ГУЩИНА О. А., КОРЖОВ А. С.

ПРИМЕНЕНИЕ АЛГОРИТМА СЛУЧАЙНОГО ЛЕСА ДЛЯ АВТОМАТИЗАЦИИ КЛАССИФИКАЦИИ КАТЕГОРИЙ ГРУНТОВ

Аннотация. В статье рассматривается создание модели машинного обучения для решения задачи классификации грунтов с использованием ансамбля случайных деревьев решений (случайного леса) для автоматизации определения с максимальной точностью категорий грунтов на основе имеющихся о них данных, включающих такие характеристики как плотность, влажность, фракционный состав и прочие. Также представлен пользовательский интерфейс разработанной программно-информационной системы для проведения предсказательной аналитики с помощью полученной модели.

Ключевые слова: дерево решений, алгоритм случайного леса, машинное обучение, предсказательная аналитика.

GUSHCHINA O. A., KORZHOV A. S.

APPLYING RANDOM FOREST ALGORITHM FOR AUTOMATING CLASSIFICATION OF SOIL CATEGORIES

Abstract. The article discusses the creation of a machine learning model to solve the problem of soil classification using an ensemble of random decision trees (random forest) to automate the determination of soil categories with maximum accuracy based on the data available about them, including such characteristics as soil density, humidity, fractional composition and others. The user interface of the developed software and information system for conducting predictive analytics using the resulting model is also presented.

Keywords: decision tree, random forest algorithm, machine learning, predictive analytics.

Постановка задачи. Пусть дана таблица из результатов проб грунта (более 500 000 экземпляров) с 32 характеристиками различных грунтов. На основе данных взятых проб грунта в Москве и Московской области необходимо определить категорию грунта и сделать заключение о пригодности почвы к застройке определенных типов зданий. Это делается потому, что тип грунта ключевым фактором, влияющим на выбор типа фундамента здания (так как некоторые типы грунта более подходят для строительства строго определенных типов зданий) [1].

Процесс решения задачи. Грунты, представляющие собой комплекс природных материалов, различаются по своему происхождению, составу (супесчаные, глинистые, песчанисто-глинистые, глинисто-песчанистые и пр.), химическому составу, физическим

свойствам (плотность, влажность, текстура, консистенция) и другим характеристикам (палеоповерхность, уровень углерода, минерализованные зоны) [2].

Для решения поставленной задачи целесообразно использовать возможности языка программирования Python, который специально используется в машинном обучении для выполнения научных и коммерческих проектов [3; 4]. Так, библиотека scikit-learn предназначена для непосредственного построения и работы с деревьями решений.

При разработке программно-информационной системы (ПИС) был использован CSV-файл с результатами анализа грунта. Структура CSV-файла представлена в таблице 1. CSV-файл расположен на Google Drive и, после загрузки, данные сохраняются в переменной 'df' типа DataFrame.

Таблица 1

Структура CSV-файла

1	index	Класс грунта
2	input_index	Входное значение класса по результатам первичной обработки экспертами (может совпадать или не совпадать с выходными значениями)
3	index_kod	Дубль поля input_index, но записан как id (число). Можно воспринимать как ранговый показатель, т.к. порядок закреплен – большие значения id лежат глубже меньших
4	prev_index	Предыдущий индекс – проинтерпретированный класс предшествующего (меньшего по глубине) слоя в данной скважине. Важный параметр, поскольку в большинстве случаев породы залегают последовательно и более древние породы не могут залегать выше более молодых
5	bottom	Абсолютные отметки кровли и подошвы слоя (метры над уровнем моря)
6	top	
7	depth_to_paleosurf	Разница между отметкой кровли слоя и поверхностью палеорельефа. Палеорельеф – результат работы по картированию ненарушенной топографии городской территории (до начала хозяйственного использования)
8	depth_to_carbon	Разница между кровлей слоя и поверхностями дочетвертичных отложений
9	depth_to_mz	Разница между кровлей слоя и поверхностями каменноугольных отложений
10	top_MZ_map	top + depth_to_mz
11	top_C_map	top + depth_to_carbon
12	paleosurf	top + depth_to_paleosurf
13	x	Координаты скважины в некоторой системе координат, в нашем случае – в Московской системе координат
14	y	
15	okrug	Округ Москвы. Введен для упрощения учета территориального фактора
16	PreQ_map	Числовая информация с разных геологических карт. Для разных геологических индексов важна разная информация
17	Carb_map	
18	Q_map	
19	Geomorf	

Продолжение таблицы 1

20	genesis	Генезис происхождение грунтов – например речные, морские, болотные, ледниковые отложения на основе входного класса. За каждым классом жестко закреплен генезис. Чаще всего генезис не меняется в процессе интерпретации. Генезис можно рассматривать как кластер более высокого уровня, один генезис всегда включает несколько индексов. Грунты одного индекса не могут быть разного генезиса
21	litol	Литологическое описание грунта Дисперсные грунты: – Глинистые грунты – Глины, Суглинки, Супеси. Относятся к одной группе глинистых грунтов, но это не значит, что это одно и то же. В некоторых слоях (древние каменноугольные отложения – генезис carbon) могут быть только глины, причём только твердые и полутвердые. – Песчаные грунты – пески различной крупности и плотности сложения. – Крупнообломочные грунты (щебень, гравий, галька и т.д.). – Скальные грунты: известняк, доломит, мергель, песчаник. Бывают только в древних отложениях (генезис – carbon, Creatcious, Jurassic). – Специфические органогенные грунты – торф, ил, сапропель. Редкие и опасные для строительства. – Техногенные грунты – отходы человеческой деятельности, строительные материалы и т.д. (генезис – technogen). – Карстовые полости – большие полости в известняках, часто заполненные дисперсными грунтами
22	podoshva	Глубина подошвы слоя
23	prochn	Прочность грунта (только для скальных пород)
24	vkluch	Наличие или отсутствие включений (случайные органические или минеральные тела или предметы, генетически не связанные с почвенными процессами)
25	vlaga	Влажность, водонасыщенность грунта
26	color_lit	Цвет грунта
27	konsist	Консистенция грунта: текучесть / пластичность / твердость и т.д. (только для глинистых пород)
28	krupnost	Крупность (только для песчаных грунтов)
29	kavern	Кавернозность – наличие полостей и пор в скальных грунтах
30	plotn	Плотность сложения песчаных грунтов
31	sohran	Степень разрушения скального грунта
32	cons	konsist в числовом формате

Для эффективной обработки и анализа результатов анализа проб грунта, необходимо реализовать процесс импорта данных из базы данных для их последующей обработки с использованием алгоритмов машинного обучения, а также процесс экспорта полученных результатов. Для этого были проведены следующие манипуляции с данными.

1. Выбор таблицы и данных из корпоративной базы данных с помощью SQL-запросов.

2. Преобразование полученной таблицы в файл CSV и его экспорт.

3. Импорт CSV-файла в среду Jupyter Lab с помощью Pandas.

4. Обработка данных с помощью моделей машинного обучения является основным этап и включает следующие подэтапы обработки информации.

4.1. Преобразование типов данных.

4.2. Обработка пропущенных значений.

4.3. Нормализация данных.

4.4. Построение моделей машинного обучения.

4.5.Преобразование в исходный формат.

5. Подготовка полученных данных к экспорту в формат CSV.

6. Загрузка файла CSV в базу данных с использованием SQL.

На рисунке 1 представлена диаграмма вариантов использования ПИС.

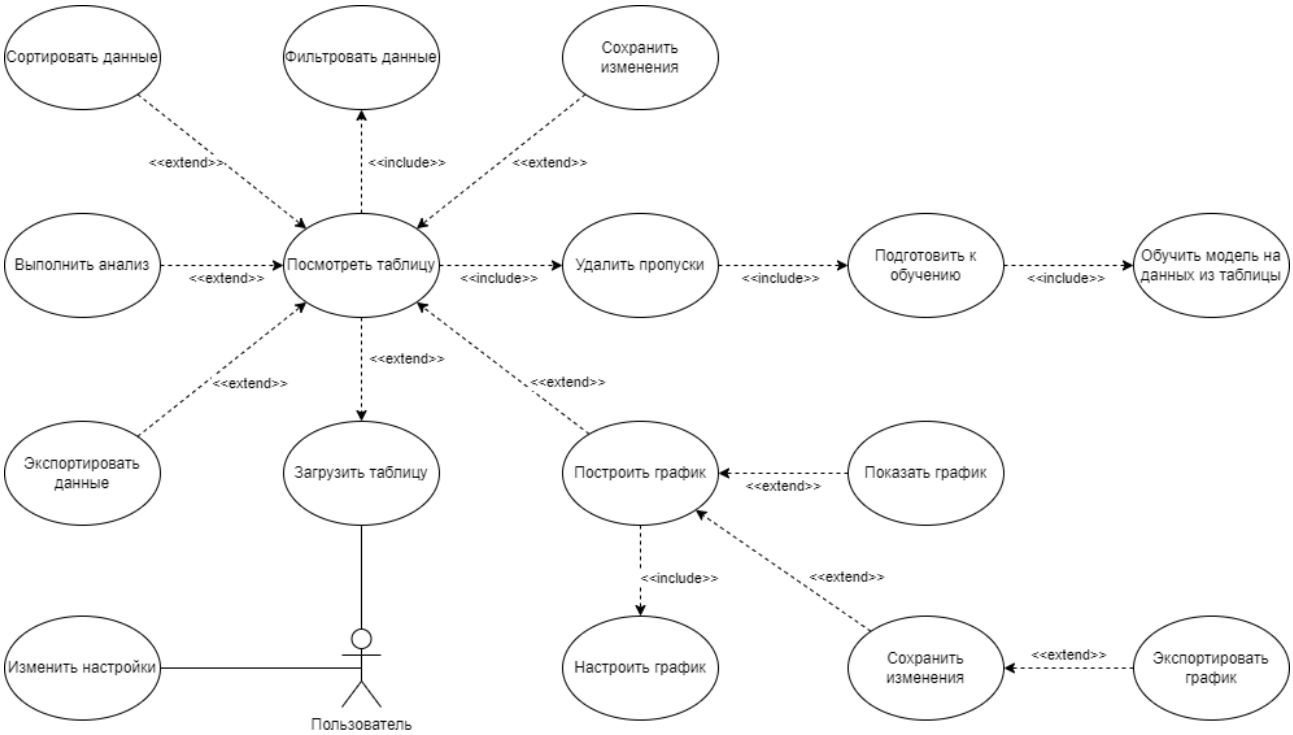


Рис. 1. Диаграмма вариантов использования программно-информационной системы для определения категории грунта.

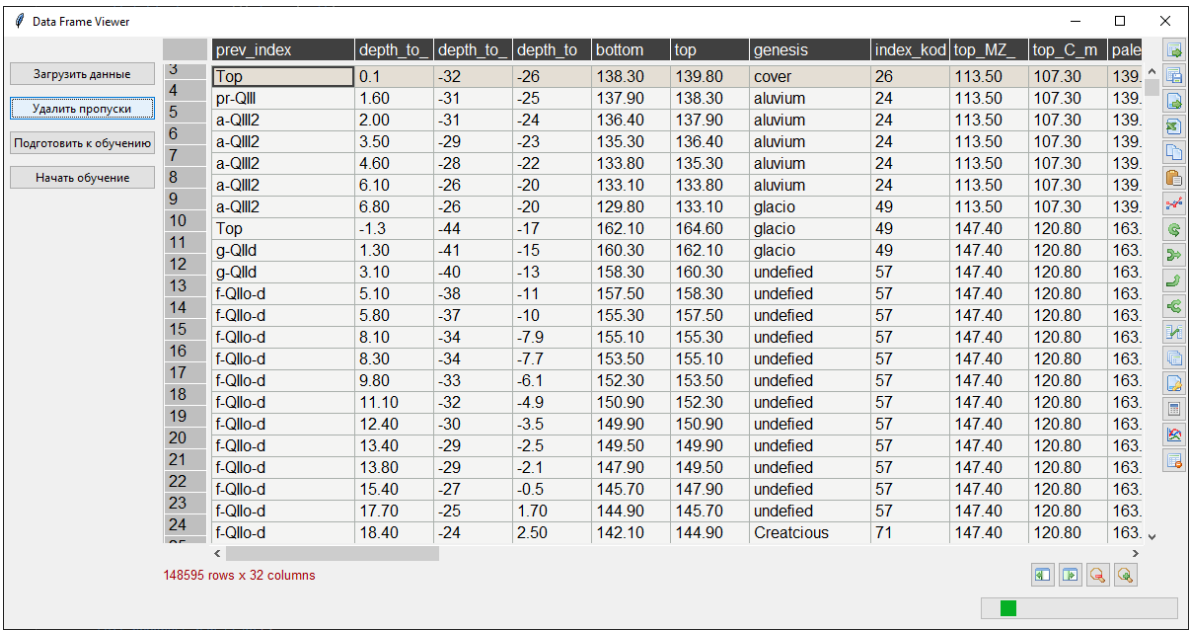
ПИС разработана на языке программирования Python с его мощными библиотеками для научных вычислений (pandas, numpy, sklearn, seaborn, matplotlib, tkinter, pandastable и time). Она выполняет перечисленную выше последовательность преобразований и визуализаций данных, а затем импортирует их в модель машинного обучения Random Forest для непосредственного определения типа грунта на основе предоставленных данных. Random Forest работает с ансамблями деревьев решений и используется для задач

классификации и регрессии. То есть он представляет собой ансамблевый метод, который комбинирует прогнозы нескольких деревьев решений для улучшения общей производительности модели.

Обучение модели является ключевым шагом, позволяющим модели извлекать закономерности и зависимости из имеющихся данных и использовать их для предсказания типов грунта. В процессе обучения модели Random Forest, каждое дерево строится на основе случайной подвыборки данных из обучающей выборки. Для каждого дерева также случайно выбирается подмножество признаков. Это помогает снизить переобучение модели и повысить ее обобщающую способность. Деревья строятся путем разделения данных на основе различных признаков и значений. Признаки и их значения используются для создания условий, по которым данные разделяются на более чистые подгруппы, соответствующие различным типам грунта. Процесс разделения продолжается до достижения определенного критерия остановки, такого как достижение определенной глубины дерева или минимального числа образцов в листьях. Каждое дерево в Random Forest производит свое собственное предсказание типа грунта для каждого образца данных. Затем, для получения окончательного предсказания, модель применяет голосование или усреднение предсказаний от всех деревьев в ансамбле. Это позволяет учесть мнение нескольких деревьев и получить более надежный результат.

Обучение модели Random Forest включает построение ансамбля деревьев и настройку их параметров. Указываем количество деревьев в ансамбле, функцию измерения качества разделения, минимальное число образцов в листьях и бутстрап. Настройка параметров может влиять на производительность и обобщающую способность модели. После завершения обучения модели Random Forest на обучающей выборке, можно провести оценку производительности модели на тестовой выборке для оценивания качества классификации и способности модели обобщать на новые данные.

На рисунке 2 приведен пример графического интерфейса пользователя ПИС при выполнении действия (о чем свидетельствует прогресс-бар в правой нижней части экрана). Он позволяет выполнить: импорт, обработку, визуализацию и экспорт данных; выбрать средства для редактирования и изменения отображения графиков; выбрать средства для обучения моделей машинного обучения.



	prev_index	depth_to_	depth_to_	depth_to	bottom	top	genesis	index_kod	top_MZ_	top_C_m	pale
3	Top	0.1	-32	-26	138.30	139.80	cover	26	113.50	107.30	139.
4	pr-QIII	1.60	-31	-25	137.90	138.30	aluvium	24	113.50	107.30	139.
5	a-QIII2	2.00	-31	-24	136.40	137.90	aluvium	24	113.50	107.30	139.
6	a-QIII2	3.50	-29	-23	135.30	136.40	aluvium	24	113.50	107.30	139.
7	a-QIII2	4.60	-28	-22	133.80	135.30	aluvium	24	113.50	107.30	139.
8	a-QIII2	6.10	-26	-20	133.10	133.80	aluvium	24	113.50	107.30	139.
9	a-QIII2	6.80	-26	-20	129.80	133.10	glacio	49	113.50	107.30	139.
10	Top	-1.3	-44	-17	162.10	164.60	glacio	49	147.40	120.80	163.
11	g-QIId	1.30	-41	-15	160.30	162.10	glacio	49	147.40	120.80	163.
12	g-QIId	3.10	-40	-13	158.30	160.30	undefied	57	147.40	120.80	163.
13	f-QIIo-d	5.10	-38	-11	157.50	158.30	undefied	57	147.40	120.80	163.
14	f-QIIo-d	5.80	-37	-10	155.30	157.50	undefied	57	147.40	120.80	163.
15	f-QIIo-d	8.10	-34	-7.9	155.10	155.30	undefied	57	147.40	120.80	163.
16	f-QIIo-d	8.30	-34	-7.7	153.50	155.10	undefied	57	147.40	120.80	163.
17	f-QIIo-d	9.80	-33	-6.1	152.30	153.50	undefied	57	147.40	120.80	163.
18	f-QIIo-d	11.10	-32	-4.9	150.90	152.30	undefied	57	147.40	120.80	163.
19	f-QIIo-d	12.40	-30	-3.5	149.90	150.90	undefied	57	147.40	120.80	163.
20	f-QIIo-d	13.40	-29	-2.5	149.50	149.90	undefied	57	147.40	120.80	163.
21	f-QIIo-d	13.80	-29	-2.1	147.90	149.50	undefied	57	147.40	120.80	163.
22	f-QIIo-d	15.40	-27	-0.5	145.70	147.90	undefied	57	147.40	120.80	163.
23	f-QIIo-d	17.70	-25	1.70	144.90	145.70	undefied	57	147.40	120.80	163.
24	f-QIIo-d	18.40	-24	2.50	142.10	144.90	Creatcious	71	147.40	120.80	163.

Рис. 2. Отображение графического интерфейса пользователя ПИС при выполнении действия.

При работе пользователи активно использует методы визуализации выбранных из таблицы данных в виде графиков прямо в интерфейсе (рис. 3). Существует возможность настройки различных атрибутов графика: написания заголовка, подписи осей, стиля линий.

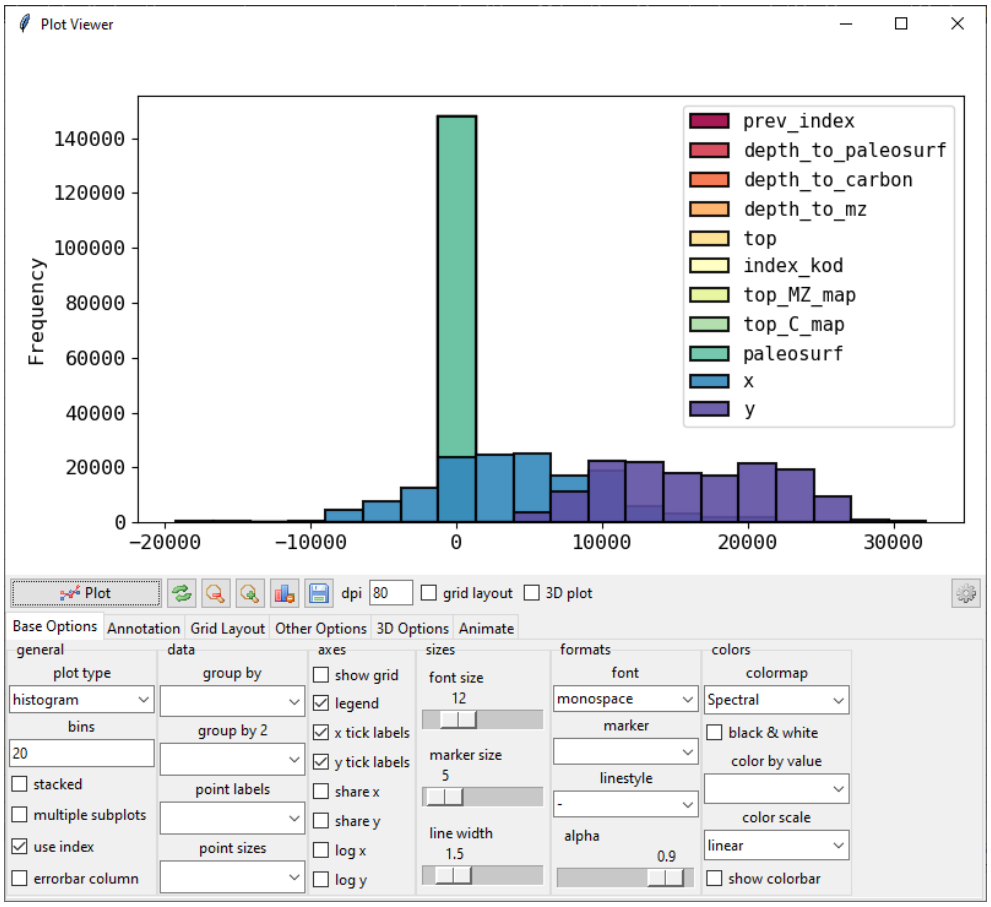


Рис. 3. Пример отображения данных и настройки его параметров.

После загрузки данных из .csv файла проводится предварительный анализ данных. Структура данных состоит из 32 столбцов (то есть 32 признаков для классификации типов грунта) и 148595 строк (то есть данных для 148595 различных экземпляров проб грунта).

Далее было произведено изучение распределения значений в каждом признаке, используя статистические метрики (минимум и максимум, среднее значение, стандартное отклонение). Также были построены гистограммы распределения значений для визуального оценивания форм распределения и выявления возможных выбросы или аномалии.

При анализе данных были выявлены пропущенных значений. Далее, используя библиотеку pandas, были идентифицированы эти пропущенные значения и к каждому была применена соответствующая стратегия обработки (в некоторых случаях пропущенные значения были заполнены неизвестными или полностью удалены некорректные данные).

Затем был выполнен анализ, включающий вычисление статистических метрик, визуализацию данных и проверку корреляции между признаками и целевой переменной. В результате были выбраны наиболее значимые признаки для классификации грунтов, то есть сформировано подмножество признаков, которые будут использоваться для обучения модели. Визуализации boxplot полезна для сравнения распределений разных признаков и выявления потенциальных выбросов или необычных значений, что важно при принятии решений о предобработке данных или выборе соответствующих статистических методов (рис. 4).

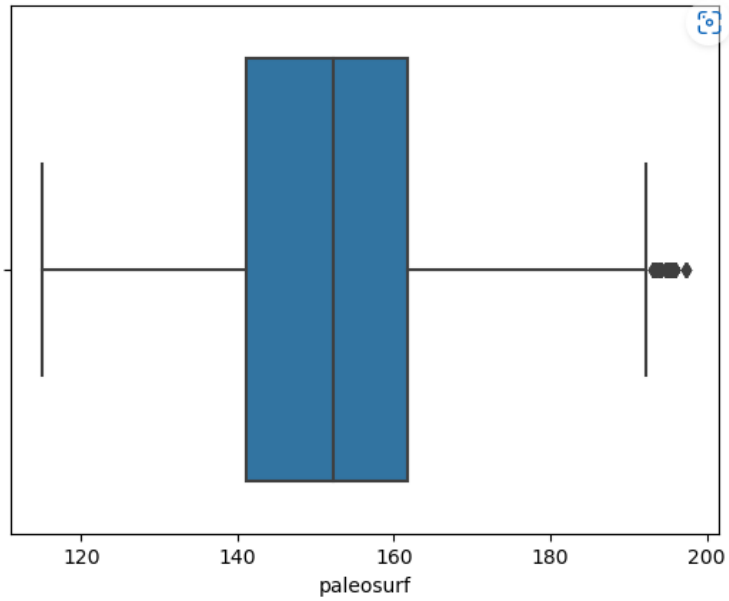


Рис. 4. Пример применения boxplot на один из признаков.

Гистограммы распределения данных использовались для визуализации частоты появления значений в различных интервалах. Она позволяет наглядно представить форму распределения данных и выявить ее особенности (моды, наличие выбросов и прочие). Это

помогает принять решения о преобразовании данных, удалении выбросов или выборе соответствующих статистических методов для дальнейшего анализа и моделирования.

Для анализа данных и подготовки их к обучению модели была использована корреляционная матрица (рис. 5). Она позволяет определить степень линейной связи между признаками и выявить сильно коррелирующие признаки. Это важно для исключения дублирования информации и снижения влияния мультиколлинеарности на процесс обучения модели. После анализа матрицы корреляции можно принимать решение о выборе наиболее значимых признаков и использовать их для обучения модели.

	depth_to_paleosurf	depth_to_carbon	depth_to_mz	top	top_MZ_map	top_C_map	paleosurf	x	y
depth_to_paleosurf	1.000000	0.679863	0.847739	-0.734667	-0.122145	-0.011433	-0.095446	-0.084972	-0.179697
depth_to_carbon	0.679863	1.000000	0.757827	-0.881950	-0.452814	0.374055	-0.626097	0.046105	-0.412643
depth_to_mz	0.847739	0.757827	1.000000	-0.755472	0.025480	0.112045	-0.275554	-0.072698	-0.382344
top	-0.734667	-0.881950	-0.755472	1.000000	0.635716	0.107225	0.745448	-0.084348	0.461933
top_MZ_map	-0.122145	-0.452814	0.025480	0.635716	1.000000	0.295630	0.812772	-0.214329	0.254355
top_C_map	-0.011433	0.374055	0.112045	0.107225	0.295630	1.000000	0.146069	-0.068706	0.038490
paleosurf	-0.095446	-0.626097	-0.275554	0.745448	0.812772	0.146069	1.000000	-0.207281	0.501189
x	-0.084972	0.046105	-0.072698	-0.084348	-0.214329	-0.068706	-0.207281	1.000000	-0.029307
y	-0.179697	-0.412643	-0.382344	0.461933	0.254355	0.038490	0.501189	-0.029307	1.000000

Рис. 5. Матрица корреляции.

Далее в четыре этапа происходит подготовка данных к обучению.

На первом этапе с помощью метода get_dummies было проведено преобразование категориальных переменных в числовые значения для подготовки данных к классификации с использованием алгоритма Random Forest (так как при наличии категориальными данными не получится их использовать в моделях машинного обучения (линейной регрессии или дереве принятия решений)).

На втором этапе было выполнено масштабирование признаков, то есть приведение значений признаков к одному и тому же диапазону или единой шкале (иначе признаки с большими значениями в своих диапазонах будут иметь большее влияние на модель, чем признаки с меньшими значениями и/или с меньшими диапазонами). Для масштабирования признаков использовались методы стандартизации (z-масштабирование). То есть при стандартизации признаков вычисляется среднее значение и стандартное отклонение каждого признака. Затем каждое значение признака вычитается из среднего значения и делится на стандартное отклонение. Это приводит к тому, что каждый признак имеет среднее значение равное нулю и стандартное отклонение равное единице.

На третьем этапе с помощью функции concat колонки объединяются и подготавливаются датафреймы к обучению.

На четвертом этапе проводится разделение данных на обучающую и тестовую выборки, необходимо для оценивания производительности модели машинного обучения и проверки ее способности к обобщению на новых данных. Пусть обучающая выборка составляет 70% от исходного набора данных, а тестовая – 30%.

После всех проведенных подготовительных работ переходим к обучению модели на обучающей выборке, что позволит ей извлекать закономерности и зависимости из имеющихся данных и использовать их для предсказания типов грунта.

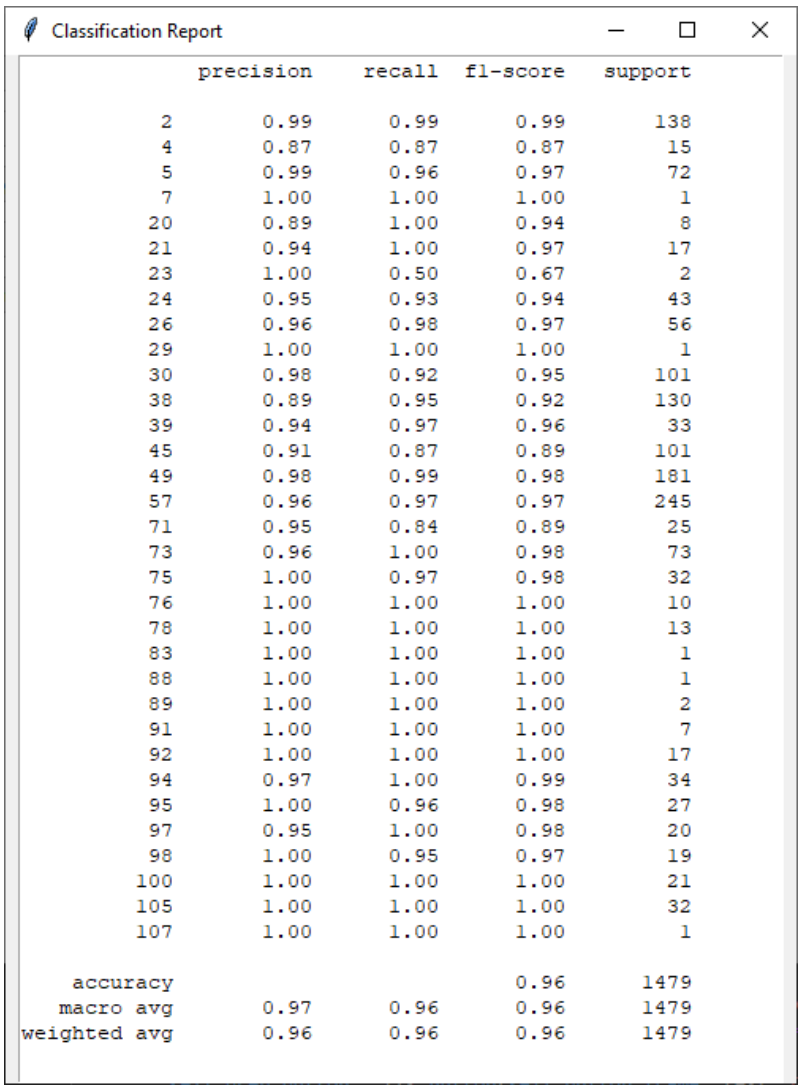
В процессе обучения модели Random Forest, каждое дерево строится на основе случайной подвыборки данных из обучающей выборки. Для каждого дерева также случайно выбирается подмножество признаков (это снижает переобучение модели и повышает ее обобщающую способность). Деревья строятся путем разделения данных на основе различных признаков и значений. Признаки и их значения используются для создания условий, по которым данные разделяются на более чистые подгруппы, соответствующие различным типам грунта. Процесс разделения продолжается до достижения определенного критерия остановки, такого как достижение определенной глубины дерева или минимального числа образцов в листьях.

Каждое дерево в Random Forest производит свое собственное предсказание типа грунта для каждого образца данных. Затем, для получения окончательного предсказания, модель применяет голосование или усреднение предсказаний от всех деревьев в ансамбле. Это позволяет учесть мнение нескольких деревьев и получить более надежный результат.

Обучение модели Random Forest включает построение ансамбля деревьев и настройку их параметров. Изначально указывается количество деревьев в ансамбле, функции измерения качества разделения, минимальное число образцов в листьях и бутстрап. Настройка параметров может влиять на производительность и обобщающую способность модели. После завершения обучения модели на обучающей выборке, можно провести оценку производительности модели на тестовой выборке.

После завершения обучения модели выполняется анализ результатов обучения, включающий оценку точности модели на тестовых данных, генерацию отчета о классификации и другие метрики оценивания модели (рис. 6).

Реализация интерфейса ПИС осуществлялась с использованием библиотеки Tkinter для создания интуитивно понятного графического пользовательского интерфейса (GUI) и PandasTable для отображения и взаимодействия с табличными данными.



	precision	recall	f1-score	support
2	0.99	0.99	0.99	138
4	0.87	0.87	0.87	15
5	0.99	0.96	0.97	72
7	1.00	1.00	1.00	1
20	0.89	1.00	0.94	8
21	0.94	1.00	0.97	17
23	1.00	0.50	0.67	2
24	0.95	0.93	0.94	43
26	0.96	0.98	0.97	56
29	1.00	1.00	1.00	1
30	0.98	0.92	0.95	101
38	0.89	0.95	0.92	130
39	0.94	0.97	0.96	33
45	0.91	0.87	0.89	101
49	0.98	0.99	0.98	181
57	0.96	0.97	0.97	245
71	0.95	0.84	0.89	25
73	0.96	1.00	0.98	73
75	1.00	0.97	0.98	32
76	1.00	1.00	1.00	10
78	1.00	1.00	1.00	13
83	1.00	1.00	1.00	1
88	1.00	1.00	1.00	1
89	1.00	1.00	1.00	2
91	1.00	1.00	1.00	7
92	1.00	1.00	1.00	17
94	0.97	1.00	0.99	34
95	1.00	0.96	0.98	27
97	0.95	1.00	0.98	20
98	1.00	0.95	0.97	19
100	1.00	1.00	1.00	21
105	1.00	1.00	1.00	32
107	1.00	1.00	1.00	1
accuracy			0.96	1479
macro avg	0.97	0.96	0.96	1479
weighted avg	0.96	0.96	0.96	1479

Рис. 6. Пример квалификационного отчета.

Заключение. В результате проведенного исследования разработано программное обеспечение (ПИС) для анализа и классификации данных проб грунтов, а также практического определению категории грунта на основе имеющихся о них данных.

СПИСОК ЛИТЕРАТУРЫ

1. Специалисты Газпром нефти научили программу исследовать образцы керна по фото [Электронный ресурс]. – Режим доступа: https://neftegaz.ru/news/standarts/637475-spetsialisty-gazprom-nefti-nauchili-programmu-issledovat-obraztsy-kerna-po-foto/ (дата обращения: 03.09.2023).
2. Комплексный анализ и определение фильтрационно-емкостных свойств геологического образца [Электронный ресурс]. – Режим доступа: https://xn--b1aghfftcbpg0bw.xn--p1ai/ (дата обращения: 11.10.2023).

3. Документация Python 3.12.0 [Электронный ресурс]. – Режим доступа: https://docs.python.org/3/ (дата обращения: 11.10.2023).
4. Груздев А. В. Прогнозное моделирование в IBM SPSS Statistics, R и Python: метод деревьев решений и случайный лес. – М.: ДМК Пресс, 2018. – 642 с.