АННОТАЦИЯ
диссертации на соискание степени доктора философии (PhD)
по специальности 6D060200 – Информатика
НУРЖАНОВА ЧИНГИЗ АСКАРОВИЧА
Проектирование информационной системы для прогнозирования и принятия решений в процессе очистки почвы, содержащей токсичные элементы

Актуальность проблемы. Компьютерная (информационная) технология открыла широкие возможности для изучения процессов, происходящих в природе. С помощью методов математического моделирования и вычислительной техники на основе прорывных технологий создаются новые методы, модели, алгоритмы и технологии для решения глобальных экологических проблем взаимодействия человека и природы. В настоящее время машинное обучение (МО), используя алгоритмы статистических методов регрессии и классификации, находит применение на современном уровне в разных сферах науки, например, в области информационной науки, экологии, сельского хозяйства. Широкомасштабная химизация в сельском хозяйстве привела не только к загрязнению, но и к снижению урожайности сельскохозяйственных культур, повышение рентабельности растениеводства зависит от экологичности производимой продукции. В связи с этим, цифровизация сельского хозяйства для отслеживания урожайности сельскохозяйственных культур в режиме реального времени, производства экологически чистой продукции занимает ведущее место в аграрном секторе.

Для обеспечения экологической безопасности разрабатываются информационные системы экологического мониторинга, за счет сбора больших данных о состоянии среды Эффективность систем мониторинга в значительной степени определяется используемыми информационными технологиями для непрерывного сбора, обработки и хранения данных традиционно решаются с использованием систем управления базами данных (БД). В Казахстане отсутствует единая компьютерная информационная БД учета уровня загрязнения антропогенно-нарушенных экосистем и их отходов. Важность создания единой БД позволит правительству страны решать стратегически важные вопросы, связанные с экологией, например, строительство высокотехнологичных сооружений для сжигания отходов промышленного производства, а главное оценить реальный уровень техногенного загрязнения, возможный риск для окружающей среды и здоровья населения.

Интерес к математическому моделированию процесса рекультивации загрязненных токсичными элементами (ТЭ) экосистемы растет из года в год, в связи с увеличением масштабов загрязнения окружающей среды, вызванного антропогенной деятельностью. Создаются алгоритмы для быстрой нейтрализации источника загрязнения и математические модели, которые позволяют найти оптимальное решение, адекватное описание процесса загрязнения почв, прогнозируют последствия нарушения почвенных процессов, их влияния на растительный и животный мир, выбор оптимальной стратегии рекультивации.

Системный анализ с привлечением математического моделирования направлены на моделирование процесса восстановления загрязненных земель, в основном углеводородов. При этом нет теории и модели, описывающих поведение других основных загрязнителей среды в почвенно-грунтовой толще, процесс их миграции в системе «почва-растение», которые могли бы стать основой разработки технологии очистки территорий, загрязненных токсичными элементами (ТЭ). В связи с этим, разработка информационной системы для технологии восстановления загрязненных ТЭ территорий страны, создание единой БД о загрязненных территориях и БД о растениях, способных к их восстановлению, станет основой для решения важных экологических проектов и задач в области охраны окружающей страны, улучшения благосостояния населения, проживающего в экологически опасных регионах.

Цель диссертационного исследования – разработка интеллектуальной информационной системы для обработки данных о почвах, загрязненных токсичными элементами для прогнозирования и принятия решений об очистке земель Республики Казахстан.

Задачи исследования

1. Изучение математического моделирования продуктивности биоэнергетического растения на почвах, загрязненных токсичными элементами.

2. Создание баз данных: о древесных и травянистых видах растений, способных к восстановлению почв; о землях, загрязненных токсичными элементами, с учетом географического месторасположения территории; о количестве устаревших пестицидов и концентрации их в почве.

3. Разработка информационной системы для прогнозирования и принятия решений в процессе очистки почвы от токсичных элементов.

4. Применение методов машинного обучения для прогнозирования продуктивности растений, поглощающих токсичные элементы из почвы.

Методы исследования: методы машинного обучения, многорядный эвристический метод самоорганизации для построения регрессионных уравнений, методы моделирования, методы регрессионного и дисперсионного многофакторного анализа. При разработке комплексного подхода к созданию информационной системы были применены теория проектирования информационных систем, методы проектирования баз данных, процессно-ориентированный подход. Объектом исследования являются данные о почвах и растениях, загрязненных токсичными элементам, о климатических условиях (на примере Алматинской области). Предметом исследования являются климатические данные AlmatyWeatherDataSet.csv, данные за 2015–2022 гг. о продуктивности растений на почвах, загрязненных токсичными элементами.

Основные результаты исследования

1. Адаптирован «Многорядный эвристический метод самоорганизации», предсказывающий биомассу растений в зависимости от окружающих условий. Наибольшее влияние на этот процесс оказывают три фактора: испарение влаги почвы, фотосинтетическая активная радиация и осадки.

1.1 Усовершенствована модель Miscancalc на основе Miscanmod для прогнозирования биомассы растений на загрязненной почве с учетом климатических данных путем расчета коэффициента разности между загрязнённой и чистой почвой.

2. Созданы два организованных хранилища данных: о древесных и травянистых видах растений, способствующие к восстановлению почвы; о количестве устаревших пестицидов и их концентрации в почве.

3. Разработана информационная система для прогнозирования и принятия решений в процессе очистки почвы от ТЭ. Определена длительность периода очистки, затраченного для каждого токсичного элемента.

4. Применены интеллектуальные методы машинного обучения и определен наилучший из них для прогнозирования концентраций загрязнения почвы. По результатам исследования XGB Regressor имеет самые низкие метрики R2 = 0.998; MSE = 421.19; MAE = 15.03; MAPE = 0.065.

4.1 Исследованы ансамблевые методы машинного обучения с целью прогнозирования производительности растений на основе климатических данных. Процесс анализа осуществлен через инструментальное программное обеспечение JupyterLab в среде Anaconda.

4.2 Проведена оценка моделей регрессии на урожайность в зависимости от климатических условий. С помощью подхода SHAP отобрано 13 информативных признаков, ответственных за продуктивность, высокая степень влияния оказывает признаки datetime_1 (месяцы) и datetime_2 (дни).

4.3 Установлена длительность периода очистки почвы 1 га почвы с глубины залегания 0-20 см с помощью растения мискантус для разных элементов. Для особо токсичных элементов: свинец - 12 лет, цинк - 7 лет.

Обоснование новизны и важности полученных результатов:

Научная новизна заключается в разработке и получении следующих выводов:

Предложена усовершенствованная модель Miscancalc на основе Miscanmod для прогнозирования биомассы растений на загрязненной ТЭ почве с учетом климатических данных.

Адаптирован метод многорядной самоорганизации, по прогнозирующим свойствам превосходящий в точности регрессионные модели, обеспечивающий автоматический отбор информативных входных переменных и выбор структуры регрессионной модели оптимальной сложности.

Расширена модель рекультивации почвы на основе методов машинного обучения с помощью интегрированного подхода библиотеки XGBoost, обладающей высокой производительностью и устойчивостью к переобучению.

Важность полученных результатов: Алгоритмы для прогнозирования биомассы растений на загрязненной ТЭ почве с учетом климатических данных обладают высокой математической точностью. Предложенный многорядный алгоритм самоорганизации по своим прогнозирующим свойствам превосходит регрессионные модели, обеспечивает автоматический отбор информативных входных переменных и выбор структуры регрессионной модели оптимальной сложности.

Теоретическая и практическая значимость результатов – фундаментальные аспекты информационной технологии и математического моделирования в задачах природопользования. Прикладная ценность – в оперативности передачи информации государственным органам, осуществляющим управление земельными фондами, при формировании рекультивационных мероприятий; в возможности использования опытными хозяйствами, занимающимися внедрением информационных технологий, а также организациям, осуществляющих агроэкологический мониторинг.

Соответствие направлениям развития науки или государственным программам (проекты) Диссертационная работа были выполнена в рамках программы на базе Института информационной и вычислительной технологии КН МНВО РК и гранта № AP19678926 «Разработка интеллектуальной системы для исследования и решения экологических проблем загрязнения почвы и воздуха с помощью методов науки о данных» (2023-2025 годы). Все результаты диссертационной работы, которые вынесены на защиту, выполнены и собраны докторантом лично автором. Кроме того, основные результаты исследований, анализы, модели, программы и информационные системы созданы докторантом.

Среди основных результатов: приложение MiscanCalc, как новое приложение модели MiscanMod, для прогнозирования урожайности растения на загрязненной ТЭ и незагрязненной почве в зависимости от климатических условий; оценка 13 моделей регрессии машинного обучения на урожайность в зависимости от климатических условий; создание базы о загрязненных ТЭ территориях и базы данных растениях казахстанской флоры, способных к очистке, загрязненной различными вредными веществами почвы; создание информационной системы для прогнозирования и принятия решений в процессе очистки почвы от ТЭ. Разработка многорядного алгоритма самоорганизации для анализа связи динамики биомассы растения с климатическими условиями среды проведена под руководством доктора физико-математических наук, профессора Т. Ж. Мазакова.

Основные положения диссертации и результаты исследования докладывались и обсуждались на научных семинарах кафедры Компьютерные науки факультета Информационных технологий Казахского национального университета имени аль-Фараби; на ученом совете Института информационной и вычислительной технологии КН МНВО РК; Международной Научной конференции ИИВТ «Современные проблемы информатики и вычислительных технологий». 28–29 июня 2016; II International Conference on Modern Problems of Computer Science and Computer Technology, MSHE RK, September 27–30, 2017; 15th International Phytotechnology conference. October 1–5, 2018. University of Novi Sad, Serbia; Международной научной конференции в области информационных технологий, посвященной 75-летию профессора У. А. Тукеева, 8 октября 2021 года и других международных конференциях.

Описание вклада докторанта в подготовку каждой публикации.

В опубликованных статьях и научных трудах описаны результаты исследования по теме диссертации. За время научной работы было написано 17

научных работ, в том числе 4 статьи в журналах, рекомендованных Комитетом по Контролю в Сфере Образования и Науки Министерства науки и высшего образования Республики Казахстан; 8 публикации в материалах международных конференций, 5 статьи в журналах, входящих базу данных Thomson Reuters и Scopus.

Структура и объем диссертации. Диссертация состоит из обозначения и сокращения, введения, пяти глав, заключения, списка использованных источников и приложения. Полный объём диссертации составляет 191 страницу, 62 рисунка и 18 таблиц и 3 приложения. Список литературы содержит 268 наименований.

Во введении обосновывается актуальность диссертации. Сформулированы цель работы, объект и предмет исследования. Выявлена научная новизна и практическая значимость. Описаны результаты исследования. Приводится информация об апробации результатов исследования и публикации.

В первой главе диссертационной работы рассматривается литературный обзор об информационной системе экологического мониторинга факторов среды на базе технологий хранилищ данных, методах моделирования в области экологии, сельского хозяйства и экологической биотехнологии.

Во второй главе для оценки продуктивности растений (на примере биоэнергетического, сельскохозяйственного вида Мискантус), произраставших на ТЭ-загрязненной в зависимости от климатических условий (на примере Алматинской области) рассмотрены три подхода: 1) использование алгоритмов 13-и регрессионных моделей МО; 2) создание приложения модели MiscanCalc на основе модификации модели MiscanMod; 3) разработка математической модели «Многорядный эвристический метод самоорганизации для построения регрессионных уравнений».

В третьей главе представлены результаты разработки комплексного подхода к созданию информационной системы на базе данных о загрязненных токсичными элементами территориях и базе данных растениях, способных к восстановлению техногенных ландшафтов. Приведены функциональные требования, этапы создания программы и их построение. Спроектирована информационная система для формирования отчетности об образовании, утилизации и хранения отходов и рекультивации почв.

В четвертой главе представлены данные о создании математической модели, описывающие накопление тяжелых металлов в вегетативных органах растений в зависимости от типа почвы и условий среды.

В пятой главе представлены данные о создании информационной системы для прогнозирования процесса очистки 1 га почвы с глубины залегания 0–20 см, с помощью растений от ТЭ. В основу концепции информационной системы легло интегрированный подход включая: сбор, передачу, накопление и обработка измерительных данных, данные БД**,** модели продуктивности растений, модели фитотоксичности почвы, модели поглощения ТЭ растением и содержания их в почве.

В заключении сформулированы полученные результаты в диссертации.