КОНКУРСНОЕ ЗАДАНИЕ

ДЛЯ ПЯТОГО ОТРАСЛЕВОГО

ЧЕМПИОНАТА

В СФЕРЕ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

DIGITAL SKILLS 2023

ПО КОМПЕТЕНЦИИ

«МАШИННОЕ ОБУЧЕНИЕ И   
БОЛЬШИЕ ДАННЫЕ»

**С1. Модуль А. Парсинг и предобработка данных**

## СОДЕРЖАНИЕ

Модуль А данного Конкурсного задания состоит из следующей документации / файлов:

1. Две папки с фрагментами спутниковых снимков (поля размечены экспертами): для обучения модели - «train\_new».
2. Приложение 1 к Модулю А.
3. Инструкция к Модулю А.

## ВВЕДЕНИЕ

**«*Единственное средство удержать государство в состоянии независимости от кого либо – это сельское хозяйство. Обладай вы хоть всеми богатствами мира, если вам нечем питаться – вы зависите от других…*» Жан Жак Руссо.**

В связи с непрерывным ростом населения планеты (которое по прогнозам ООН достигнет 8,5 миллиардов человек уже к концу этого десятилетия), чтобы его прокормить, необходимо использовать интенсивные технологии ведения сельскохозяйственной деятельности. Данные технологии характеризуются высокой автоматизацией операций и интенсивным использованием агрохимикатов, что позволяет получать более высокие урожаи на единицу площади пахотных угодий. Но, в свою очередь, создает проблемы, связанные с сохранением плодородия почв, а также борьбой с резистентными сорняками, вредителями и болезнями.

Сельское хозяйство является важной отраслью экономики РФ, которая показывает последние годы стабильный рост и непосредственно влияет на продовольственную безопасность страны. В число наиболее экономически выгодных для выращивания культур в нашей стране входят пшеница, подсолнечник, кукуруза. Под пшеницу отводится самая большая (среди всех культур) доля посевных площадей, составляющая порядка 36%. Доля подсолнечника составляет порядка 12% посевных площадей, а под кукурузу отводится менее 4% посевных площадей.

Одной из актуальных проблем современного сельского хозяйства является борьба с болезнями и вредителями с целью сохранения урожаев сельскохозяйственных культур. Эффективность этой борьбы тем выше, чем раньше удается идентифицировать проблему и выбрать наиболее эффективные средства воздействия на нее (например, пестициды). Раннему выявлению болезней сельскохозяйственных культур может способствовать дистанционный мониторинг и автоматический анализ данных с последующей выдачей рекомендаций.

В рамках данного конкурсного задания предлагается разработать интеллектуальную информационную систему, которая позволит идентифицировать сельскохозяйственные поля, их текущий статус, определить сегменты вероятного заражения сельскохозяйственных культур, а также риск пожарной опасности.

Для обучения модели используется набор изображений и масок из папки «train\_new». В подпапке «images» находятся фрагменты спутниковых снимков (всего более 4х тысяч) с полями. Разрешение снимков - 10м. В подпапке «labels» находятся контуры размеченных полей (разметка выполнена экспертами в ручном режиме).

## ИНСТРУКЦИЯ УЧАСТНИКУ

В рамках данного модуля Вам предстоит загрузить предоставленные разработчиком задания данные, создать дополнительный набор данных. Выполнить предобработку данных.

## ПРАКТИЧЕСКИЕ РЕЗУЛЬТАТЫ

|  |
| --- |
| **1. Загрузка данных и создание дополнительного набора данных** |
| 1. Загрузите полностью основной набор изображений и масок из папки «train\_new». 2. Создайте дополнительный набор данных. Необходимо получить набор данных EO\_Data из «EO Browser» (описан подробнее в **Приложении 1**), содержащий фрагменты территорий суши с с/х угодиями, каждый размером не менее 224х224 пикселя с разрешением таким же, как у основного набора. Создать набор не менее 500 изображений TrueColor и NDVI в форматах PNG. 3. Добавленные изображения EO\_Data необходимо преобразовать к разрешению 1 пиксел = 10 м. |
| **2. Предобработка данных. Анализ и корректировка изображений** |
| * + - 1. Предложите алгоритм нахождения облачности на изображении. Обоснуйте Ваше предложение.       2. Найдите все изображения из основного набора, содержащие облачность. Оцените долю облачности на каждом найденном изображении.       3. В качестве проверки выведите исходные изображения из основного набора (не менее 10 шт.), содержащие облачность и полученные маски облачности для этих изображений. А также процент пикселей, принадлежащих облачности для каждого изображения.       4. Выведите изображения из дополнительного набора EO\_Data (не менее 5 шт.), содержащие облачность и полученные маски облачности для этих изображений. А также процент пикселей, принадлежащих облачности для каждого изображения.       5. Подготовьте описание данных (основной набор) по следующей форме  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | **№** | **Файл** | **Дата** | **Есть ли с/х поле?** | **% с/х угодий** | **Количество областей (с/х поля)** | **Наличие облачности** | **Доля облачности** | |  |  |  |  |  |  |  |  |  * + - 1. Сохраните полученную таблицу в doc или pdf-файл Description\_№, где № - номер Вашего рабочего места. |
| **3. Отчет** |
| 1. Предоставлен отчет о проделанной работе. Внимание: оценка сессии будет проводиться на основании отчета. Отчет предлагается писать в Jupyter Notebook или аналогичной среде, где участник может последовательно представить, как описание проделанной работы, так и часть программного кода, и результат работы программы. 2. Отчет должен быть предоставлен в папке C1[X]\_MA, где [Х] – номер участника, который должен быть размещен на рабочем столе. Папка должна содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков кода. Отчет также должен быть в формате \*.pdf или \*.html |

Приложение 1 к Модулю А

EO Browser

EO Browser обеспечивает открытый доступ к коллекциям полноразмерных спутниковых снимков, включая снимки в исходном разрешении Landsat 5, 7, 8 и Sentinel 2, 3 в различных комбинациях каналов. Доступен по адресу: <https://apps.sentinel-hub.com/eo-browser>.



В рамках конкурсного задания используются только снимки Sentinel-2. Он должен быть выбран на первой странице EO Browser и, после нажатия кнопки «Search» появится список доступных снимков за текущую дату. После выбора снимка появляется список доступных для просмотра слоев (комбинации различных спектральных каналов). При выполнении задания используются слои TrueColor и NDVI.

По умолчанию EO Browser осуществляет поиск снимков за последнюю доступную дату. Для изменения временного диапазона поиска доступны две опции:

- Single Date позволяет выбрать конкретную дату для поиска снимков.

- Timespan позволяет выбрать диапазон дат для поиска снимков.

В правой части экрана находится список инструментов, один из которых позволяет сохранять изображение на экране браузера в файлы формата JPG, PNG. Этот инструмент можно использовать для получения фрагментов спутниковых снимков.

**C1. Модуль B. Разведочный анализ данных**

## СОДЕРЖАНИЕ

Модуль B данного Конкурсного задания состоит из следующей документации / файлов:

1. Инструкция к Модулю B.

## ИНСТРУКЦИЯ УЧАСТНИКУ

В рамках данного модуля Вам предстоит расширить набор данных, а также сегментировать спутниковые снимки и найти на них поля сельскохозяйственного назначения.

## ПРАКТИЧЕСКИЕ РЕЗУЛЬТАТЫ

|  |
| --- |
| **1. Расширение набора данных с помощью методов аугментации** |
| * + - 1. Провести аугментацию для 500 снимков с разными комбинациями признаков: например, «нет облачности, малая доля (3%) с/х угодий» и т.п. (см. построенную в Модуле А таблицу Description\_№). Сформировать не менее 5 комбинаций.   При аугментации:   * + - 1. Выполнены зеркальные отражения.       2. Выполнены повороты.       3. Выполнены приближения/отдаления (зумы).       4. Выполнена смена яркости.       5. Выполнена смена контрастности.       6. Используются обрезки (кропы).       7. Используются наклоны.       8. Используются вращения. |
| **2. Сегментация изображений «без учителя»** |
| * + - 1. Для сегментации (кластеризации) областей (основной набор данных, дополненный изображениями, полученными в результате аугментации) предложить не менее 3-х вариантов алгоритмов. Для реализации выбрать один. Обосновать выбор.       2. Выполнить кластеризацию областей (основной набор данных, дополненный изображениями, полученными в результате аугментации) – не менее 5 кластеров. Сделать предположения о характере каждого кластера (с/х угодия, лес, облака, вода и т.д.).       3. Визуализируйте результат кластеризации для не менее 5 изображений.       4. Выполнить кластеризацию областей (дополнительный набор EO\_Data TrueColor) – не менее 5 кластеров. Сделать предположения о характере каждого кластера (с/х угодия, лес, облака, вода и т.д.) – для этого понадобится EO\_Data NDVI.       5. Сделайте вывод – различаются ли найденные кластеры в основном и дополнительном наборах?       6. Выполнить расчет индексов IoU и Dice для оценки полученного кластеризатора (основной набор данных, дополненный изображениями, полученными в результате аугментации) с точки зрения определения с/х угодий с использованием масок. |
| **3. Формирование итогового набора расширенных и предобработанных данных** |
| * + - 1. Сформирован набор данных для обучения модели, включающий основной набор данных, дополненный изображениями, полученными в результате аугментации.       2. Итоговый набор данных содержит разметку для всех изображений.       3. Итоговый набор данных изображений и разметок сохранен в архив с расширением .zip |
| **4. Отчет** |
| 1. Предоставлен отчет о проделанной работе. Внимание: оценка сессии будет проводиться на основании отчета. Отчет предлагается писать в Jupyter Notebook или аналогичной среде, где участник может последовательно представить, как описание проделанной работы, так и часть программного кода, и результат работы программы. 2. Отчет должен быть предоставлен в папке C1[X]\_MB, где [Х] – номер участника, который должен быть размещен на рабочем столе. Папка должна содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков кода. Отчет также должен быть в формате \*.pdf или \*.html |

**C2. Модуль C. Построение, обучение и оптимизация модели**

## СОДЕРЖАНИЕ

Модуль C данного Конкурсного задания состоит из следующей документации / файлов:

1. Построенная в Модуле А таблица Description\_№.
2. Набор данных, полученный в Модуле В.
3. Архивы спутниковых снимков для областей Краснодар и Белгород. Снимки за период с 04.2023 по 08.2023 г. Каждый снимок содержит изображение TrueColor в формате PNG и значения спектральных каналов Red (band 4, res 10m), NIR (band 8, res 10m), SWIR1 (band 11, res 20m), каждый из которых в формате TIF.
4. Приложение 2 к Модулю С.
5. Инструкция к Модулю C.

## ИНСТРУКЦИЯ УЧАСТНИКУ

В рамках данного модуля Вам предстоит построить, обучить и оптимизировать модель сегментации с/х угодий, а также оценить ее качество.

## ПРАКТИЧЕСКИЕ РЕЗУЛЬТАТЫ

|  |
| --- |
| **1. Сегментация изображений «с учителем»** |
| * + - 1. Разделить набор данных, **полученный в Модуле В**, на обучающую, валидационную и тестовую выборки. Обосновать разбиение.       2. Выбрать и обосновать выбор модель сегментации. Рассмотреть не менее 3-х моделей.       3. Выполнить построение и обучение выбранной модели сегментации с помощью обучающей и валидационной выборок.       4. Визуализируйте результаты обучения модели/моделей в процессе настройки. Сделайте выводы.       5. Выполнить расчет индексов IoU и Dice для тестового набора данных.       6. Если индексы IoU и Dice больше 0,7, то хорошо.       7. Выполнить визуализацию для не менее 10 разных изображений (с разными комбинациями признаков: например, «нет облачности, малая доля (3%) с/х угодий» и т.п. Сформировать не менее 5 комбинаций. Таблица Description\_№ из Модуля А). |
| **2. Сегментация набора из Модуля С** |
| * + - 1. Выполнить сегментацию для заданного набора данных по Краснодарской и Белгородской областям (данные Модуля С). Отобрать только те снимки, на которых есть поля. |
| **3. Классификация с/х угодий** |
| * + - 1. Разбить поля на следующие группы: 1) не засеяны (под паром)/не используются, 2) используются для выращивания с/х культур. Для разделения полей используются два основных индикатора: цветовой и средний индекс NDVI на поле. Индекс NDVI рассчитывается как комбинация каналов NIR и Red (в Приложении 2 представлена краткая справка по индексу NDVI).   *Среднее значение индекса NDVI менее 0.1 означает, что растительности практически нет на поле. Высокая неоднородность значений индекса в пределах одного поля может означать, что поле не окультурено. Также, открытая почва сильно контрастирует по цвету с зелеными полями*   * + - 1. Визуализируйте результаты классификации.       2. Рассчитать общую площадь полей под с/х культурами – для каждого снимка отдельно и общую площадь по всем снимкам.       3. Предложите алгоритм для классификации полей исходя из предположения, что там растут одинаковые с/х культуры. Критерий однотипности полей – анализ графика изменения среднего (для поля) значения индекса NDVI за период 04.2023 по 08.2023.       4. Реализуйте предложенный алгоритм.       5. В качестве примера работы алгоритма приведите снимки одной выбранной области за период 04.2023 по 08.2023, укажите на ней поля.       6. Постройте графики изменения среднего (для каждого поля выбранной области) значения индекса NDVI за период 04.2023 по 08.2023.       7. Визуализировать результаты разделения полей по типам культур (например, Культура 1 – желтый цвет, Культура 2 – синий цвет и т.п.).   Пример:     * + - 1. Рассчитать площадь полей, засеянных одинаковыми культурами. |
| **4. Классификация с/х угодий** |
| * + - 1. Рассчитать индекс NDMI как комбинацию каналов NIR и SWIR1. (При расчете необходимо учитывать, что пространственное разрешение канала SWIR1 нужно привести к 10м). Индекс подробнее описан в Приложении 3.       2. Визуализируйте (выделите) зоны предполагаемого аномального развития растений (болезни). Приведите в качестве примера не менее 5 снимков.   *Индикатором зоны предполагаемого аномального развития растений (болезни) является диапазон изменения индекса NDVI 0.-0.3 при среднем для поля значении индекса больше 0.5.*   * + - 1. Визуализируйте (выделите) зоны на полях, где растения испытывают дефицит влаги. Рассчитайте площадь этих зон. Выделите зоны повышенной пожароопасности и рассчитайте их площадь.   *Пожароопасные зоны – NDMI в диапазоне 0.–0.2. Зоны дефицита влаги – NDMI в диапазоне 0.–0.4.* |
| **5. Отчет** |
| 1. Предоставлен отчет о проделанной работе. Внимание: оценка сессии будет проводиться на основании отчета. Отчет предлагается писать в Jupyter Notebook или аналогичной среде, где участник может последовательно представить, как описание проделанной работы, так и часть программного кода, и результат работы программы. 2. Отчет должен быть предоставлен в папке C2[X]\_MС, где [Х] – номер участника, который должен быть размещен на рабочем столе. Папка должна содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков кода. Отчет также должен быть в формате \*.pdf или \*.html |

Приложение 2 к Модулю С

Индекс NDVI (Normalized Difference Vegetation Index)

Нормализированный вегетационный индекс NDVI (NDVI) – это стандартизированный индекс, показывающий наличие и состояние растительности (относительную биомассу). Этот индекс использует контраст характеристик двух каналов из набора мультиспектральных растровых данных – поглощения пигментом хлорофилла в красном канале и высокой отражательной способности растительности в инфракрасном канале (NIR). Один из самых распространенных и используемых индексов для решения задач, использующих количественные оценки растительного покрова.

Индекс рассчитывается по следующей формуле:

NDVI = ((NIR - Red)/(NIR + Red))

NIR – значения пикселов из ближнего инфракрасного канала.

Red – значения пикселов из красного канала.

Этот индекс принимает значения в диапазоне -1.0 и 1.0. Для растительности значение индекса меняется от 0 до 1.0. Примеры значений спектральных каналов и индекса NDVI приведены в таблице.

|  |  |  |  |
| --- | --- | --- | --- |
| **Тип объекта** | **Отражение в красной области спектра** | **Отражение в инфракрасной области спектра** | **Значение NDVI** |
| Густая растительность | 0.1 | 0.5 | 0.7 |
| Разряженная растительность | 0.1 | 0.3 | 0.5 |
| Открытая почва | 0.25 | 0.3 | 0.025 |
| Облака | 0.25 | 0.25 | 0 |
| Снег и лед | 0.375 | 0.35 | -0.05 |
| Вода | 0.02 | 0.01 | -0.25 |
| Искусственные материалы (бетон, асфальт) | 0.3 | 0.1 | -0.5 |

**C2. Модуль D. Разработка программного продукта**

## СОДЕРЖАНИЕ

Модуль D данного Конкурсного задания состоит из следующей документации / файлов:

1. Инструкция к Модулю D.

## ИНСТРУКЦИЯ УЧАСТНИКУ

В рамках данного модуля Вам предстоит разработать API, а также программный продукт в виде приложения для заказчика.

## ПРАКТИЧЕСКИЕ РЕЗУЛЬТАТЫ

|  |
| --- |
| **1. Разработка API** |
| * + - 1. Разработано API       2. Разработана и реализована функция загрузки изображения/изображений.       3. Разработана и реализована функция обработки изображения.       4. Разработана и реализована функция для сегментации изображений (поля).       5. Разработана и реализована функция визуализации результатов сегментации.       6. Разработана и реализована функция определения площади полей.       7. Разработана и реализована функция определения полей с одинаковыми культурами.       8. Разработана и реализована функция формирования зон предполагаемого аномального развития растений (болезни) с выводом значения соответствующего индикатора и определением масштаба этих зон. |
| **2. Разработка графического интерфейса пользователя (приложение)** |
| * + - 1. Интерфейс разработан в виде приложения.       2. Интерфейс имеет красивый внешний вид, легко читаются слова, легко понять где, что и куда нажимать, подобрана цветовая палитра.       3. Интерфейс использует методы и функции из API предыдущего модуля.       4. Интерфейс позволяет загружать данные с одного изображения.       5. Интерфейс позволяет загружать данные с нескольких изображений.       6. Интерфейс позволяет проводить сегментацию изображений (определение полей). Если полей нет, то выдается соответствующее предупреждение.       7. Интерфейс позволяет выделять контуры полей.       8. Интерфейс позволяет определять площади полей.       9. Интерфейс позволяет определять поля с одинаковыми культурами.       10. Интерфейс позволяет определить зоны предполагаемого аномального развития растений (болезни) с выводом значения соответствующего индикатора и определением масштаба этих зон.       11. Интерфейс позволяет определить зоны на полях, где растения испытывают дефицит влаги с выводом значения соответствующего индикатора и определением масштаба этих зон. |
| **3. Отчет** |
| 1. Предоставлен отчет о проделанной работе. Внимание: оценка сессии будет проводиться на основании отчета. Отчет предлагается писать в Jupyter Notebook или аналогичной среде, где участник может последовательно представить, как описание проделанной работы, так и часть программного кода, и результат работы программы. 2. Отчет должен быть предоставлен в папке C2[X]\_MD, где [Х] – номер участника, который должен быть размещен на рабочем столе. Папка должна содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков кода. Отчет также должен быть в формате \*.pdf или \*.html |

**C3. Модуль E. Разработка средств интеграции и поддержки готового решения. Презентация результатов работы**

## СОДЕРЖАНИЕ

Модуль E данного Конкурсного задания состоит из следующей документации / файлов:

1. «Секретная» выборка – папка valid\_new
2. Инструкция к Модулю E.

## ИНСТРУКЦИЯ УЧАСТНИКУ

В рамках данного модуля Вам предстоит оценить точность построенной модели на секретной выборке, оценить результат. Также необходимо будет подготовить продукт для передачи заказчику – оформить инструкцию для пользователя, презентацию и сделать доклад.

## ПРАКТИЧЕСКИЕ РЕЗУЛЬТАТЫ

|  |
| --- |
| **1. Оценка точности разработанной модели на «секретной» выборке** |
| * + - 1. Определить точность разработанной модели сегментации, используя проверочную выборку из папки «valid\_new».       2. Выполнить расчет индексов IoU и Dice для набора данных. Сделать вывод о точности. |
| **2. Разработка инструкции для пользователя (по запуску и эксплуатации приложения)** |
| * + - 1. Разработана инструкция для пользователя по работе с разработанной системой (файл с названием Инструкция в формате docx или txt).       2. Инструкция содержит описание полного функционала работы системы.       3. Инструкция оформлена профессионально (в соответствии с профессиональными стандартами). |
| **3. Подготовка презентации** |
| * + - 1. Презентация, охватывает все результаты выполнения конкурсного задания.       2. В презентации указано ёмкое описание результатов работы с обоснованием выбора того или иного решения.       3. В презентации имеются скриншоты работы программы и выводимых результатов.       4. Презентация включает клиентоориентированную часть.       5. В презентации присутствуют перспективы улучшения модели и дальнейшего использования результатов в промышленности. |
| **4. Доклад** |
| * + - 1. Регламент – не более 5 минут.       2. Доклад построен логически верно, грамотно, нет слов-паразитов.       3. Доклад сделан четко, без повторений сказанного и запинок.       4. Доклад сделан профессионально – лексикон предметной области, лектор – профессионал своего дела. |
| **5. Итоговый отчет** |
| 1. Предоставлен отчет по всей проделанной работе. Отчет предлагается писать в Jupyter Notebook или аналогичной среде, где участник может последовательно представить, как описание проделанной работы, так и часть программного кода, и результат работы программы. 2. Отчет должен быть предоставлен в папке C3[X]\_ME, где [Х] – номер участника, который должен быть размещен на рабочем столе. Папка должна содержать все результаты выполнения модуля E, а также все необходимые файлы для запуска и проверки участков кода. Отчет также должен быть в формате \*.pdf или \*.html |