Санкт-Петербургский Государственный Университет

Математическое обеспечение и администрирование
информационных систем
Кафедра информационно-аналитических систем

Пиккио Полина Феличе

Применение методов машинного обучения для классификации и анализа геологических артефактов

Бакалаврская работа

Научный руководитель:
к.ф.-м.н, доцент Графеева Н. Г.

Рецензент:
младший программист первой категории в
ООО “Пегасис”
Железняков И. Э.

Санкт-Петербург
2021

SAINT-PETERSBURG STATE UNIVERSITY

Software and Administration of Information Systems
Information Systems and Data Bases

Pikkio Polina

Application of machine learning methods for classification and clustering of geological artifacts

Bachelor's Thesis

Scientific supervisor:
Ph.D., associate professor Grafeeva Natalia

Reviewer:
Junior developer in 'Pegasys'
Zheleznyakov Ivan

Saint-Petersburg
2021

Оглавление

Введение **4**

Постановка задачи **5**

Обзор предметной области **6**

Методология **10**

Логистическая регрессия 10

Линейный дискриминантный анализ 10

Метод k-ближайших соседей 10

Сравнение алгоритмов 11

Исходные данные **12**

Реализация приложения **14**

Инструкция для пользователя **15**

Применение приложения к первому датасету **18**

Результаты работы с первым датасетом **20**

Применение приложения ко второму датасету **21**

Результаты работы со вторым датасетом **22**

Заключение **23**

Список литературы **24**

Введение

В настоящее время невозможно представить научные исследования одной области деятельности без участия и поддержки других областей. Так и в нашем случае — данные, полученные и анализируемые геологами, также обрабатываются с помощью компьютерных технологий для уточнения, обобщения и привнесения новых результатов. Исследования, которые геологи проводят с помощью приборов, мы проведем с помощью алгоритмов и кластерного анализа: исследуем полезные схемы группирования объектов, разработаем классификацию, выделим закономерности на основе исследуемых данных.

Мы разработали и согласовали тему совместно с геологическим факультетом Санкт-Петербургского государственного университета: создать приложение, простое и удобное в использовании, позволяющее определить принадлежность образца (см. Рис. 1) той или иной кластерной группе и в конечном итоге — промышленную полезность.

Рис. 1: Исследуемый образец.

Постановка задачи

Целью данной дипломной работы является создание приложения для пользования геологов в самых простейших обстоятельствах со следующей функциональностью: классификацией гранитов на основе химического анализа и определения промышленной ценности образцов. Для реализации цели были поставлены следующие задачи:

- выбор алгоритма
- разработка дизайна приложения
- поиск и обработка исходных данных
- создание приложения
- создание инструкции для пользователей приложения
- внедрение и применение приложения
- определение промышленной ценности образцов

Обзор предметной области

В своих исследованиях геологи часто ставят задачи определения химического состава образцов и его принадлежности тому или иному классу. Благодаря этому они узнают много важной информации: определяют породообразующие минералы образца, делают выводы о принадлежности определенной местности и даже о времени формирования образца. К сожалению, приборы не всегда дают точные результаты: бывают пропущенные значения — они выходят за рамки допустимых пределов обнаружения и прибор не может их определить, бывает, что пропущен важный атрибут образца. Также бывает, что неизвестно, откуда получен тот или иной образец. Мы можем восстанавливать пропущенные значения и определять принадлежность образца тому или иному классу, что требуется для работы геологов — благодаря этому результаты анализа будут более точными, результаты работы приборов будут подтверждены компьютерной программой.

Данные, которые получают и исследуют геологи, представлены на Рис. 2 и Рис. 3.

	Саханай			Дурулгуй		
	Бт	Бт	Бт	Бт	Бт	Му
Химический элемент	C-2798	C-2793	C-2800	Д-2749	D-2749	D-272
F	1500	1500	1300	1200	1100	
Li	50,87	188	188	129,01		
Be	10,08		15,4	8,27		
Sc	1,72			2,72		
Ti	558,21			1002,33		
V	7,99	20,7		10,83	17,4	
Cr	4,69	14,2		5,52	14,3	
Mn	313,28			437,62		
Co	1,39	2,68		1,89	2,54	

Рис. 2: Исходные данные по цирконам.

W%

Название спектр	O	F	Na	Mg	Al	Si	P	S	Cl	K	Ca	Ti	Mn	Fe	Zn
Спектр 1093	31,55	3,81					15,68				35,7		1,18	0,74	
Спектр 1094	42,7					45,58									
Спектр 1095	38,94		7,7		10,07	31,7									
Спектр 1096	38,32				9,62	29,92				13,84					
Спектр 1097	33,58	3,28			13,94	20,76				8,69		0,23	1,55	9,03	
Спектр 1098	36,25	2,52			15,05	21,69				8,99			1,02	5,43	
Спектр 1099	38,14		0,33		9,49	30,1				13,58					
Спектр 1100	32	3,4			1,12	3,68	15,94			0,93	31,97		2,8		
Спектр 1101	28,08	3,29					16,9				33,04		4,82		
Спектр 1102	30,79	0,77					14,35				25,72		1,99		
Спектр 1103	25,71				11,09	17,36				7,17			0,85	5,82	
Спектр 1104	36,25		0,2		9,09	28,53				13,19					
Спектр 1105	36,37		7,07		9,55	29,86									
Спектр 1106	33,62	2,29	0,13		13,82	19,97				8,12			1,01	7,15	

Рис. 3: Представление анализа микрозонда.

Также следует отметить, что все исследования минералов, которые проводят геологи, в конечном итоге направлены на то, чтобы понять, как в реальной жизни можно использовать минералы: делать из них украшения или зубные коронки, изготавливать краску или использовать в электротехнике, строительстве или медицине. Для этого геологи изучают породообразующие минералы образца: сначала смотрят на строение массива, затем на химический состав, потом восстанавливают химическую формулу минерала и после этого понимают, какому классу породообразующих минералов принадлежит образец.

Выделяют три основные группы породообразующих минералов: мусковит, биотит и двуслюдяные. С точки зрения геологии, мусковит — породообразующий минерал из группы слюд подкласса слоистых силикатов, KAl2([AlSi3O10](OH,F)2. Мусковит обладает следующими свойствами: может быть белого, серого, светло-коричневого или зеленоватого цвета (см. Рис. 4). В тонких спайных листах бесцветен, но часто с желтоватым, сероватым, зеленоватым и редко красноватым оттенком. Фуксит ярко-зелёный. Блеск стеклянный, на плоскостях спайности перламутровый и серебристый. Мусковит является отличным изолятором для электрических токов обычного напряжения и обладает достаточно высоким сопротивлением пробою.

Рис. 4: Изображение мусковита.

Биотит — породообразующий минерал из класса водных алюмосиликатов, группа слюд. Химическая формула: K(Mg,Fe)3[AlSi3O10](OH,F)2. Для биотита характерны неметаллический блеск, небольшая твердость (не царапает стекло), черный цвет, весьма совершенная спайность и листоватые, чешуйчатые агрегаты (см. Рис. 5).

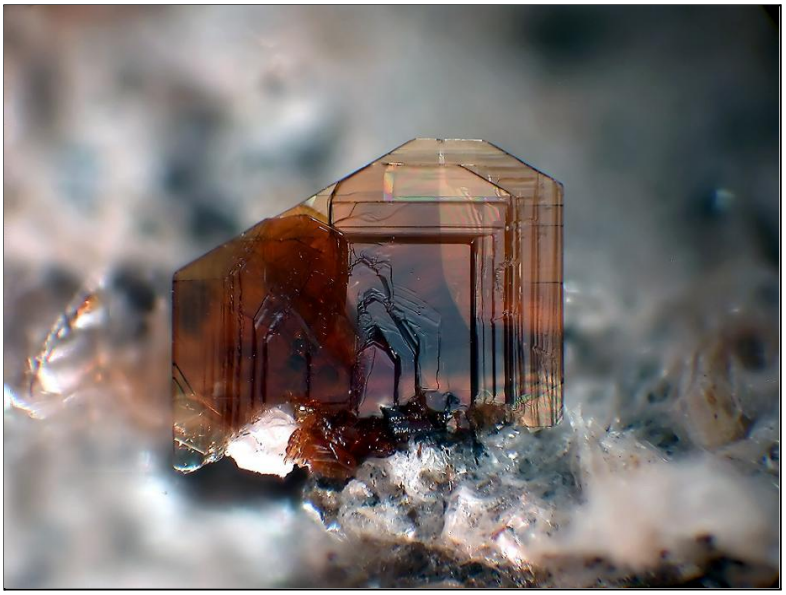


Рис. 5: Изображение биотита.

Рассмотрим кластерную группу “Двуслюдяные”. Двуслюдяной сланец — слюда, представленная мусковитом и биотитом почти в равном соотношении (см. Рис. 6).

Рис. 6: Изображение двуслюдяного сланца.

Определение промышленного назначения образцов с помощью породообразующих минералов — одна из главных задач нашей предметной области.

Методология

Рассмотрим методологию, которая применяется в данной работе. Изучим и сравним три самых популярных алгоритма машинного обучения, которые используются для кластеризации объектов: алгоритм логистической регрессии, линейный дискриминантный алгоритм и алгоритм k-ближайших соседей.

Логистическая регрессия

Логистическая регрессия — алгоритм, пришедший в машинное обучение прямиком из статистики. В этом алгоритме требуется найти значения коэффициентов для входных переменных, выходное значение преобразуется с помощью нелинейной или логистической функции. Его хорошо использовать для задач бинарной классификации (это задачи, в которых на выходе мы получаем один из двух классов).

Линейный дискриминантный анализ

Представление линейного дискриминантного анализа состоит из статистических свойств данных, рассчитанных для каждого класса. Для каждой входной переменной это включает среднее значение для каждого класса и дисперсию, рассчитанную по всем классам. Предсказания производятся путём вычисления дискриминантного значения для каждого класса и выбора класса с наибольшим значением. Предполагается, что данные имеют нормальное распределение, поэтому перед началом работы рекомендуется удалить из данных аномальные значения. Это простой и эффективный алгоритм для задач классификации.

Метод k-ближайших соседей

Теперь рассмотрим наиболее популярный среди своих аналогов метод k-ближайших соседей, который применяется для автоматической

классификации объектов. При использовании метода объект присваивается тому или иному классу. Идея заключается в том, что близким объектам в признаковом пространстве соответствуют похожие метки. Каждый из объектов в задаче классификации представляется в виде вектора в n-мерном пространстве, каждое измерение в котором представляет собой описание одного из признаков объекта. Для реализации метода необходимо выполнить алгоритм, состоящий из трех шагов:

1) Вычислить расстояние от нового объекта до всех объектов обучающей выборки.
2) Отобрать k объектов из обучающей выборки, расстояние до которых минимально.
3) Наш объект принадлежит тому классу, который наиболее часто встречается среди классов k выбранных объектов.

Функция для расчета расстояния должна удовлетворять трем свойствам:

1) $(x,y) \geq 0$, $d(x,y) = 0$ тогда и только тогда, когда $x = y$;
2) $d(x,y) = d(y,x)$;
3) $d(x,z) \leq d(x,y) + d(y,z)$, при условии, что точки x, y, z не лежат на одной прямой.

Где x, y, z — векторы признаков сравниваемых объектов.

Сравнение алгоритмов

Логическая регрессия не соответствует нашим требованиям, поскольку она используется, когда нужно отнести образец к одному из двух классов. При использовании линейного дискриминантного анализа на практике средние и ковариации классов неизвестны. Они могут быть оценены по тренировочному набору, используя либо метод максимального правдоподобия, либо метод оценки апостериорного максимума вместо точного значения в обоих равенствах. Хотя оценки ковариации могут в некотором смысле считаться оптимальными, это не значит, что дискриминант, полученный подстановкой этих значений, оптимален в любом смысле, даже если предположение о нормальном распределении классов верно. Алгоритм k-ближайших соседей же идеально подходит для решения поставленной задачи.

Исходные данные

Для исследования и создания приложения были получены два датасета размерами 26 строк и 1237 строк: данные по цирконам и данные анализа микрозонда одних и тех же местностей. В каждом из датасетов содержится информация о названии образца, идентификаторе образца, процентное содержание химического элемента и атрибуты.

Образцы, с которыми была проведена первая часть работы: цирконы — минералы, извлеченные из гранитов различных горных массивов. Цирконы бывают различных цветов: от коричневого, коричнево-красного до кристально прозрачного, — и различных назначений, в зависимости от содержания примесей. Циркон является ценным минералом, который используется как основной минерал-источник циркония и гафния, в ювелирном деле, как источник редкого элемента урана, цирконовый концентрат используется при производстве огнеупоров. Циркон является дополнительным акцессорным минералом в гранитах, то есть входящим в состав горных пород в очень малых количествах и поэтому не влияющим на классификацию породы, однако по нему можно узнать много полезной информации.

Первый датасет — данные, на основе которых произведен анализ образцов циркона — полученное методами аналитической химии содержание в цирконах химических элементов.

Вследствие химического анализа были получены значения по 40-ка химическим элементам для каждой характеристики каждого образца. Все цирконы разбиты на группы по принадлежности различным горным массивам. Если какие-то ячейки пустые, значит либо эти элементы не присутствуют в породе, либо прибор, которым анализируется данная проба, имеет какие-то пределы обнаружения, в которые не входит концентрация данного элемента.

Например, конкретный микрозонд может определять концентрации от 0,005 до 100 ppm. Если у нашего образца показатель 0,00000076 ppm, он не отобразится, следовательно, этот химический элемент мало значит в формировании самого образца и не является значительным.

В первом датасете представлены данные о 26-ти различных цирконах 5-ти различных горных массивов: позднее каменноугольного гранодиорит-гранитного жип кашинского комплекса, нового Дурулгуя, горы Сангар-Хая местности Санахай, Барун-Ундура и Зун-Ундура.

Во втором датасете представлены данные анализа микрозонда тех же горных массивов, которые были получены с помощью измерения электронным спектрометром.

Эта часть работы связана уже с основными минералами и породами, которые содержатся в гранитах. В число основных породообразующих минералов гранита входят кварц, слюда и полевые шпаты (см. Рис. 7).

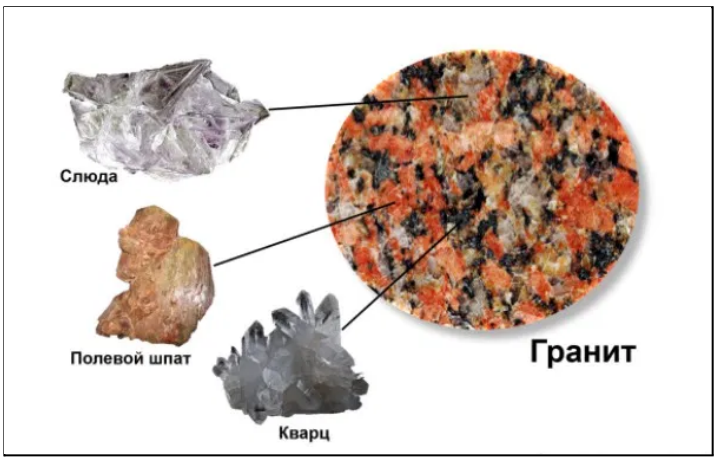


Рис. 7: Наглядное представление минерального состава гранита.

Этот прибор оценивает концентрацию элементов в образце. На рисунке 3 изображена таблица, в которой присутствуют названия химических элементов, номера точек, в которых было произведено измерение, и значения — содержание в конкретной точке конкретного химического элемента. Этот датасет более информативный и дает больше знаний о горном массиве, поскольку в нем представлены не только цирконы, но и основные породы и минералы, которые содержат граниты. Датасет содержит в себе данные по 1237-ми спектрам, информацию о сумме и участке.

Реализация приложения

Для начала в приложение были загружены обработанные исходные данные (см. Рис. 8) и выполнена подготовительная работа для успешной реализации метода.

B	C	D	E	F	G	H	I	J
	Саханай			Дурулгуй				
	Бт	Бт	Бт	Бт	Бт	Му	Двусл	
Химический элемент	C-2798	C-2793	C-2800	Д-2749	D-2749	D-2724	D-2728	Ж-3017
F	1500	1500	1300	1200	1100	2100	1600	2300
Li	50,87	188	188	129,01	129,01	129,01	129,01	188
Be	10,08	12,74	15,4	8,27	8,27	8,27	8,27	23,2
Sc	1,72	1,72	1,72	2,72	2,72	2,72	2,72	0
Ti	558,21	558,21	558,21	1002,33	1002,33	1002,33	1002,33	0
V	7,99	20,7	9,41	10,83	17,4	7,42	8,03	15,6
Cr	4,69	14,2	9,445	5,52	14,3	11,9	13,9	17,2
Mn	313,28	313,28	313,28	437,62	437,62	437,62	437,62	0

Рис. 8: Обработанные данные.

Затем была сконструирована область для заполнения данных по химическому составу неопознанного образца (см. Рис. 9). После применения метода в ячейках появляется результат, например, тот или иной горный массив, однако при начальной работе эти ячейки являются пустыми.

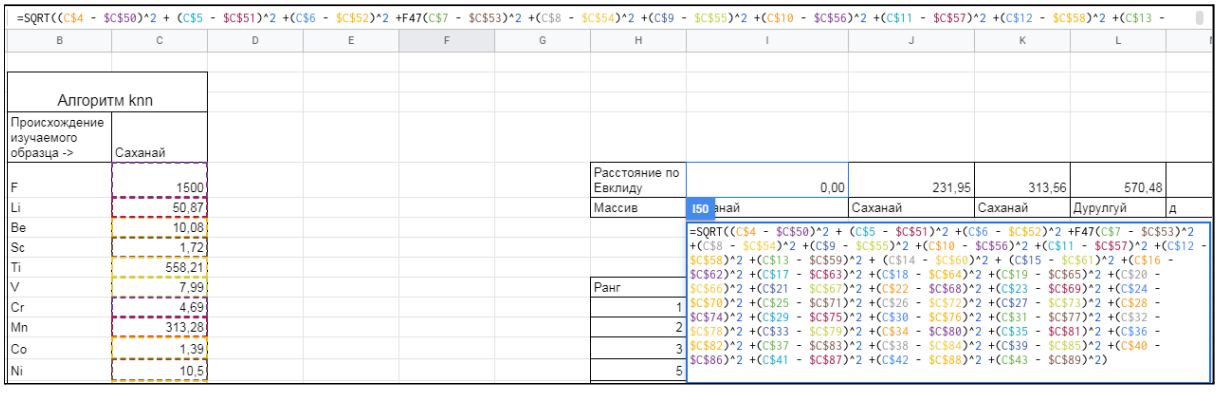


Рис. 9: Наглядное представление приложения.

Приложение помогает геологам за несколько секунд на основе уже имеющихся данных определять по химическому составу принадлежность образцов тому или иному кластеру.

Инструкция для пользователя

Приложение с помощью встроенного алгоритма классифицирует минералы по атрибутам. Для того, чтобы получить результат, необходимо проводить работу с каждым образцом по описанной инструкции:

1. Загрузка данных о неопознанном образце.

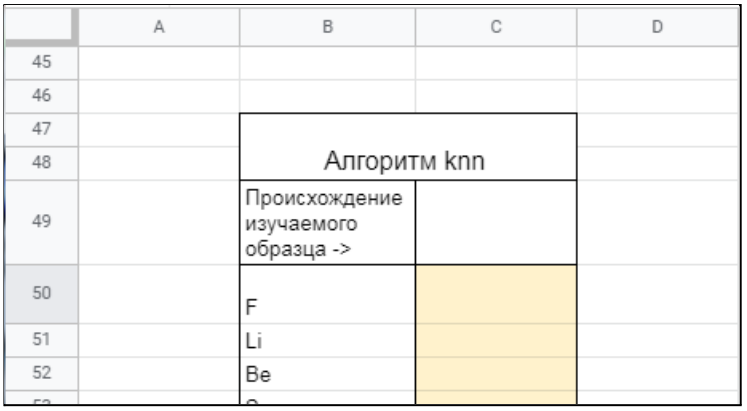
Для начала работы необходимо с помощью химического анализа получить данные об образце в виде таблицы из 2-х столбцов и 40-ка строчек. Первый столбец отвечает за химический элемент из таблицы Менделеева, второй — за количественное содержание в нашем образце того или иного химического элемента (см. Табл. 1).

Химический элемент	Внутреннее название образца
F	1500
Li	50,87
Be	10,08
Sc	1,72
Ti	558,21
V	7,99
Cr	4,69
Mn	313,28
Co	1,39
Ni	10,5
Zn	19,53
Ga	7,59
Ge	1,47
As	6,09
Rb	335,14

Sr	52,59
Y	11,35
Zr	36,52
Nb	9,1
Mo	0,4
Sn	8,06
Cs	11,79
Ba	115,18
La	18,8
Ce	32,13
Pr	4,59
Nd	15,5
Sm	3,09
Eu	0,35
Gd	2,37
Tb	0,33
Dy	2,05
Ho	0,38
Er	1,17
Tm	0,19

Табл. 1: Представление данных об образце.

Данные готовы для загрузки. Переносим данные из второго столбца в область C50:C89. На рисунке 10 видно, что изначально эта область является пустой.



	A	B	C	D
45				
46				
47		Алгоритм knn		
48				
49		Происхождение изучаемого образца ->		
50		F		
51		Li		
52		Be		

Рис. 10: Область для заполнения данными.

2. Получение результата.

После загрузки данных о неопознанном образце результат автоматически появится в ячейке C49. На рисунке 11 видно, что неопознанный образец принадлежит горному массиву Саханай.

	A	B	C	D
46				
47		Алгоритм knn		
48				
49		Происхождение изучаемого образца ->	Саханай	
50		F	1500	
51		Li	50,87	
52		Be	10,08	
53		Sc	1,72	
54		Ti	558,21	
55		V	7,99	
56		Cr	4,69	
57		Mn	313,28	
58		Co	1,39	
59		Ni	10,5	
60		Zn	19,53	
61		Ga	7,59	
62		Ge	1,47	

Рис. 11: Наглядное представление полученного результата.

Применение приложения к первому датасету

В первую очередь с помощью построенного алгоритма был проведен анализ 8-ми неопознанных образцов на принадлежность тому или иному горному массиву. Все образцы принадлежат Жипкошинскому штоку — результат полностью совпадает с анализом, который был проведен с помощью химической экспертизы.

В первом датасете имеются также данные о внутренней классификации некоторых гранитов: они могут состоять из преобладающего количества биотитовой слюды (кальция, алюминия, железа и магния), мусковитной слюды (калия) или же быть двуслюдяными — биотит и мусковит в почти равном соотношении.

B	C	D	E	F	G	H	I	J
	Саханай			Дурулгуй				
	Бт	Бт	Бт	Бт	Бт	Му	Двусл	
Химический элемент	C-2798	C-2793	C-2800	Д-2749	D-2749	D-2724	D-2728	Ж-3017
F	1500	1500	1300	1200	1100	2100	1600	2300
Li	50,87	188	188	129,01	129,01	129,01	129,01	188
Be	10,08	12,74	15,4	8,27	8,27	8,27	8,27	23,2
Sc	1,72	1,72	1,72	2,72	2,72	2,72	2,72	0
Ti	558,21	558,21	558,21	1002,33	1002,33	1002,33	1002,33	0
V	7,99	20,7	9,41	10,83	17,4	7,42	8,03	15,6
Cr	4,69	14,2	9,445	5,52	14,3	11,9	13,9	17,2
Mn	313,28	313,28	313,28	437,62	437,62	437,62	437,62	0

Рис. 12: Внутренняя классификация гранитов.

С помощью сконструированного ранее алгоритма мы получили автоматическую классификацию для еще не классифицированных гранитов (21-го образца).

БТ	БТ	БТ	БТ	МУ	МУ	МУ	МУ	МУ	МУ	МУ	МУ	МУ	МУ	МУ	МУ	МУ	МУ	МУ	МУ	МУ
Ж-3017	Ж-2490	Ж-3021	Ж-2495	Ж-3011	Ж-2494	Ж-2838	Ж-2838	Ж-2837	Ж-2840	Ж-3025	Ж-3026	Ж-3030	В-2835	В-2831	Z-3043	Z-1757	З-2807	З-2808	Z-1744	З-2823

Табл. 2: Классификация гранитов по породообразующим минералам.

В таблице 2 представлены результаты. Первая строка — классификация, преобладание биотита (БТ) или мусковита (МУ), вторая

строка — наше внутреннее название образцов. Результат совпадает с результатом, который получили геологи, на 90 процентов.

Результаты работы с первым датасетом

Как уже было сказано в предыдущей главе, с помощью сконструированного knn-алгоритма мы получили автоматическую классификацию для еще не классифицированных гранитов (см. Табл. 2).

Можно сделать вывод, что классифицированные образцы Ж-3011, Ж-3017 и т.д. можно использовать в строительстве и как камень специального назначения (кислотоупорный и пр.). Гранит с высоким содержанием калиевого полевого шпата (породообразующим которого является мусковит) — полевошпатовое сырьё, которое используется как флюс при производстве стекла и тонкой керамики.

В промышленности граниты, породообразующими которых является биотит, используются для приготовления бронзовой краски и жаростойких масс, в электротехнике (флогопит).

Применение приложения ко второму датасету

Во втором датасете также имеются данные о внутренней классификации некоторых гранитов по преобладающему количеству биотитовой слюды, мусковитной слюды или же содержанию биотита и мусковита в почти равном соотношении. На основе уже известной классификации (136-ти гранитах) с помощью приложения мы распределили оставшиеся 1101 гранит (см. Табл. 3) по трем кластерным группам: “Биотит”, “Мусковит” и “Двуслюдяные”.

БТ	БТ	БТ	БТ	МУ	МУ	МУ	ДВ	ДВ	ДВ
Спектр 1506	Спектр 1507	Спектр 1508	Спектр 1509	Спектр 1274	Спектр 1275	Спектр 1276	Спектр 1279	Спектр 1280	Спектр 1281

Табл. 3: Часть классификации второго датасета.

Результаты работы со вторым датасетом

В главе “Результаты работы с первым датасетом” подробно описано промышленное применение гранитов, у которых породообразующими минералами являются либо мусковит, либо биотит. Во втором датасете появилась также группа “Двуслюдяные”. В промышленности минералы, принадлежащие этому классу, используются для изготовления кровельных материалов: крошкой из двуслюдяного сланца покрывают поверхности некоторых видов рубероидов. Рубероид — рулонный кровельный и гидроизоляционный материал, изготовляемый пропиткой кровельного картона легкоплавкими нефтяными битумами с последующим покрытием его слоем тугоплавкого битума и защитной посыпкой асбестом, тальком, песком и т. д.

Заключение

В результате работы было создано пользовательское приложение, которое с помощью алгоритма knn определяет принадлежность образца гранита той или иной кластерной группе на основе химических данных по соответствующему образцу. С помощью приложения мы определили принадлежность гранитов тому или иному горному массиву, узнали породообразующие элементы группы образцов и сделали выводы о промышленной полезности исследуемых образцов. Датасеты и приложение выложены в открытый доступ: https://vk.cc/c1XA4c, https://vk.cc/c1XA6m, https://vk.cc/c1XA1k.

Список литературы

[1] В.Н. Лодочников. Главнейшие породообразующие минералы — 1955.

[2] К.В. Воронцов. Машинное обучение (курс лекций): http://www.machinelearning.ru/wiki/index — 2019.

[3] Андрей Бурков. Машинное обучение без лишних слов — 2020.

[4] С.Г. Скублов, А.В. Березин, Н.Г. Бережная. Общие закономерности состава цирконов из эклогитов по редким элементам применительно к проблеме возраста эклогитов Беломорского подвижного пояса — 2012.

[5] Никита Прияцелюк. Обзор самых популярных алгоритмов машинного обучения — 2018.

[6] А.А. Иванова, Л. Ф. Сырицо, Е.В. Баданина, А.М. Сагитова. Циркон полиформационного Тургинского массива с амазонитовыми гранитами (Восточное Забайкалье) и его петрогенетическое значение — 2018.

[7] Ewan Pelletera, Alain Cheilletza, Dominique Gasquet, Abdellah Mouttaqid, Mohammed Annich, Abdelkhalek El Hakour, Etienne Deloule, Gilbert Féraud. Hydrothermal zircons: A tool for ion microprobe U–Pb dating of gold mineralization (Tamlalt–Menhouhou gold deposit — Morocco) — 2017.

[8] Е.В. Трегер. Таблицы для оптического определения породообразующих минералов — 1958.

[9] А.М. Плякин, В.А. Жемчугова, Н.П. Минова. Породообразующие минералы и горные породы — 1999.

[10] У.А. Дир, Дж. Зусман, Р.А. Хауи. Породообразующие минералы. Том 5. Несиликатные минералы — 1965.