Alcalinity	Magnesiu	Total phen	Flavanoid	Nonflavar	Proanthoc	Color inte	Hue	OD280/OD	Proline	Wine
2	14,23	1,71	2,43	15,6	127	2,8	3,06	0,28	2,29	5,64	1,04	3,92	1065	Портвейн
3	13,2	1,78	2,14	11,2	100	2,65	2,76	0,26	1,28	4,38	1,05	3,4	1050	Портвейн
4	13,16	2,36	2,67	18,6	101	2,8	3,24	0,3	2,81	5,68	1,03	3,17	1185	Портвейн
5	14,37	1,95	2,5	16,8	113	3,85	3,49	0,24	2,18	7,8	0,86	3,45	1480	Портвейн
6	13,24	2,59	2,87	21	118	2,8	2,69	0,39	1,82	4,32	1,04	2,93	735	Портвейн
7	14,2	1,76	2,45	15,2	112	3,27	3,39	0,34	1,97	6,75	1,05	2,85	1450	Портвейн
8	14,39	1,87	2,45	14,6	96	2,5	2,52	0,3	1,98	5,25	1,02	3,58	1290	Портвейн

Рис. 5. База «Wine» с новым атрибутом – сорт вина («Портвейн», «Херес» «Мадера»)

Для устранения неопределенности идентификации необходимо создать обучающую выборку на основе заведомо установленных (определенных и привязанных) сортов винограда и их биохимических составов. Для этого три имеющиеся группы исследований, разбитых автоматически в исходном файле примера, условно (гипотетически, как учебный кейс) определим следующим образом: 1 группа – «Портвейн», 2 группа – «Херес» и 3 группа – «Мадера». Для этого создадим файл базы в Excel с новым атрибутом «Wine» – сорт вина, являющийся теперь целевой переменной строчного типа (рис. 5), который будем использовать в качестве обучающей выборки. Кроме этого, создадим второй контрольный файл (с «неизвестными» сортами вина), тоже в Excel на котором будем проводить «опыт» (рис. 6).



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Alcohol	Malic Acid	Ash	Alcalinity	Magnesiu	Total phe	Flavanoid	Nonflavar	Proanthoc	Color inte	Hue	OD280/OD	Proline		
2	12,79	2,67	2,48	22	112	1,48	1,36	0,24	1,26	10,8	0,48	1,47	480		Мадера
3	13,11	1,9	2,75	25,5	116	2,2	1,28	0,26	1,56	7,1	0,61	1,33	425		Мадера
4	13,48	1,81	2,41	20,5	100	2,7	2,98	0,26	1,86	5,1	1,04	3,47	920		Портвейн
5	13,28	1,64	2,84	15,5	110	2,6	2,68	0,34	1,36	4,6	1,09	2,78	880		Портвейн
6	13,23	3,3	2,28	18,5	98	1,8	0,83	0,61	1,87	10,52	0,56	1,51	675		Мадера
7	12,37	1,13	2,16	19	87	3,5	3,1	0,19	1,87	4,45	1,22	2,87	420		Херес
8	12,17	1,45	2,53	19	104	1,89	1,75	0,45	1,03	2,95	1,45	2,23	355		Херес
9	13,07	1,5	2,1	15,5	98	2,4	2,64	0,28	1,37	3,7	1,18	2,69	1020		Портвейн
10	12,37	1,17	1,92	19,6	78	2,11	2	0,27	1,04	4,68	1,12	3,48	510		Херес
11	13,56	1,71	2,31	16,2	117	3,15	3,29	0,34	2,34	6,13	0,95	3,38	795		Портвейн
12	13,84	4,12	2,38	19,5	89	1,8	0,83	0,48	1,56	9,01	0,57	1,64	480		Мадера
13	14,22	3,99	2,51	13,2	128	3	3,04	0,2	2,08	5,1	0,89	3,53	760		Портвейн
14	13,34	0,94	2,36	17	110	2,53	1,3	0,55	0,42	3,17	1,02	1,93	750		Херес
15	13,05	1,65	2,55	18	98	2,45	2,43	0,29	1,44	4,25	1,12	2,51	1105		Портвейн
16	12,37	1,21	2,56	18,1	98	2,42	2,65	0,37	2,08	4,6	1,19	2,3	678		Херес
17	13,11	1,01	1,7	15	78	2,98	3,18	0,26	2,28	5,3	1,12	3,18	502		Херес
18	12,58	1,29	2,1	20	103	1,48	0,58	0,53	1,4	7,6	0,58	1,55	640		Мадера
19	13,17	5,19	2,32	22	93	1,74	0,63	0,61	1,55	7,9	0,6	1,48	725		Мадера

Рис. 6. Контрольная база для проверки работы программы
(маркеры (овал) вин расположены справа в таблице и добавлены позднее)

Сформируем новый рабочий поток, где кроме уже известных виджетов «Tree» и «Tree Viever» будет новый виджет – «Логистическая регрессия» («Logistic Regression») – алгоритм классификации логистической регрессии и виджет «Предсказания» («Predictions»), который показывает прогнозы моделей на данных из обучающего файла.

Виджет получает набор данных и один или несколько «предикторов» (прогностические модели, а не алгоритмы обучения). Он выводит данные и прогнозы. То есть, он позволит нам на основе данных алгоритма «Дерево» и/или «Логистической регрессии» (как альтернативный вариант модели прогноза) предсказать по имеющимся данным, к какому вину относится объект исследования в зависимости от предикторов (в нашем случае наборов биохимических параметров) (рис. 7).

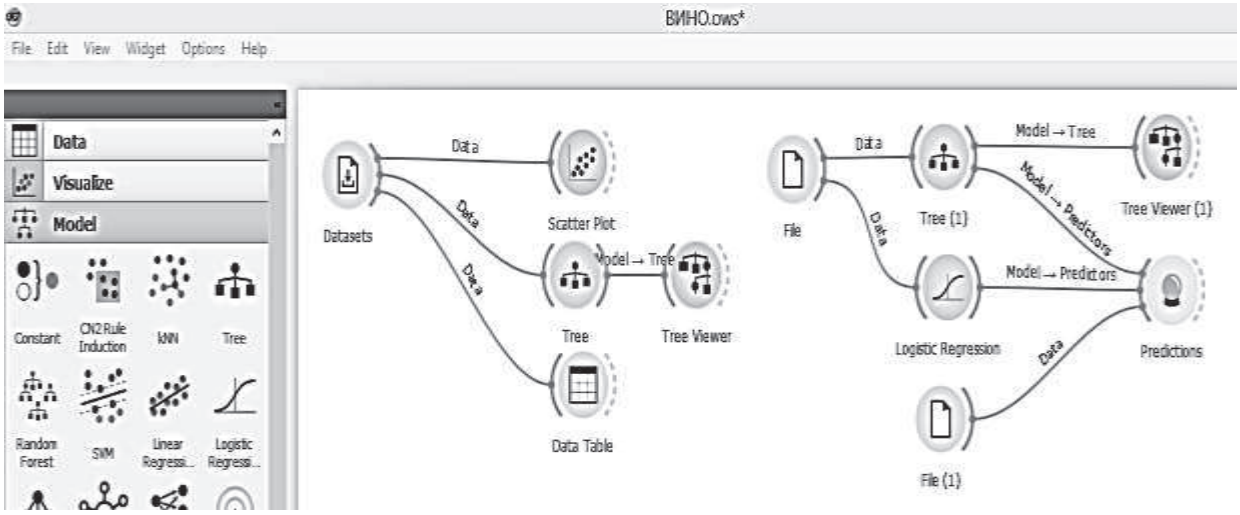


Рис. 7. Новый «рабочий поток» (правая часть рисунка)

Сделаем новую визуализацию алгоритма «дерево» (рис. 8), из которого виден процесс классификации по имеющимся химическим переменным, присущим (искусственно нами привязанные) тому или иному сорту и далее запустим процесс предвидения – классификации «неизвестных» сортов вина, имеющих те или иные биохимические характеристики из контрольного файла.

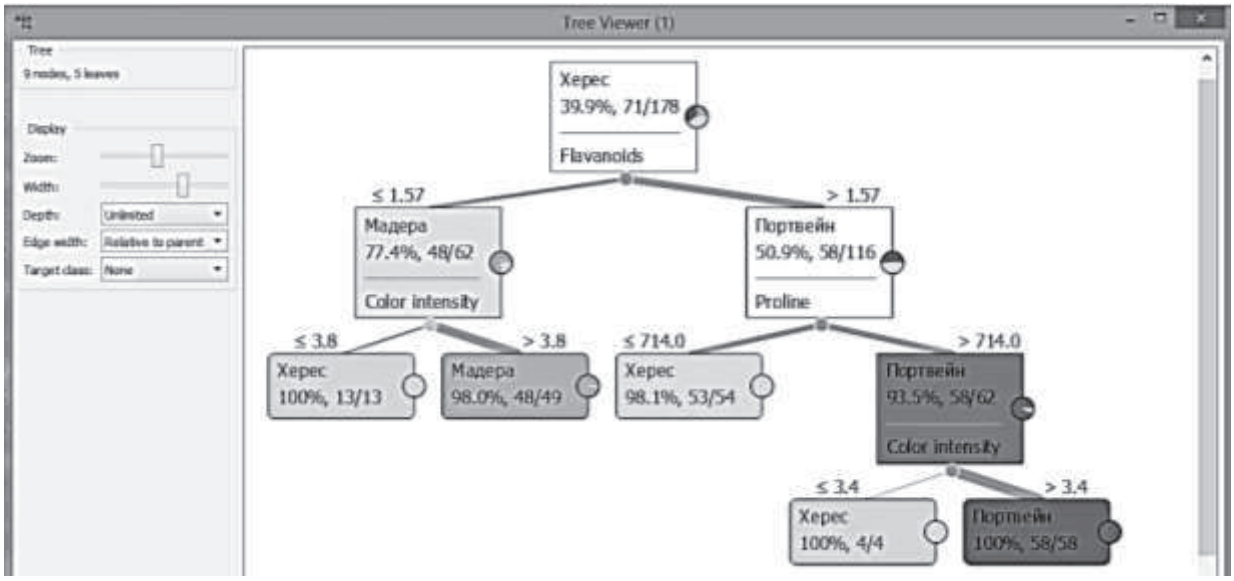


Рис. 8. Визуализация «деревьев» (виджет «Tree Viever») с новыми придикторами (в нашем случае химические параметры вина)

Обсуждение результатов. В результате имеем таблицу вида (рис. 9), где представлены результаты работы программы. Как видно, алгоритм «Дерево решений» по сравнению с Логистической регрессией более точен, что подтверждает реальную работу именно

алгоритмов классификации, а не простое сравнение с исходной базой. Имеется в виду, что в случае получения новых реальных эмпирических данных о новых сортах предложенный метод будет также работать. Единственное требование: должны быть чёткие оценки сортов вина, связанные с переменными, используемые в качестве предикторов (сделанные в единой методике). Естественно, с набором большего объёма эмпирических данных по реальным сортам, точность оценки будет повышаться, и сам прогноз будет универсальным (а не только по трём «искусственно» установленным нами сортам).

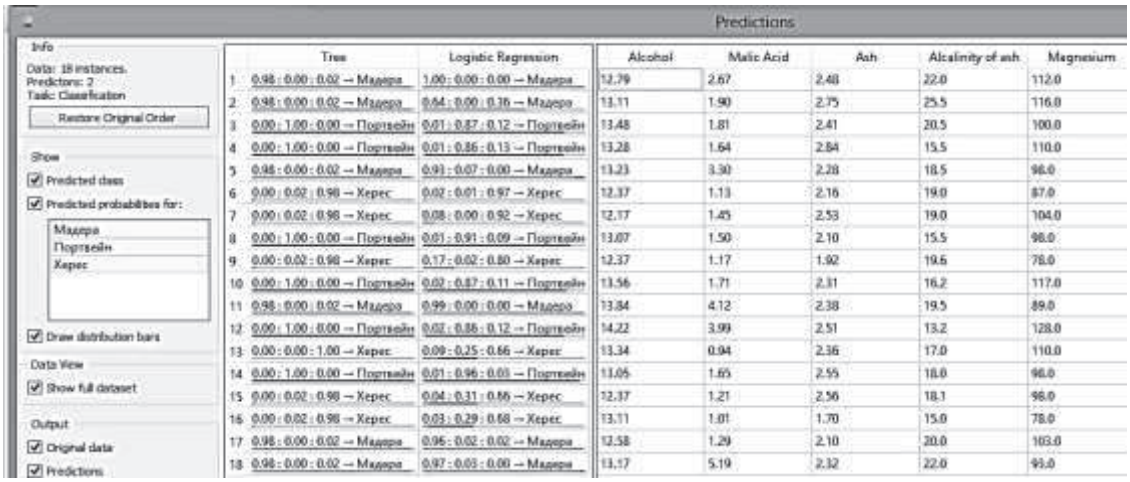


Рис. 9. Результаты работы виджета «Предвиденье»
(классификация сорта вина на основе двух альтернативных алгоритмов)

Сравним полученные результаты с нашими данными по контрольному файлу (см. рис. 6). Как видно их таблицы Excel, куда позднее были добавлено указание сорта вина для последующего контроля, программа справилась по использованным алгоритмам на сто процентов, ошибок нет. Конечно, это идеальные условия, в реальности ошибки возможны, что связано с работой программы по определенному алгоритму и обучающей выборке, но при увеличении эмпирических данных ошибки будут уменьшаться.

Кратко остановимся также на возможностях данной программы по анализу графических объектов. Аналогично сформируем новый рабочий поток. Для этого из встроенных наборов загрузим комплексную базу по локализации дрожжевого белка, имеющую как биологические характеристики, так и графические объекты (рис. 10).

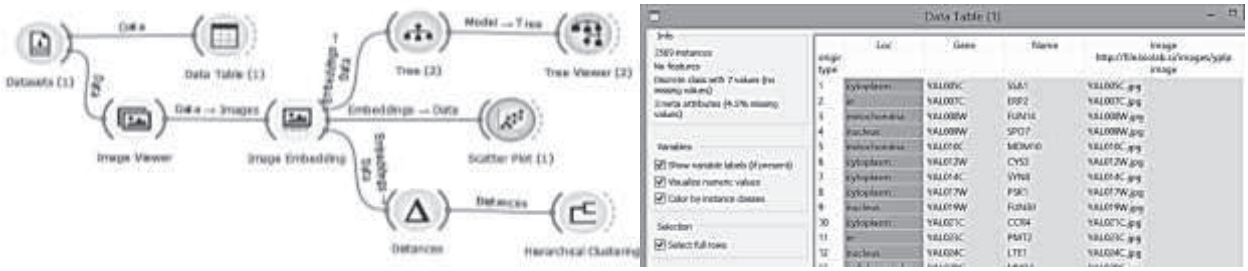


Рис. 10. «Рабочий поток» и характеристика объекта – биологические характеристики и микрографические снимки локализации дрожжевого белка (правая часть рисунка)

Из рисунка 10 следует, что нами были задействованы новые виджеты: «Image Viewer» («Просмотр рисунков») который отображает изображения из набора данных, хранящихся локально или в Интернете. Виджет будет искать атрибут с *type = image* в третьей

строке заголовка. Он может быть использован для сравнения изображений при поиске сходств или расхождений между выбранными экземплярами данных (например, рост бактерий). К этим данным может также применятся ряд инструментов, которые для краткости охарактеризуем «Дерево решений» по осуществлению классификации и проведем кластерный анализ.

В результате использования первого инструмента получим следующую картину (рис. 11). При осуществлении классификации ограничились только 5 уровнем (дерева, левая часть рис. 11), что сделано для повышения наглядности, все уровни по 2569 рисункам просто не поместятся. Для сравнения построим диаграмму рассеяния, где также заметны выделенные нами группы в зависимости от выбранных исследователем вариантов соотношения биологических показателей (красные, жёлтые и зелёные точки).

Рис. 11. Визуализация «деревьев» (виджет «Tree Viever», левая часть рисунка)
и диаграмма рассеяния (виджет «Scatter Plot», правая часть рисунка)

Кластерный анализ также возможно осуществлять по графическим объектам, однако необходимо сделать расчёт дистанций – виджет «Distances» (рис. 13).

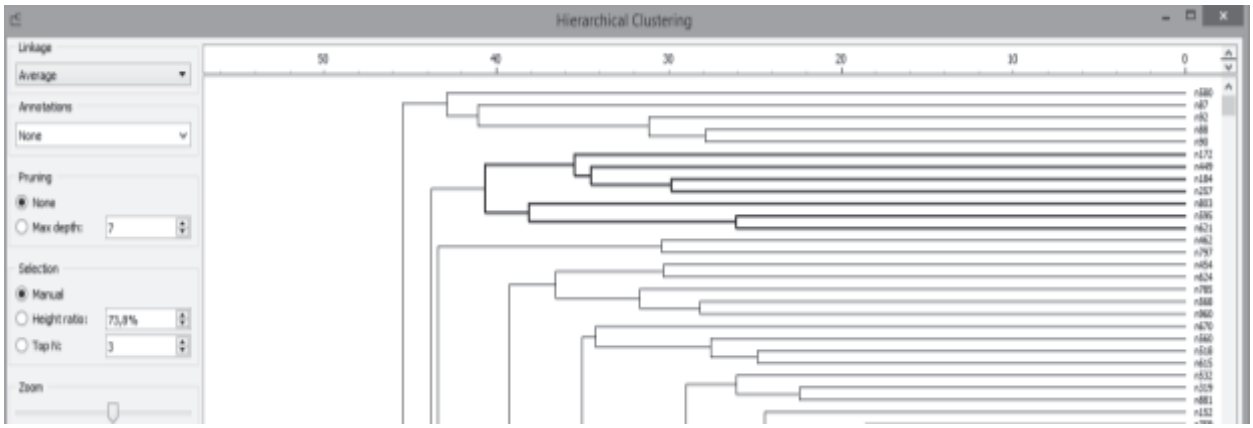


Рис. 13. Кластерный анализ визуальных объектов

Сжато охарактеризуем раздел биоинформатики в Orange. Orange Bioinformatics расширяет Orange – пакет программного обеспечения для интеллектуального анализа данных, с общими функциями для биоинформатики, предоставляет доступ к общедоступным данным, таким как наборы данных GEO, Biomart, GO, KEGG, Atlas, ArrayExpress и база данных PIPAx, что позволяет сделать отбор генов, контроль качества, оценку расстояний между экспериментами с несколькими факторами и др. Все функции могут быть объединены в единую систему с мощными методами визуализации, исследования сети и интеллектуального анализа данных (из среды интеллектуального анализа данных Orange) [9]. Данное направление будет объектом дальнейшего исследования нашего коллектива, результаты которого представим на следующей конференции.

Научная новизна исследования состоит в обосновании нового подхода ситуационной коррекции методов анализа биохимических объектов в диверсифицированных условиях необходимой гибкости междисциплинарных исследований. Новый метод заключается в создании сложных системных конструкций анализа разнородных объектов, полученных на основе слияния визуальных объектов, объектов специальных баз данных биохимических исследований и традиционных методов статистики, реализуемых в среде компьютерной обработки информации уровня «Биг Дата» и специальных программ (типа Orange).

Заключение. Применение инструментов Orange сделало возможным проведение оценок по целевой выборке в автоматическом режиме с использованием соответствующего программного продукта. В Orange аналитик объединяет основные вычислительные единицы, называемые виджетами, в схемы аналитики потока данных. Два модуля-виджета могут быть связаны, если они совместно используют тип данных. По сравнению с другими популярными инструментами виджеты Orange представляют собой высокоуровневые, интегрированные потенциально сложные задачи, но достаточно специфичны для самостоятельного использования. Даже сложные конструкции анализа редко состоят из десяти и более виджетов, а такие задачи, как кластеризация, могут выполняться с использованием до пяти виджетов.

При построении схемы каждый виджет контролируется независимо от настроек, но настройки концептуально не обременяют аналитика. Полученные результаты и предложенный подход при соответствующих доработках можно использовать при решении научной задачи – классификации различных сортов вина, винных продуктов и других объектов виноградарства и виноделия.

Литература

1. Интерактивная визуализация данных [Электронный ресурс]. Режим доступа: https://orange.biolab. si/home/interactive datavisualization/ 2019.

2. Кунин Е. Суп из гвоздя. Ведущий эволюционист рассказал о Мультивселенной и антропном принципе. // Lenta.ru, 1 декабря 2012 [Электронный ресурс]. Режим доступа: https://lenta.ru/articles/2012/11/30/koonin/

3. Ivan Y. Torshin Bioinformatics in the Post-Genomic Era: The Role of Biophysics, Novapublishers, (2006), ISBN 1-60021-048-1.

4. Hogeweg P. The roots of bioinformatics in theoretical biology. (англ.) // Public Library of Science for Computational Biology. 2011. Vol. 7, no. 3. P. e1002021. DOI:10.1371/journal.pcbi.1002021.)

5. Назипова Н.Н., и др. Большие данные в биоинформатике / Н.Н. Назипова [и др.]. // Математическая биология и биоинформатика, 2017. Т. 12. № 1. С. 102-119. DOI: 10.17537/2017.12.102

6. Orange (программное обеспечение) [Электронный ресурс]. Режим доступа: https://en.wikipedia.org/wiki/ Orange (software). 2019.

7. Orange: набор инструментов для интеллектуального анализа данных в Python (PDF) / Демшар Я. Керк Т. Ерявец А. [и др.]. (2013). JMLR. 14 (1): 2349-2353.

8. Буторин, Денис Н. MS Agent и Speech API в Delphi: производственно-практическое издание. Санкт-Петербург : БХВ-Петербург, 2005. 448 с. : ил. (Профессиональное программирование). ISBN 5-94157-502-5 : 191.44 р.

9. Последняя версия Orange для Windows [Электронный ресурс]. Режим доступа: https://orange.biolab.si/download/#windows – 2019.

10. Malyshenko, K.A., Malyshenko, V.A. & Anashkina, M.V. (2018). Methodical Approaches to Forecasting Dynamics of the Stock Market based on "Text-Mining". Proceedings of the 32nd International Business Information Management Association Conference - Vision 2020: Sustainable Economic Development and Application of Innovation Management from Regional expansion to Global Growth, 15-16 November 2018, Seville, Spain. pp. 1030-1044.

11. Новости синхротронного излучения / Топлак М., Бирарда Г., Рид С. [и др.], № 30, р. 40-45 (2017). https://doi.org/10.1080/08940886.2017.1338424