МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра информатики и программирования

ИДЕНТИФИКАЦИЯ МЕЛОВЫХ ОТЛОЖЕНИЙ ДЛЯ ОПРЕДЕЛЕНИЯ ВЕРОЯТНЫХ УЧАСТКОВ РАСПРОСТРАНЕНИЯ КРАСНОКНИЖНЫХ РАСТЕНИЙ-КАЛЬЦЕФИЛОВ С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Кондратовой Александры Дмитриевны

Научный руководитель:
зав. кафедрой, доцент, к.ф.-м.н. _____ М. В. Огнева

подпись, дата

Зав. кафедрой:
доцент, к.ф.-м.н. _____ М. В. Огнева

подпись, дата

Саратов 2026

ВВЕДЕНИЕ

Актуальность темы. Методы машинного обучения сегодня играют ключевую роль в развитии множества научных и прикладных областей. Их ценность обусловлена способностью обрабатывать огромные массивы данных, выявлять скрытые закономерности и строить точные прогнозы – то, что практически невозможно сделать традиционными методами в разумные сроки. Алгоритмы машинного обучения эффективно работают с разнородными источниками информации: спутниковыми снимками, климатическими рядами, геоданными и т. д., что обеспечивает их применение в науках о Земле, биологии, медицине, промышленности и многих других сферах.

В науках о Земле такие алгоритмы используются для моделирования рельефа, прогнозирования природных катастроф, анализа урбанизации и климатических изменений на основе ГИС-данных. В биологии методы машинного обучения применяются для расшифровки геномов, классификации видов, изучения миграций животных и моделирования экосистем. В частности, в задачах сохранения биоразнообразия машинное обучение позволяет прогнозировать распространение редких растений и животных, связывая их ареалы с геологическими и климатическими факторами, как в случае с растениями-кальцефилами и меловыми отложениями.

Растения-кальцефилы – виды, экологически привязанные к карбонатным породам, часто являются эндемиками или реликтами, формируют уникальные растительные сообщества и служат индикаторами особых геологических условий. На севере Саратовской области существуют участки выходов меловых отложений, на которых произрастают такие редкие краснокнижные растения, однако подобные участки лишь частично обследованы биологами и географами СГУ и размечены вручную по космоснимкам. Применение методов машинного обучения для идентификации меловых отложений открывает новые возможности в

прогнозировании вероятных мест распространения растений-кальцефилов с целью определения территорий для дальнейших полевых исследований, что обуславливает актуальность темы работы.

Цель бакалаврской работы – с помощью методов машинного обучения на основе полевых и камеральных данных выявить закономерности расположения меловых участков с вероятным произрастанием растений-кальцефилов.

Поставленная цель определила **следующие задачи**:

1. выполнить обзор источников;
2. изучить методы обучения и метрики качества в задачах классификации и кластеризации;
3. исследовать и предобработать набор данных с точками меловых отложений;
4. применить методы одноклассовой классификации для идентификации точек с меловыми отложениями;
5. применить методы кластеризации для идентификации точек с меловыми отложениями;
6. применить бинарную классификацию для идентификации точек с меловыми отложениями;
7. сравнить полученные результаты и сделать выводы.

Методологические основы применения методов машинного обучения для идентификации геологических и биологических объектов представлены в работах Н.А. Корецкого, В.И. Сахнюка, И.С. Степанова, Т.А. Оксенчук, А.В. Ваганова, А.В. Васильева, Е.В. Попова и В.В. Брыкина.

Практическая значимость бакалаврской работы. заключается в том, что полученные результаты могут быть использованы специалистами в области географии и ландшафтоведения при планировании полевых исследований территорий с вероятным распространением краснокнижных растений-кальцефилов, а также при выявлении новых участков выходов меловых отложений на основе доступных дистанционных и

картографических данных.

Теоретическая значимость работы состоит в сравнительном анализе эффективности различных подходов машинного обучения (одноклассовой классификации, кластеризации и бинарной классификации) для решения задачи идентификации меловых отложений, а также в выявлении наиболее информативных признаков (рельефных и спектральных), определяющих качество классификации подобных геопространственных объектов.

Структура и объём работы. Бакалаврская работа состоит из введения, 2 разделов, заключения, списка использованных источников и 4 приложений. Общий объем работы – 85 страниц, из них 62 страницы – основное содержание, включая 34 рисунка и 3 таблицы, 4 приложениями, список использованных источников информации – 27 наименований.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Примеры задач географии и методы машинного обучения» посвящен обзору применения методов машинного обучения в прикладных задачах географии и геологии, а также рассмотрению моделей машинного обучения, используемых для решения задачи идентификации меловых отложений.

В разделе рассмотрены примеры использования машинного обучения для анализа геологических данных, прогнозирования природных катастроф, мониторинга растительности и оценки факторов, влияющих на распространение видов растений и животных. Отдельное внимание уделено особенностям растений-кальцефилов и условиям формирования меловых отложений, благоприятных для их произрастания.

Также в разделе описаны основные модели машинного обучения, применимые для решения поставленной задачи: бинарная классификация (метод k-ближайших соседей, градиентный бустинг), одноклассовая классификация (метод опорных векторов OneClassSVM) и кластеризация (методы k-средних и алгоритмы, основанные на плотности данных, –

DBSCAN, HDBSCAN, OPTICS). Для каждого метода рассмотрены основные гиперпараметры и метрики оценки качества. Сделан вывод, что для решения задачи идентификации точек с меловыми отложениями могут быть использованы методы бинарной и одноклассовой классификации, а также кластеризации.

Принцип работы метода одноклассовой классификации OneClassSVM заключается в построении границы вокруг области нормальных (типичных) объектов в пространстве признаков; объекты, не попавшие в эту область, рассматриваются как объекты другого класса. Такой подход может применяться как для поиска аномалий, так и для решения задачи бинарной классификации на сильно несбалансированных данных, когда более многочисленный класс выявляется как «типичный». Рассмотрены также два подхода к кластеризации: методы, требующие заранее заданного числа кластеров (KMeans, MiniBatchKMeans), и методы, формирующие кластеры на основе плотности распределения объектов без указания их числа (DBSCAN, HDBSCAN, OPTICS).

Для оценки качества классификации в разделе рассмотрена матрица ошибок, на основе которой вычисляются метрики accuracy, precision, recall и F1-мера, а также метрика ROC-AUC, измеряющая способность модели ранжировать положительные примеры выше отрицательных независимо от выбранного порога классификации. Среди реализаций градиентного бустинга рассмотрены библиотеки XGBoost, CatBoost (разработка компании “Яндекс”) и LightGBM (разработка Microsoft), которые эффективно работают с табличными данными и категориальными признаками. Для оценки качества кластеризации выделены внешние метрики, сравнивающие результат с эталонным разбиением (индекс Рэнда и его скорректированный вариант ARI), и внутренние метрики, не требующие эталонной разметки (коэффициент силуэта).

Второй раздел «Практическая часть» посвящен реализации алгоритмов машинного обучения для идентификации точек с вероятными

меловыми отложениями на основе геопространственного набора данных по территории Предволжья Саратовской области, включающего признаки геолокации, спектральные характеристики (RGB-каналы со снимков Bing и многоспектральные данные Landsat 8,9), геологический признак и параметры рельефа (FABDEM).

Исходный набор данных содержал более 16,5 млн объектов и 54 признака, при этом доля объектов с подтверждённым наличием меловых отложений составляла лишь 0,118% (около 19,5 тыс. объектов), что свидетельствует о крайне выраженном дисбалансе классов. В разделе описаны этапы исследования и предобработки данных: анализ баланса классов и уменьшение мажоритарного класса методом RandomUnderSampler, статистический анализ рельефных и спектральных признаков, выявление и удаление выбросов (в частности, по признаку экспозиции склона Expo), применение пространственно-ориентированных методов балансировки данных (undersampling по сетке, выравнивание классов по ячейкам, oversampling), преобразование координат из системы UTM в WGS84 и построение интерактивной карты с полигонами вероятных меловых выходов, построенными тремя методами (Dissolve, Convex Hull, Buffer).

Далее для идентификации точек с меловыми отложениями были самостоятельно реализованы и сопоставлены три подхода: одноклассовая классификация (OneClassSVM, в двух вариантах обучения), кластеризация (KMeans, MiniBatchKMeans, DBSCAN, HDBSCAN, OPTICS) и бинарная классификация (KNeighborsClassifier, CatBoostClassifier, XGBClassifier, LGBMClassifier). Для каждого метода были подобраны гиперпараметры, проведены вычислительные эксперименты и оценено качество с помощью соответствующих метрик (recall, F1-score, ROC-AUC, коэффициент силуэта, индекс Рэнда).

Наилучшее качество показала модель XGBClassifier с кодировщиком OneHotEncoder (recall=0.97, F1-score=0.95, ROC-AUC=0.996). Сравнение метрик качества рассмотренных моделей бинарной классификации приведено

в таблице 1.

Таблица 1 – Метрики оценки качества моделей бинарной классификации

Модель	recall	F1-score	ROC-AUC
KNeighborsClassifier	0.94	0.92	0.99
CatBoostClassifier	0.93	0.94	0.99
XGBClassifier + OrdinalEncoder	0.96	0.95	0.996
XGBClassifier + OneHotEncoder	**0.97**	**0.95**	**0.996**
LGBMClassifier	0.94	0.94	0.996

Анализ значимости признаков выявил, что наибольший вклад в результат вносят уклон рельефа (Ugly), геологический признак (geology) и спектральные каналы за май-июнь, отражающие разреженность растительности на меловых участках. Кластеризация не позволила разделить данные по признаку наличия меловых отложений ни методами с заданным числом кластеров, ни методами, основанными на плотности данных. Одноклассовая классификация показала промежуточный результат (recall=0.76, F1-score=0.70) при обучении модели только на данных класса меловых отложений.

ЗАКЛЮЧЕНИЕ

В работе был выполнен обзор источников на предмет использования методов машинного обучения в географии и геологии. В теоретической части работы описаны методы машинного обучения, которые можно применить для решения задачи предсказания областей с наличием мела, и метрики оценки качества для этих моделей.

Были исследованы и предобработаны данные датасета с описанием почв в Саратовской области. В наборе данных наблюдался сильный дисбаланс классов, который мог негативно повлиять на обучение методов машинного обучения. В наборе обнаружены и удалены выбросы.

Для идентификации точек с меловыми отложениями были использованы три подхода: одноклассовая классификация, кластеризация и бинарная классификация.

При сравнении качества применённых методов машинного обучения было выявлено, что наилучшее качество показали методы бинарной классификации. Наилучший результат показала модель XGBClassifier с использованием кодировщика OneHotEncoder, она показала значение метрик recall=0.97, F1-score=0.95, ROC-AUC=0.996. По каждой модели градиентного бустинга были выявлены наиболее значимые признаки. По результатам исследований оказалось, что наибольшее внимание стоит обращать на признаки: Ugly, May_B3, June_B3, June_B4, geology, признанные наиболее важными лучшей моделью XGBClassifier.

Результаты работы могут быть в дальнейшем использованы при полевых исследованиях местности специалистами в географии, ландшафтоведами.

Основные источники информации:

1. Ваганов, А.В. Моделирование потенциального ареала обитания растений методами машинного обучения / А.В. Ваганов, В.Ф. Зайков, О.С. Кротова [и др.] // Известия АлтГУ. – 2022. – №4 (126). – С. 85–92.

2. Оксенчук, Т.А. Применение методов машинного обучения в геологии // Вестник науки. – 2025. – № 6 (87).

3. Степанов, И.С. Использование методов машинного обучения в геоинформационных моделях при решении задач геофизической разведки // Вестник СГУГиТ. – 2024. – Т.29. – №2. – С. 108–117.

4. Васильев, А.В. Стратегии сохранения биоразнообразия: региональный аспект / А.В. Васильев, В.М. Васюков, Т.Д. Зинченко [и др.] // Самарская Лука: проблемы региональной и глобальной экологии. – 2021. – №3. – С. 5–22.

5. Сахнюк, В.И. Применение методов машинного обучения в обработке данных геофизических исследований скважин отложений викуловской свиты / В.И. Сахнюк, Е.В. Новиков, А.М. Шарифуллин [и др.] // Георесурсы. – 2022. – Т. 24. – № 2. – С. 230–238.

6. Корецкий, Н.А. Оценка подверженности территории центрального черноземья к развитию овражной эрозии с применением методов машинного обучения / Н.А. Корецкий, А.С. Горбунов, В.Н. Бевз // Географическая среда и живые системы. – 2025. – №2. – С. 74–91.

7. Попов, Е.В. Распознавание и классификация изображений природных ресурсов на спутниковых снимках с помощью нейросетевого алгоритма / Е.В. Попов, П.В. Юрченко // Вестник ЮУрГУ. Серия: Строительство и архитектура. – 2025. – №2. – С. 72–82.

8. Брыкин, В.В. Анализ состояния растений с применением технологий искусственного интеллекта / В.В. Брыкин, М.Я. Брагинский, И.О. Тараканова, Д.В. Тараканов // Вестник кибернетики. – 2022. – №4 (48). – С. 6–11.