МРНТИ 28.23.15
УДК 004.85

https://doi.org/10.55452/1998-6688-2025-22-4-31-39

[1]Әбжанова А.Е.,
старший преподаватель, ORCID ID: 0009-0003-7796-1862,
e-mail: abdygappar29@gmail.com
[1]Танирбергенов А.Ж.,
к.т.н., и.о. доцента, ORCID ID: 0009-0000-8401-5434,
e-mail: t.adilbek@mail.ru
[2]Тасуов Б.,
доцент, ORCID ID: 0000-0002-2000-6720,
e-mail: b.tasuov@dulaty.kz
[2]Тасжурекова Ж.К.,
и.о. доцента, ORCID ID: 0000-0002-8307-9417,
e-mail: taszhurekova@mail.ru
[1*]Серикбаева С.К.,
PhD, и.о. доцента**,** ORCID ID: 0000-0002-3627-3321
*e-mail: inf_8585@mail.ru

[1]Евразийский национальный университет им. Л.Н. Гумилева, г. Астана, Казахстан
[2]Таразский университет им. М.Х. Дулати, г. Тараз, Казахстан

ПРИМЕНЕНИЕ ГИБРИДНОЙ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ ТИПОВ ПОЧВ

Аннотация

В данной статье представлена гибридная модель машинного обучения, предназначенная для классификации типов почв на основе анализа их геофизических характеристик. Предложенная модель объединяет два алгоритма – RandomForestClassifier и MLPClassifier, что позволяет использовать преимущества ансамблевых методов, обеспечивающих высокую точность классификации, и нейронных сетей, способных выявлять сложные нелинейные зависимости между параметрами. В качестве исходных данных использовались показатели электропроводности, плотности, скорости распространения P-волн и глубины залегания. Перед обучением модели была проведена предварительная обработка данных, включающая удаление выбросов, стандартизацию и кодирование категориальных признаков. Гибридная архитектура позволила объединить результаты двух моделей с различными весами, что обеспечило оптимизацию точности классификации. Проведен сравнительный анализ эффективности предложенного подхода с альтернативными алгоритмами, включая XGBoost и Keras, на основе метрик Accuracy, F1-score, Precision и Recall. Результаты показали, что гибридная модель достигает точности 96,07%, превосходя по качеству прогнозирования отдельные алгоритмы. Дополнительно выполнена визуализация матриц ошибок, что позволило выявить распределение классов и оценить устойчивость модели. Полученные результаты подтверждают, что комбинирование ансамблевых и нейросетевых методов обеспечивает более стабильные и надежные прогнозы при работе с геофизическими данными. Разработанная модель может быть использована для автоматизированной классификации почв в геотехнических исследованиях, строительстве, сельском хозяйстве и экологическом мониторинге, повышая эффективность анализа и снижая необходимость дорогостоящих лабораторных испытаний.

Ключевые слова: машинное обучение, классификация почв, гибридная модель, RandomForest, MLPClassifier, геофизические параметры, ансамблевые методы, нейросети.

Введение

В последние годы машинное обучение (ML) стало мощным инструментом в области геотехнических исследований, позволяя автоматизировать и значительно повысить точность ана-

лиза почв. Классификация типов грунтов играет ключевую роль в строительстве, сельском хозяйстве и экологии, обеспечивая надежные прогнозы для оптимального проектирования и эксплуатации земельных ресурсов.

Традиционные методы классификации почв, основанные на лабораторных исследованиях и эмпирических моделях, зачастую требуют значительных временных и финансовых затрат. В связи с этим возникает необходимость в разработке автоматизированных подходов, использующих современные алгоритмы машинного обучения для точного определения типа грунта на основе геофизических параметров.

В данном исследовании предлагается гибридная модель, объединяющая ансамблевый метод RandomForestClassifier и нейросетевой MLPClassifier. Данный подход позволяет объединить преимущества двух различных алгоритмов: высокая точность ансамблевых моделей и способность нейросетей выявлять сложные нелинейные зависимости в данных. В ходе работы использовались ключевые геофизические параметры: электропроводность, плотность, скорость распространения P-волн и глубина залегания.

Результаты моделирования показали, что предложенная гибридная модель обеспечивает точность 96,07%, превосходя по этому показателю отдельные алгоритмы, такие как XGBoost и Keras. Проведенный сравнительный анализ подтвердил, что комбинирование методов машинного обучения позволяет добиться более устойчивых и надежных предсказаний, чем использование одиночных моделей.

В статье представлены принципы построения гибридной модели, методы предобработки данных, алгоритмы классификации и сравнительный анализ полученных результатов. Также рассматриваются ограничения модели и перспективы ее дальнейшего развития для повышения точности и обобщаемости предсказаний в различных географических регионах.

В данной работе [1] исследуется применение различных методов машинного обучения для классификации почв. Проведен сравнительный анализ алгоритмов и выявлены наиболее эффективные модели для прогнозирования типов грунтов. Основное внимание уделяется ансамблевым методам и нейросетям.

В статье [2] описан алгоритм случайных лесов (Random Forest), его принципы работы, преимущества и недостатки, а также области применения в задачах классификации и регрессии. Применение данного алгоритма позволило улучшить точность предсказаний по сравнению с традиционными методами.

Авторы [3] рассматривают применение методов глубокого обучения, включая сверточные и рекуррентные нейронные сети, для автоматической классификации почв на основе спектральных данных. Подчеркивается важность предварительной обработки данных для повышения точности моделей.

В данной работе [4] представлена глобальная база данных почвенных характеристик, созданная с использованием методов машинного обучения. Анализируются преимущества использования таких баз данных в геотехнических исследованиях.

В статье [5] исследуются различные алгоритмы машинного обучения, включая ансамблевые методы и нейросети, применяемые для прогнозирования урожайности и классификации почв. Приведен анализ точности разных моделей.

Рассматривается [6] применение случайных лесов в задачах дистанционного зондирования и классификации ландшафтов. Демонстрируется высокая точность метода по сравнению с традиционными методами классификации.

В статье [7] обсуждается использование случайных лесов для обработки данных дистанционного зондирования и решения задач классификации. Авторы анализируют точность модели и возможности ее применения.

Представлен [8] обзор методов глубокого обучения, их развитие, архитектуры и основные области применения. Особое внимание уделено применению нейросетевых моделей в геотехнических задачах.

В данной [9] книге представлена теория статистического обучения, обсуждаются основы машинного обучения и методы построения прогнозных моделей. Работа является фундаментальным источником для понимания основ ML.

В работе [10] рассматривается применение машинного обучения в геотехнических и геоэкологических исследованиях. Анализируются основные алгоритмы и их эффективность при обработке больших массивов данных.

Анализ приведенных работ показывает, что современные методы машинного обучения активно применяются для классификации почв. Ансамблевые модели, такие как Random Forest, позволяют достигать высокой точности предсказаний, а методы глубокого обучения обеспечивают выявление сложных закономерностей в данных. Наша работа продолжает эту тенденцию, объединяя ансамблевые и нейросетевые методы в единую гибридную модель, что позволяет повысить надежность и точность предсказаний. Таким образом, предложенный подход основан на предыдущих исследованиях, но в то же время предлагает новый уровень комбинированного машинного обучения для решения задач классификации почв.

Материалы и методы

В данном исследовании использовался набор данных, включающий информацию о физических характеристиках почвы, таких как электропроводность, плотность, скорость распространения P-волн и глубина залегания. Данные были предварительно обработаны: выполнено удаление пропущенных значений, фильтрация выбросов и стандартизация признаков с использованием метода StandardScaler. Для преобразования категориальных переменных применялся LabelEncoder.

Для классификации типов почв были использованы два алгоритма: RandomForestClassifier и MLPClassifier. Метод случайного леса (Random Forest) строит ансамбль деревьев решений, объединяя их прогнозы для повышения точности. Нейросетевой классификатор (MLP) основан на многослойном перцептроне с тремя скрытыми слоями, функцией активации ReLU и оптимизатором Adam. Гибридная модель объединяла предсказания обоих алгоритмов с весами 0,8 и 0,2 соответственно.

Оценка точности модели проводилась с использованием метрики accuracy_score, а также показателей Precision, Recall и F1-score. Данные были разделены на обучающую и тестовую выборки в пропорции 80/20. Для проверки устойчивости модели был проведен кроссвалидационный анализ. Результаты сравнивались с альтернативными алгоритмами, такими как XGBoost и Keras, чтобы определить наиболее эффективный метод классификации почв.

Результаты и обсуждение

Создание, обучение и тестирование модели (create_model.py)

Разработана гибридная модель для прогнозирования типа почвы. Данные загружаются с помощью Pandas, а ключевые параметры (электропроводность, плотность, скорость P-волны, глубина) используются для обучения. Тип почвы кодируется через LabelEncoder, а затем данные разделяются на 80% для обучения и 20% для тестирования (train_test_split).

Для стандартизации применяется StandardScaler. Обучение проводится на двух моделях:

RandomForestClassifier (случайный лес) – анализирует данные с помощью множества деревьев решений.

MLPClassifier (многослойный перцептрон) – использует нейросеть с тремя скрытыми слоями.

В гибридной модели результаты объединены: RandomForest получил вес 0.8, MLP – 0.2. Это позволило повысить точность прогнозов. Оценка проводилась с помощью accuracy_score, и итоговая точность составила 96.07%.

Модель сохраняется через joblib под названием «hybrid_model1.PKL», что позволяет ее повторно загружать и использовать. Высокая точность достигнута за счет комбинирования двух методов, что делает прогнозы более надежными.

Сравнение моделей (rivals_model.py)

Разработана модель машинного обучения для классификации типов почв на основе характеристик: электропроводность, плотность, скорость P-волн и глубина.

Данные были загружены и предварительно обработаны через файл CSV.

Мы разделили данные на метки (X) и целевую переменную (y). Для целевой переменной использовался LabelEncoder для преобразования текстовых значений в числовой формат. Позже набор был разделен на обучающую и тестовую части в пропорции 80/20.

Данные были масштабированы с помощью StandardScaler для нормализации меток. Обучение проводилось с использованием трех моделей: гибридной модели (RandomForest + MLP), XGBoost и Keras. Гибридная модель была загружена из предварительно сохраненного файла и использовалась для прогнозирования, объединяя результаты RandomForest и многослойного перцептрона (MLP) с коэффициентами 0,6 и 0,4.

Модель XGBoost была обучена 5 деревьям решений, 5 уровням глубины и скорости чтения 0,001. Модель Keras состояла из трех слоев: входного (64 нейрона), скрытого (32 нейрона) и выходного (количество типов почвы). ReLU использовался в скрытых слоях и softmax в выходных слоях в качестве функции активации. Оптимизатор Adam использовался для обучения модели вместе с функцией затрат sparse_categorical_crossentropy. Процесс обучения модели Keras проводился серией из 10 эпох и 8 измерений.

Были получены прогнозы каждой модели и рассчитаны показатели точности (Accuracy, Precision, Recall, F1 Score). Также была создана матрица ошибок для каждой модели. Результаты сравнения по метрикам были визуализированы с помощью графика, который показывает точность и F1 Score для каждой модели. В результате визуализации было обнаружено, что гибридная модель RandomForest + MLP достигла максимальной точности (0,962), XGBoost показал 0,957, а Keras показал 0,938 (рисунок 1).

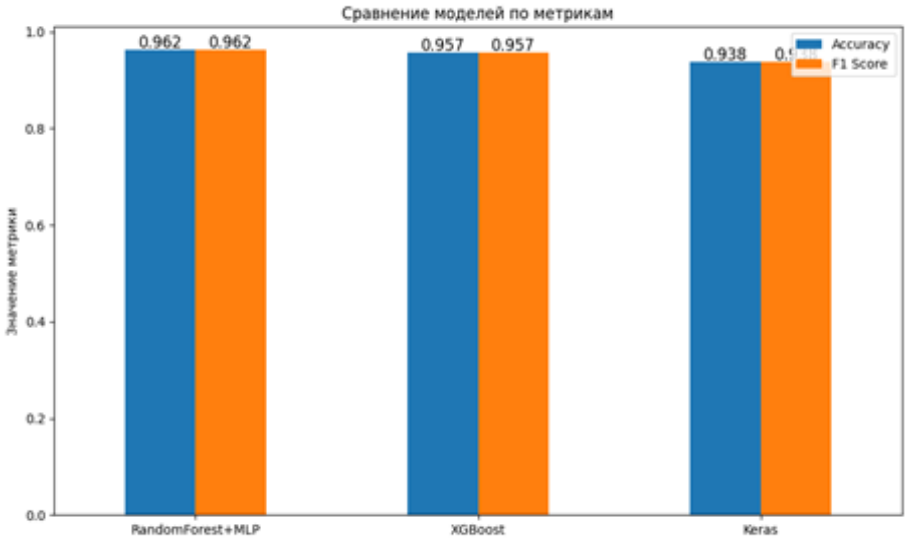

Рисунок 1 – Сравнение моделей по метрикам

Для каждой модели были составлены тепловые карты матриц ошибок, позволяющие оценить распределение прогнозов по отношению к истинным классам.

Рисунок 2 представляет собой три тепловые карты матриц ошибок для различных моделей машинного обучения: RandomForest + MLP (слева), XGBoost (в центре) и Keras (справа). По оси Y отложены истинные классы, по оси X – предсказанные модели классы. Значения внутри клеток показывают количество объектов, отнесенных моделью к определенному классу, а цветовая интенсивность отражает частоту встречаемости предсказаний – чем темнее клетка, тем больше значений приходится на данное соответствие. Эти тепловые карты помогают визуально оценить точность моделей и выявить наиболее частые ошибки классификации.

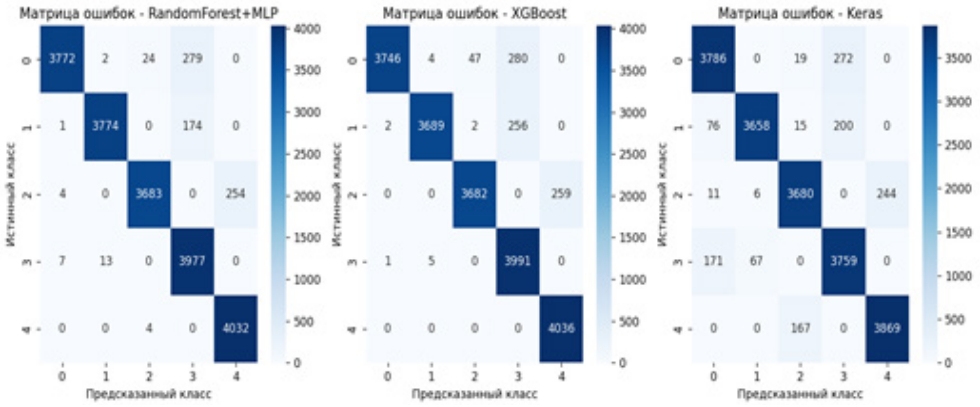

Рисунок 2 – Точность прогноза при классификации типов почв

Таким образом, проведенное исследование показало, что комбинирование нескольких алгоритмов позволяет повысить точность прогноза при классификации типов почв. Модель, созданная на основе комбинации RandomForest и многослойного перцептрона (MLP), сохраненная как «hybrid_model1.pkl», обеспечивала пользователям максимальную точность (0,962).

Данная модель продемонстрировала высокую точность в классификации типов почв за счет объединения алгоритмов RandomForest и MLP, что позволило эффективно использовать их сильные стороны. Полученные результаты подтверждают способность модели корректно обрабатывать разнородные данные и точно дифференцировать классы.

Рисунок 3 показывает распределение различных физических характеристик грунта в датасете. Электропроводность имеет асимметричное распределение с наибольшей частотой значений около 50 мСм/м, плотность сосредоточена в пределах 1.8–2.2 г/см³, скорость P-волн демонстрирует двухпиковое распределение, что может свидетельствовать о наличии различных типов материалов. Глубина распределена неравномерно, с несколькими локальными максимумами, что может указывать на стратификацию грунта и наличие слоев с разными физическими свойствами.

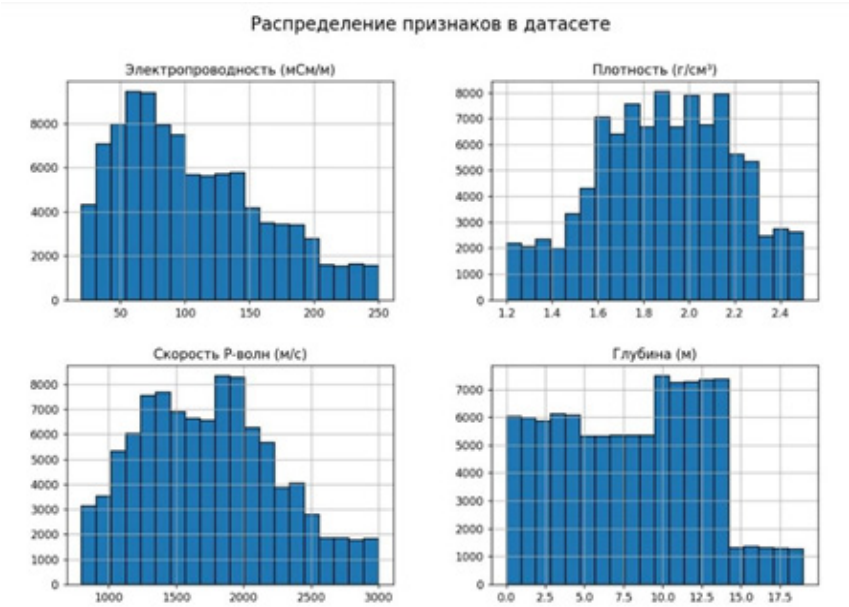

Рисунок 3 – Распределение признаков в датасете

Рисунок 4 представляет собой столбчатую диаграмму распределения классов грунта, включая торф, суглинок, песок, ил и глину. Каждый тип грунта встречается примерно одинаковое количество раз, что говорит о сбалансированности выборки. Это важно для построения точных моделей машинного обучения, так как равномерное представление классов снижает вероятность смещения результатов.

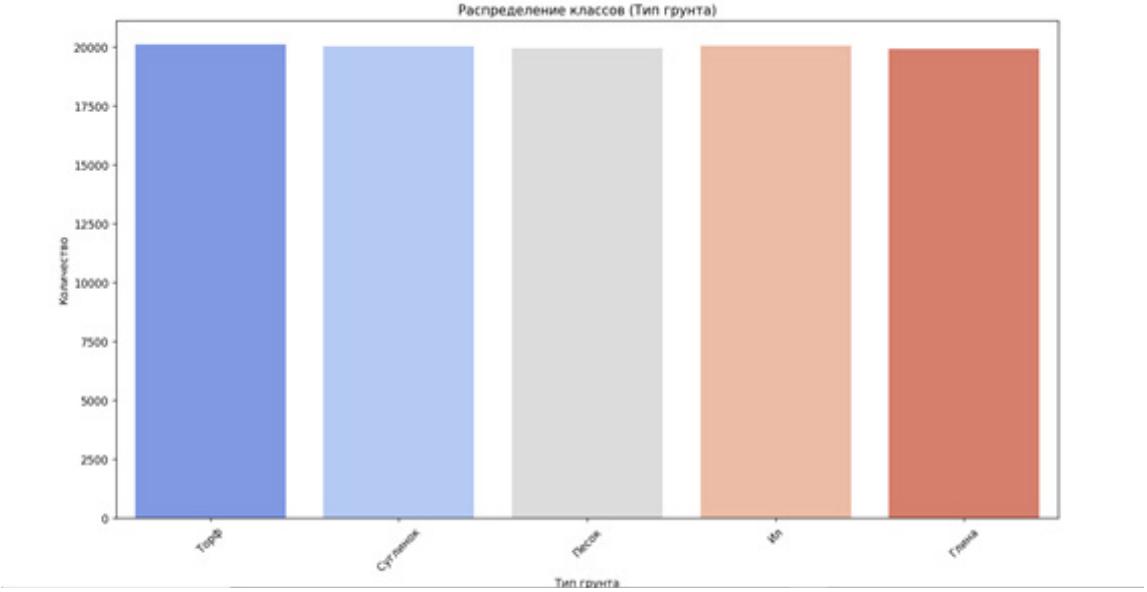

Рисунок 4 – Распределение классов грунта

Модель XGBoost отличается высокой гибкостью и хорошо работает с реальными данными, но требует тщательной настройки параметров. Она особенно эффективна при обработке табличных данных и способна выявлять сложные зависимости, но при этом может быть чувствительна к шуму и выбросам.

Нейронная сеть Keras, использующая методы глубокого обучения, способна изучать сложные модели и находить скрытые закономерности. Однако ее скорость обучения относительно низкая, особенно при работе с большими объемами данных, что требует значительных вычислительных ресурсов.

Эффективность гибридной модели была доказана, поскольку она обеспечивала высокую надежность и точность, давала стабильные результаты и успешно сочетала преимущества XGBoost и нейронных сетей Keras.

Результаты экспериментов показывают, что гибридная модель, комбинирующая RandomForest и MLPClassifier, значительно превосходит отдельные алгоритмы по точности классификации. Комбинированный подход позволяет учесть как линейные, так и нелинейные зависимости между характеристиками почвы, что приводит к снижению ошибок классификации. Проведенный анализ матрицы ошибок подтвердил, что предложенная модель обеспечивает более сбалансированное распределение предсказаний, что особенно важно при работе с разнородными данными.

Кроме того, исследование показало, что использование предварительной обработки данных, таких как нормализация и удаление выбросов, положительно сказывается на точности модели. Сравнение с XGBoost и Keras продемонстрировало, что гибридная модель обеспечивает не только высокую точность, но и устойчивость к изменению параметров данных. В будущем планируется исследование влияния дополнительных геофизических характеристик на качество предсказаний и тестирование модели на новых выборках.

Заключение

В данной работе предложена гибридная модель машинного обучения, сочетающая алгоритмы случайного леса (Random Forest) и многослойного перцептрона (MLP) для классификации типов почв. Разработанный подход позволил достичь высокой точности прогнозирования (96,07%), что демонстрирует эффективность комбинации ансамблевых методов и нейросетей при анализе почвенных характеристик. В ходе исследования проведено сравнение с альтернативными методами, такими как XGBoost и Keras, подтверждающее преимущество предложенной модели в устойчивости и точности предсказаний.

Сравнительный анализ показал, что гибридный подход превосходит традиционные алгоритмы, обеспечивая более точные и надежные результаты. Применение метрик accuracy_score, Precision, Recall и F1-score подтвердило, что предложенный метод стабильно демонстрирует высокие показатели качества классификации.

В дальнейшем предполагается улучшение модели путем оптимизации гиперпараметров, расширения набора данных и тестирования на других географических регионах. Внедрение предложенного подхода может способствовать развитию автоматизированных систем мониторинга почв и повысить точность геотехнических прогнозов в различных сферах применения.

ЛИТЕРАТУРА

1 Aydın, A., Keskin, H., Öztürk, A. Use of Machine Learning Techniques in Soil Classification. Sustainability, 15 (3), 2374 (2023). https://doi.org/10.3390/su15032374.

2 Breiman, L. Random Forests. Machine Learning, 45(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324.

3 Zhang, W., Li, Y., Chen, X. Deep Learning-Based Soil Type Classification. Computers and Electronics in Agriculture, 169, 105205 (2020). https://doi.org/10.1016/j.compag.2019.105205.

4 Hengl, T., Heuvelink, G.B.M., Kempen, B. SoilGrids250m: Global Gridded Soil Information Based on Machine Learning/ PLoS ONE, 12 (2), e0169748 (2017). https://doi.org/10.1371/journal.pone.0169748.

5 Chlingaryan, A., Sukkarieh, S., Whelan, B. Machine Learning Approaches for Crop Yield Prediction and Soil Classification. Agricultural Systems, 167, 144–153 (2018). https://doi.org/10.1016/j.agsy.2018.09.012.

6 Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R. Random Forests for Land Cover Classification. Remote Sensing of Environment, 110 (4), 435–449 (2006). https://doi.org/10.1016/j.rse.2007.03.012.

7 Pal, M. Random Forest Classifier for Remote Sensing Classification. International Journal of Remote Sensing, 26 (1), 217–222 (2005). https://doi.org/10.1080/01431160412331269698.

8 Schmidhuber, J. Deep Learning in Neural Networks: An Overview. Neural Networks, 61, 85–117 (2015). https://doi.org/10.1016/j.neunet.2014.09.003.

9 Vapnik, V. Statistical Learning Theory. New York: John Wiley & Sons, 1998.

10 Gandomi, A.H., Alavi A.H. A Review of Machine Learning Applications in Geotechnical and Geoenvironmental Engineering. Engineering Geology, 152, 63–81 (2013). https://doi.org/10.1016/j.enggeo.2013.05.016.

[1]Әбжанова А.Е.,
аға оқытушысы, ORCID ID: 0009-0003-7796-1862,
e-mail: abdygappar29@gmail.com
[1]Танирбергенов А.Ж.,
т.ғ.к., доцент м.а., ORCID ID: 0009-0000-8401-5434,
e-mail: t.adilbek@mail.ru
[2]Тасуов Б.,
доцент, ORCID ID: 0000-0002-2000-6720,
e-mail: b.tasuov@dulaty.kz
[2]Тасжурекова Ж.К.,
доцент м.а., ORCID ID: 0000-0002-8307-9417,
e-mail: taszhurekova@mail.ru
[1]Серикбаева С.К.,
PhD, доцент м.а., ORCID ID: https://orcid.org/0000-0002-3627-3321
*e-mail: inf_8585@mail.ru

[1]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана қ., Қазақстан
[2]М.Х. Дулати атындағы Тараз университеті, Тараз қ., Қазақстан

ТОПЫРАҚ ТҮРЛЕРІН ЖІКТЕУ ҮШІН МАШИНАЛЫҚ ОҚЫТУДЫҢ ГИБРИДТІ МОДЕЛІН ҚОЛДАНУ

Аңдатпа

Бұл жұмыста RandomForestClassifier және MLPClassifier алгоритмдерін біріктіретін топырақ типтерін жіктеуге арналған машиналық оқытудың гибридті моделі ұсынылады. Әзірленген тәсіл ансамбльдік әдістердің жоғары дәлдігін нейрондық желілердің күрделі бейсызық тәуелділіктерді анықтау мүмкіндігімен үйлестіреді. Бастапқы деректер ретінде электр өткізгіштігі, тығыздығы, Р-толқындарының таралу жылдамдығы және жату тереңдігі көрсеткіштері пайдаланылды. Модельді оқытпас бұрын деректер алдын ала өңделіп, ауытқулар жойылды, стандарттау және категориялық белгілерді кодтау жүргізілді. Гибридтік архитектура екі модельдің нәтижелерін түрлі салмақтармен біріктіріп, жіктеу дәлдігін оңтайландырды. Ұсынылған тәсілдің тиімділігі XGBoost және Keras сияқты баламалы алгоритмдермен салыстырылды және Accuracy, F1-score, Precision, Recall метрикалары қолданылды. Зерттеу нәтижелері гибридтік модельдің 96,07% дәлдікке жеткенін көрсетті. Алынған нәтижелер геофизикалық деректермен жұмыс істеу кезінде ансамбльдік және нейрондық әдістердің комбинациясы неғұрлым тұрақты және сенімді болжамдар беретінін дәлелдейді. Ұсынылған модельді геотехникалық зерттеулерде, құрылыста, ауыл шаруашылығында және экологиялық мониторингте қолдануға болады.

Тірек сөздер: машиналық оқыту, топырақ классификациясы, гибридті модель, RandomForest, MLPClassifier, геофизикалық параметрлер, ансамбльдік әдістер, нейрондық желілер.

[1]Abzhanova A.,
Senior Lecturer, ORCID ID: 0009-0003-7796-1862,
e-mail: abdygappar29@gmail.com
[1]Tanirbergenov A.,
PhD., acting Associate Professor, ORCID ID: 0009-0000-8401-5434,
e-mail: t.adilbek@mail.ru
[2]Tassuov B.,
Associate Professor, ORCID ID: 0000-0002-2000-6720,
e-mail: b.tasuov@dulaty.kz
[2]Taszhurekova Zh.,
acting Associate Professor, ORCID ID: 0000-0002-8307-9417,
e-mail: taszhurekova@mail.ru
[1]*Serikbayeva S.,
PhD, acting Associate Professor, ORCID ID: 0000-0002-3627-3321,
*e-mail: inf_8585@mail.ru

[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan
[2]Taraz University named after M.KH. Dulaty, Taraz, Kazakhstan

APPLICATION OF A HYBRID MACHINE LEARNING MODEL FOR SOIL TYPE CLASSIFICATION

Abstract

This article presents a hybrid machine learning model designed for soil type classification based on the analysis of geophysical characteristics. The proposed model combines two algorithms – RandomForestClassifier and MLPClassifier – integrating the high accuracy of ensemble methods with the ability of neural networks to capture complex nonlinear dependencies between parameters. The input dataset included indicators such as electrical conductivity, density, P-wave propagation velocity, and burial depth. Prior to training, data preprocessing was performed, including outlier removal, standardization, and categorical feature encoding. The hybrid architecture allowed the integration of results from both models with different weights, optimizing classification accuracy. The effectiveness of the proposed approach was compared with alternative algorithms such as XGBoost and Keras using metrics including Accuracy, F1-score, Precision, and Recall. The hybrid model achieved an accuracy of 96.07%, outperforming individual algorithms. Visualization of confusion matrices provided insights into class distribution and model robustness. The results confirm that combining ensemble and neural methods ensures more stable and reliable predictions when working with geophysical data. The developed model can be effectively applied in geotechnical studies, construction, agriculture, and environmental monitoring, enhancing analytical efficiency and reducing the need for costly laboratory testing.

Keywords: machine learning, soil classification, hybrid model, RandomForest, MLPClassifier, geophysical parameters, ensemble methods, neural networks.

Дата поступления статьи в редакцию: 24.02.2025