***ФГБОУ ВО «Кубанский государственный аграрный университет имени И.Т. Трубилина»,***

***Российская Федерация***

**161 Инженирия знаний и интеллектуалные системы. Лабораторная 8. 2020-10-26**

**Инструкции по Лабораторной работе №8 в системе AIDOS: Разработка приложения, регистрация и технические аспекты**

**Резюме**

Данный текст представляет собой запись онлайн-занятия (лабораторной работы №8) по дисциплине "Инженерия знаний и интеллектуальные системы" для группы Б17-01 от 26 октября 2020 года, проводимого профессорами Луценко Е.В. и Аршиновым Г.А.

1. **Цель Лабораторной работы №8**: Разработка собственного интеллектуального приложения с использованием системы AIDOS-X. Это ключевая работа, демонстрирующая освоение компетенций курса.
2. **Предварительные требования**:
   * Студенты должны зарегистрироваться в системах РИНЦ (eLibrary) и ResearchGate. Регистрация в РИНЦ может занять время и требует получения SPIN-кода для заключения договора на размещение публикаций.
   * Необходимо выбрать уникальную тему для приложения, проверив ее по списку уже существующих работ, доступному по ссылке в инструкции.
3. **Процесс разработки и сдачи**:
   * Выбрать тему и исходные данные (желательно объемом до 10 МБ для размещения в облаке AIDOS).
   * Попробовать самостоятельно создать модель в системе AIDOS.
   * Проверить достоверность модели (режим 3-4). Если достоверность высокая, прислать скриншот распределения частот преподавателю на утверждение темы.
   * Если возникают трудности с созданием модели или конвертацией данных, прислать ссылку на данные и описание темы преподавателю для помощи и утверждения.
   * После утверждения темы подготовить описание приложения по одному из предложенных шаблонов.
   * Разместить готовое приложение и его описание в облаке AIDOS, а также в ResearchGate и РИНЦ (после получения SPIN-кода).
4. **Технические аспекты и обновления системы AIDOS**:
   * **Обновление системы**: Необходимо обновить систему AIDOS до последней версии, так как в ней исправлены ошибки и добавлены новые функции.
   * **CSV Конвертер**: Добавлен улучшенный CSV-конвертер (режим 2-3-2-2 и отдельный 2-3-2-10), который корректно обрабатывает различные форматы CSV, в отличие от Excel (который может неверно интерпретировать числа как даты). Он позволяет вводить данные из CSV, DBF, старых и новых XLS файлов. Режим 2-3-2-10 позволяет конвертировать CSV в DBF для анализа структуры данных (заголовков) в Excel или текстовом редакторе.
   * **Большие данные**: При работе с очень большими наборами данных (гигабайты) следует использовать подмножество данных (до 10 МБ) для разработки и размещения в облаке AIDOS. Полные данные в облако не поместятся.
   * **Совместимость с Windows**: Для избежания проблем совместимости с Windows 10 рекомендуется использовать виртуальную машину (VirtualBox, VMware Player) с установленной Windows 7, где система AIDOS работает без дополнительных настроек. Работа под Windows 10 возможна, но может требовать отключения защитных механизмов или цифровой подписи исполняемых модулей (что платно).
5. **Обсуждение и взаимодействие**: Занятия посвящены помощи студентам в разработке приложений. Можно задавать вопросы, присылать данные и промежуточные результаты в чате или по почте. Возможна совместная работа над одним приложением (вдвоем).

**Детальная расшифровка текста**

**I. Введение и организационные моменты**

* **Приветствие и проверка связи**:
  + Здравствуйте, ребята.
  + Добрый вечер.
  + Добрый вечер.
  + Здравствуйте.
  + Здравствуйте, здравствуйте.
* **Дата и время занятия**:
  + Сегодня у нас 26 октября 2020 года.
  + Шестая пара.
  + 17:20 - 18:50.
* **Тема и группа**:
  + Лабораторная работа номер 8 с группой Б17-01.
  + По дисциплине Инженерия знаний и интеллектуальные системы.
* **Преподаватели**:
  + Занятия ведут:
  + Профессор Луценко Евгений Вениаминович
  + и профессор Аршинов Георгий Александрович.

**II. Обзор пройденного материала и постановка задачи ЛР №8**

* **Обзор предыдущих ЛР**:
  + Так, ребята. Значит, вы сейчас должны всё видеть на экране.
  + Вот. Сейчас мы занимаемся только одним делом все.
  + То есть уже мы основные виды лабораторных работ изучили, которые показывают, как работать с табличными данными, числовыми и текстовыми, как работать с текстовыми данными (интеллектуальная обработка текстов), как решать реальные задачи с обработкой табличной информации (методики риэлторской оценки мы посмотрели).
  + И даже мы посмотрели, как работать с изображениями и осуществлять, так сказать, все операции интеллектуальной обработки изображений: обобщение, идентификацию, кластерно-конструктивный анализ изображений и самих значений факторов, в качестве которых выступают цвета, цветовые диапазоны. Всё это тоже мы посмотрели.
* **Цель ЛР №8: Разработка собственного приложения**:
  + Поэтому вы готовы уже по своей квалификации к тому, чтобы разработать собственное приложение и разместить его в AIDOS-облаке, а также в ResearchGate и РИНЦ.
  + Поэтому сейчас мы этим и занимаемся.

**III. Требования к выполнению ЛР №8 и обратная связь**

* **Необходимость обратной связи**:
  + Пожалуйста, мне дайте информацию обратной связи в чате.
* **Посещаемость**:
  + Маловато всё-таки людей для лабораторной работы ключевого, ключевой, понимаете?
  + Да, добрый вечер, добрый вечер, ребята. Ключевой работы, которая вот мы рассматриваем, как создавать приложение собственное.
  + Вот. В чате...
  + Здравствуйте.
  + Что-что?
  + Ещё раз скажи.
  + Мало, подключаются сейчас.
  + Мм. Да, хорошо.
* **Запрос статуса выполнения**:
  + В общем, пожалуйста, в чате мне напишите, кто, ну, я не знаю, что там от вас сейчас, сразу чтоб всё вы написали. Одним плюсиком всё сразу не напишешь.
  + Ну, допустим, напишите: в ResearchGate зарегистрировался, в РИНЦе зарегистрировался, да или нет?
  + В РИНЦ – да, там, в ResearchGate – да, нет.
  + Тему выбрал, да, нет?
  + Вот такое вот напишите мне.
  + Вот.
  + Голосом можно?
  + Ну, давайте голосом. Просто когда сразу так вот смотришь, видно, видна картина.

**IV. Обсуждение прогресса студентов**

* **Данил Дворников**:
  + Ну хорошо, давайте тогда по очереди. Тогда вот, допустим, Дворников Данил.
  + Что у вас?
  + Я выбрал тему по исследованиям на COVID-19. Подскажите, пожалуйста, вот там, получается, данные есть в CSV формате, в TXT формате и тому подобное. То есть мне всё это вам нужно будет сейчас скинуть?
* **Екатерина Ильинова**:
  + Так, ну, с этим понятно. Значит, ну, Екатерина Ильинова, которая.
  + Что у вас?
  + В принципе, ResearchGate я ещё не регистрировался, я всё время тему выбирал.
  + Мм. В общем, дело в том, что РИНЦ это занимает довольно много времени. Надо РИНЦ и Science Index сразу регистрироваться.
  + Ну-ну, давайте.
  + Можно ещё вопрос? Смотрите, там получается 22 ГБ данных. Они же, получается, на почту не влезут. Я могу...
  + Сколько, сколько данных гигабайт?
  + 22 ГБ.
  + Ну это перебор явный. Тогда надо либо на подмножестве этих данных делать, либо вообще другую тему. Но если вам эта тема интересна, то можете взять данных, ну, я не знаю, там. Ну, в общем, чтоб файл исходных данных был где-то в пределах 10 МБ. Иначе они иначе они в облаке не разместятся эти данные.
  + А могу я вам тогда ссылку отправить на эти данные, чтобы вы посмотрели?
  + Да. Я ж про это и говорю.
  + Обычно мы только тренировочную выборку используем. Мы ж не собираемся там в конкурсах участвовать на этом Кагле. Хотя я не исключаю, что вас это как-то увлечёт, и в будущем, возможно, вы будете чемпионами, там, гроссмейстерами на этом Кагле и будете вспоминать времена, как 20 лет назад вам Евгений Вениаминович там, Георгий Александрович показывали этот Кагл.
  + Так, ну, Екатерина, какие у тебя успехи? Как насчёт темы, как насчёт регистрации в РИНЦ, в ResearchGate?
  + Тему я ещё выбираю. А в сайтах я зарегистрировалась. Не приняли и в РИНЦе.
  + Так, тему ты выбираешь. В РИНЦе зарегистрировалась, да? И получилось? Или начала регистрацию?
  + Я начала регистрацию, пока всё как бы получается.
  + То есть просто ты ждёшь, когда там они среагируют, да?
  + Ну да.
  + А в ResearchGate зарегистрировалась, да?
  + Да.
  + Молодец. Значит, теперь давай выбирай тему. Если есть какие-то вопросы по теме, из каких выбирать, там, где выбирать, то ты меня сейчас вот спроси, прямо вот сейчас. Я могу или вот на занятии просто, я подскажу. Мы это подробно рассматривали вообще-то. Это пункт, пункт у нас есть четвёртый инструкции. Вы ж видите, да, инструкцию? Вот. Видите? Вот пункт четвёртый инструкции посвящён выбору темы. И пункт пятый тоже – утверждению темы. То есть вы ищете тему, вот, пробуете, как получается. Если получается нормально, утверждаем.
  + Хорошо.
  + Вот.
  + Ну, как тебе удалось зарегистрироваться в ResearchGate? Ты использовала какой-то адрес корпоративный, да? Или как? Или вам помог кто-то?
  + Ну... Там, в принципе, всё, как вы рассказывали. Я через сайт заходила, регистрировалась. Там правда всё на английском языке, так понять можно.
  + Я имею в виду, там же почту нужно корпоративную использовать. Они посылают на неё ссылочку, вы даёте подтверждение, и вас тогда регистрируют. А у вас почта есть корпоративная, да? Ну там с доменом KubSAU.ru? Или какой?
  + Да, да.
  + У вас такая почта была, да?
  + Да.
  + А кто вам её дал? Как это получилось, что она у вас появилась такая почта? Потому что обычно ни у кого нету. Почему меня это интересует, где вы её взяли? Ну вот почту взяли, а связь не очень, да? Где почту взяли, Кать? Рассказывайте, признавайтесь. Чтоб другие тоже могли это сделать, может быть. Ну? Опять не поняли, что я спрашиваю, да?
* **Ольга Лемякина**:
  + Ну, Ольга Лемякина.
  + У вас как? Насчёт регистрации и насчёт выбора темы?
  + У меня просто интернет плохой. Я зарегистрировалась, но ещё не зарегистрировалась.
  + Ну, значит, в РИНЦ там можно начинать в любой момент регистрацию, там ничего не требуется. А в ResearchGate я вот не знаю, как вот Екатерине удалось зарегистрироваться. Она где-то взяла корпоративный адрес, да, Екатерина?
  + А я...
  + Да.
  + А если я не успею зарегистрироваться вместо?
  + Ну, если не удастся вам зарегистрироваться, это я уже говорил, ребята. Если не удастся, то как крайний случай, можно меня включить в соавторы, и я тогда это сделаю. Но это нежелательный вариант, потому что у меня идёт перебор самоцитирований. Вот. Мне нежелательно сейчас, ну, как ещё добавлять статьи со ссылками на свои работы, короче. Мне было бы более ценно, если бы вы сами там смогли это сделать. Но если будет, так сказать, ну, нет худа без добра. Ну больше публикаций будет тогда. Но не будут в том же ключе, что и раньше. То есть я, ну, как скажем, таких публикаций у меня перебор, понимаете? Ну можно будет, ну можно. Я про это говорил, что если вы сами будете затрудняться в этом вопросе, не будет у вас получаться, ну тогда я это сделаю. Вот так скажем.
  + Можно ещё уточнить по моей теме? У меня анализ катастроф и вы говорили...
  + Да, Алёна, я вас тоже слышу через слово, вы знаете. Ну напишите в чат, на почту. К сожалению, действительно ничего не слышно, плохо слышно очень.
  + Ну, давайте дальше. Как дела у Виктории?
* **Виктория**:
  + У меня пока всё в процессе. И регистрация, и выбор темы.
  + Мм. Ну давайте, не забывайте, что это нужно сделать, потому что это в этом и заключается подтверждение вами освоения учебного плана и компетенций, понимаете? Это как раз и подтверждает, что вы успешно прошли обучение, приобрели соответствующие компетенции. И остальные, ребят, вы тоже на это обратите внимание. Потому что в других, когда вот идут занятия очные в аудиториях, там преподаватели видят, как у вас получается. Видят, что вы сидите за компьютерами и делаете всё, понимаете? Я могу прямо сразу поставить вам аттестацию. Я уже знаю, у кого как получается. А сейчас я даже не знаю, как вы выглядите, только голос там иногда там изредка там что-то там кто-то покажет своё личико. А так, в общем, получается очень слабая обратная связь. То есть я не знаю, как у вас получается, понимаете? И вот это вот материалы эти, которые вы должны разработать, то есть вы должны решить задачу, описать её, разместить. Это и является подтверждением того, что вы успешно прошли обучение.

**V. Технические инструкции и обновления AIDOS**

* **Необходимость обновления системы**:
  + Значит, смотрите, ребята, значит, сначала дам вам новости некоторые.
  + Значит, вот нам надо всем обновить сейчас систему по одной очень простой причине. Я, значит, разместил только что новую версию, причём только что в буквальном смысле, ну где-то в обед.
  + Вот. Там была до этого версия тоже не старая, по 26-го числа. Я ещё одну версию 26-го числа разместил позже, чем ту. Ту я вчера разместил, ну или, можно сказать, ночью. А эту днём сегодня.
* **Новый CSV Конвертер**:
  + Значит, в ней хороший CSV-конвертер в режиме 2-3-2-2. То есть там можно экселевские файлы вводить старого-нового Экселя стандарта, DBF-файлы можно вводить, это и раньше было. И теперь можно вводить CSV-файлы в качестве исходных данных.
  + Но дело в том, что для того, чтобы их вводить, для этого нужно и прямо создавать приложение, для этого нужно сначала понять, где там у нас факторы, а где результаты их влияния. То есть что мы хотим узнать и на основе чего. А это для этого нужно диапазоны указать классификационных и описательных шкал. Вот для этого это нужно, собственно, это узнать.
  + Вот. Поэтому, чтобы это узнать, я сделал ещё отдельный CSV/DBF конвертер в режиме 2-3-2-10. Не только включил в состав 2-3-2-2 режима его конвертера, то есть этого, то есть общего универсального интерфейса ввода внешних данных в систему AIDOS, но также включил его прямо в виде отдельного режима.
  + Для чего? Для того, чтобы вы могли CSV-файл ввести в формат DBF. DBF-формат открывается в Экселе. И вы можете, ну, хотя бы частично, то есть он может быть размером до десятков миллионов записей. Я этот режим тестировал на файле, в котором 4 млн записей. В Экселе такое никак невозможно открыть, там до миллиона записей. Вот. Ну и то это много даже, и миллион. Но даже если открыть его в старом Экселе (новый, по-моему, не открывает DBF-файлы), но там больше 65.536 строк. Но всё-таки там будет заголовок. Можно заголовок почитать, посмотреть. Точно так же заголовок можно вытащить с самого CSV-файла, это первая строка. Открыть его в каком-нибудь редакторе текстовом и просто взять первую строку и проанализировать эти заголовки. После этого можно его вводить.
  + Вот. Этот конвертер нормально работает, в отличие от Экселя. Если Excel попробовать использовать для ввода CSV-файлов, и в режиме "Данные", можно вводить текстовые из текстов данные. И там получается, что хорошая, лучший враг хорошего. То есть Excel иногда некоторые числа, ну такие типа числа такие, вот 26 10 20 20, например. Вот такое число, если просто ввести подряд все эти цифры 2 6 1 0 там 2 0 2 0 и попробовать в Экселе его ввести файл, в котором такие есть числа, то эти числа будут заменены на даты, понимаете? Что совершенно неуместно.
  + Вот. Поэтому этот вот конвертер, который я сейчас там в системе сделал, он очень нужен. Конечно, можно ещё использовать онлайн-конвертеры, но надо смотреть, что там получилось в результате.
  + Так что обновляйте версию.
* **Совместимость с Windows и Виртуализация**:
  + Значит, теперь, вот в этом описании, где мы, где я описываю, как настраивать систему Windows, я добавил одиннадцатый пункт. Вот. Предлагаю вам, мне подсказали, ребята. Ну я так и раньше тоже это понимал, что можно использовать виртуальную машину. Но сейчас вроде как это несколько проще стало, чем раньше. Раньше VMware Player для этой цели использовался в те времена когда-то. А сейчас вот есть VirtualBox в Windows 10. Вот если поставить туда семёрку, то вообще все вопросы снимаются, никаких настроек не требуется, всё работает.
  + Вот. Но я сейчас вплотную стал заниматься тем, чтобы она была нормально работала под Windows 10 система. Ну, говорят, что надо для этого подписывать свои разработки. А ключи для подписания разработок стоят 80 там, около 90 долларов в год. Ну как бы не бог весть сколько, но всё-таки там тысяч восемь надо платить, понимаете? Вот. Так что, в общем, возникает вопрос, надо ли это делать. Потому что, ну, в общем, если её подписать, экзе-модули системы подписать, то тогда, возможно, Windows 10 будет нормально к ним относиться.
  + Поэтому, ну, проще, конечно, поставить семёрку по VirtualBox. Вот об этом я вам хотел сказать.

**VI. Процесс выбора и утверждения темы ЛР №8**

* **Выбор темы**:
  + Значит, теперь, когда вы выбираете тему, что нужно делать после этого? После выбора темы вы должны попробовать создать модель самостоятельно. И так, как вот я показывал вам на лабораторных работах.
  + И посмотреть, насколько эта модель будет достоверна. Если она будет иметь высокую достоверность, присылайте мне картиночку распределения частотного из режима 3-4. И, значит, я смотрю, насколько она достоверная получилась модель, разумная. Если она получилась хорошая модель, то есть это пункт пятый я рассказываю инструкции. Вы сейчас всё это видите, должны видеть. То тогда, значит, почему бы и нет. Тогда вполне можно эту тему утвердить и дальше делать описание. Описание тоже мне показать потом в результате, что вы сделали.
* **Утверждение темы**:
  + Вот. И когда вы мне можете, если у вас не получается сделать приложение или конвертировать данные в систему AIDOS, то тогда можете прислать мне просто ссылочку. Написать мне, что, Евгений Вениаминович, значит, я сейчас я вам пришлю свою почту. Вот я бы хотел вот эту тему выбрать, но вот у меня не получается ввести данные и сделать модель. Вот. Не могли бы посмотреть. Ну я посмотрю. У меня это не столько времени занимает, сколько у вас. Быстро я могу это посмотреть. И скажу, что принимается, и вы начинаете эту тему описывать уже. Ну то есть делаете модель и сразу описываете по шаблону по какому-то, по одному из шаблонов.
* **Использование списка существующих тем**:
  + Значит, теперь вот я сам задам вопрос. Вот Орешкина Кристина спрашивает: "А можно ссылку?" Вот я с удовольствием, только я не знаю, какую тебе ссылку нужно, ссылку на что?
  + Это я в начале спрашивала, когда вы...
  + Теперь, теперь, а теперь ты уже знаешь.
  + Просто я...
  + Да, ну то есть теперь ты знаешь, какую ссылку, да?
  + Ну вот с регистрацией у меня не получилось. То есть, если был бы пример чётко по шагам, что, где, куда, потому что мне не получалось. А с темой я примерно определилась, но точную формулировку ещё нет. То есть это будет либо с картинками, либо с Экселем что-то, а тема, ну что мы, может быть, что-то из медицины, а может быть, что-то из продукции, ну, различий.
  + Ну давайте, думайте. В любом случае выбирайте такую тему, которая вам интересна самим.
  + Хорошо. А вы подскажите, пожалуйста, а ссылку на то, что размещено, вы можете отправить, чтобы там как-то подробные инструкции или ещё что-то?
  + Ссылку на регистрации где?
  + Ну вот то, что вы говорили про регистрацию.
  + Я имею в виду ResearchGate регистрацию и РИНЦ. Вот. Значит, здесь у нас в инструкции есть вот пункт восемь, где как раз описано, как там, куда мы должны перейти, чтобы зарегистрироваться.
  + А вот у вас есть пункты, вот эта вот ссылка общая на все. Вы мне можете отправить? Ну, нам отправить?
  + На саму инструкцию, в смысле, да?
  + Да.
  + Ну конечно, конечно, могу.
  + Спасибо вам большое.
  + Вам. Вот ссылка сейчас длинненькая. Вот видите? Вот это application, вот это последняя application. Вот это и есть ссылка на инструкцию. Она на моём сайте находится.
  + Вот. А здесь вот, видите, есть ссылочки, до этого я написал, на ResearchGate и eLibrary. eLibrary - это и есть РИНЦ. Это их сайт так называется. И там просто нажимаете регистрация, и я почему не писал инструкцию по регистрации? Я ссылки дал на эти системы, где там регистрироваться. А инструкции не давал на, вернее, инструкцию по регистрации не писал по одной очень простой причине. Там совершенно стандартная регистрация, абсолютно такая же, как вот в любом сервисе, на любом форуме, в любую почту, в почтовом сервисе, в форуме там. Абсолютно такая же регистрация там. Единственное, в РИНЦ в этом, там есть особенность, нужно птичку поставить, чтобы в системе Science Index зарегистрироваться. А почему? Потому что, когда вы в системе Science Index зарегистрируетесь, через какое-то время, там, через неделю, через две, вам будет присвоен SPIN-код вот такой. И после этого вы получите право на заключение договора с РИНЦ на размещение публикаций. Вот. И это всё описано вот в этом вордовском, кстати, вот есть ссылочка на вордовский файл, подробнее смотри здесь, видите, вот написано. Там как раз это написано всё.
  + А это что, это прошлые работы?
  + Это работы, которые были сделаны раньше и мной, и очень часто студентами, магистрантами. И у них, значит, вот видите, когда сделано, в 2017 году. Есть здесь работы, которые раньше были сделаны. Вот, но раньше не было режима для их размещения в AIDOS-облаке. Облака не этого самого режима размещения в облаке не было, понимаете? А теперь есть. И он этот режим есть с семнадцатого года. Так что вы можете брать, вот, смотреть описание этих работ. Ну вот, допустим, здесь не очень удобно, таблица широкая. Вот, но вы можете посмотреть вот слева, например, вот вас интересует, скажем, анализ матчей в игре Rocket League, да? 196-е приложение, 196-е. Двигайтесь вправо, вправо до конца. Ищете 196-е описание 196-го приложения и смотрите, как люди делали до вас. Это считай предыдущее поколение студентов. То есть это было примерно год назад.
  + Ну да, я помню, вы это рассказывали. А она, получается, тоже там будет?
  + Да, да. Поэтому надо не ударить в грязь лицом, а именно вот нормально всё делать. И вы можете оттуда взять описание, кстати. Как оно выглядит. Но я вам советую брать только последнее. Да, кстати, здесь у меня 207 показано, их 208 на самом деле, но у меня этот Google Chrome, в кэш у него там застряло 207. В других браузерах сразу 208 показывает. Вот. То есть надо вот здесь посмотреть, чтобы здесь его не было. Если же здесь есть, то, например, анализ питомцев World of Warcraft, да? А вы берёте не питомцев, например, а берёте анализ цен там или противоборств там или ещё что-нибудь там. Понимаете? То есть тоже можно по этой же игре, но другое что-то. Ну, чтобы не совпадало просто-напросто. Чтобы не было полного совпадения.

**VII. Работа с системой AIDOS: Демонстрация и пояснения**

* **Просмотр списка приложений в AIDOS**:
  + Теперь сейчас я запущу систему, покажу вам, где там в системе это. Значит, мы зайдём в режим 1-3. И в режиме 1-3 есть, видите, "Добавить лабораторную работу" есть, да, кнопочка? А есть "Скачать приложение из облака". Вот я беру, нажимаю "Скачать приложение из облака". Вот эти 208 приложений. Видите? Всё здесь работает то же самое. То есть можно два раза кликнуть по строчке, вызовется вот это вот приложение или просто по ссылочке.
* **Каталог обсуждений**:
  + Есть каталог обсуждений вот здесь вот. Этот каталог обсуждений содержит информацию о том, когда какие приложения скачивались. Начиная с 2017 года, видите, вот начало здесь, и до недавнего времени. Вот последний раз 10 октября скачивалось приложение 208 "Анализ влияния финансовых и технологических факторов на результаты работы фермерских хозяйств". Это статья, причём в очень престижном журнале, входящем в ядро РИНЦ, которую я не писал. Но я соавтору, своим соавторам объяснил, как как провести такой анализ. Ну, учил их, грубо говоря, ну примерно, как вот я сейчас вас учу. И вы знаете, они неплохо освоили и прекрасно провели этот анализ. Вот. Ну я им показывал на примерах, как это делается, ссылки давал и всё такое. И потом я почитал статью, сделал несколько замечаний, их исправили и опубликовали.
* **Обсуждение конкретного приложения**:
  + Смотрите, вот, значит, 14 сентября скачано было приложение "Сценарный АС-анализ десятичных знаков" там. А вот 14-го тоже сообщение появилось в определённом, значит, приложении, в 207-м приложении. А нет, новое сообщение в 204-м, в 204-м приложении. Это я показывал, как можно обсуждать эти приложения. Вот вы здесь поместили приложение, да? Это курсовая работа, кстати. И вот нажимаю "Обсуждение AIDOS-приложения". Скачивается файлик текстовый, и там есть обсуждение. Ну я просто показывал студентам, что можно что-то там написать. Вот. И можно получить ссылки на файлы этого приложения. То есть абсолютные гиперссылки. Вот эту гиперссылку можно скопировать в строку адреса и вот уже сегодня я получил, да? Что-то не получилось. А ну-ка. А, вот они те гиперссылки. Вот берём, значит, просто любой из этих вот адресов. Вот, допустим, есть файл. Я, по возможности, файлы вордовские тоже туда записываю. В системе AIDOS есть режим, позволяющий туда записывать приложения. Эта вся информация из папочки indata берётся. Вот видите, выделил абсолютную гиперссылку на вордовский файл в этом приложении 204-м. Вот я её беру, в строку адреса нажимаю. И этот вордовский файл скачивается. Что-то как-то не особо так он быстро скачивается. Вот. Вот скачался, 2 секунды, всё. 5 МБ. То есть так можно скачать любое описание.
* **Конвертер CSV/DBF**:
  + Значит, теперь смотрите, ребят, когда мы в режим 2-3-2-2 входим, вот сюда, вот здесь вот появился такой вот конвертер CSV/DBF-конвертер. Что он позволяет делать? Если исходные данные в CSV-файле, это файл, который чаще всего как стандарт он используется в Кагле, ну и вообще часто он используется. То вот здесь вы можете прочитать описание структуры этого файла, что туда входит, как он должен называться, где он должен находиться, там и всякие у него там ограничения есть. И всё, и, пожалуйста, можно вводить из CSV-формата. И ещё я сделал режим, смотрите, 2-3-2-10, вот он он. Это прямо конвертер CSV-файлов. Но она пишет, что такого файла исходного, исходных данных нету там в папочке, и пишет путь, по которому он должен быть. Вот. Если бы он там был, она бы сейчас его в DBF-файл преобразовала бы сразу. DBF-файл считывается в Экселе. Правда, только в старой версии. Ну можно его конвертировать онлайн. Онлайн-конвертер CSV в XLS, а можете DBF попробовать. DBF. Вот. Для чего это может потребоваться? Если у вас нет Экселя 2003 на компьютере, тогда вот этот новый Excel, начиная с 2007, в нём конвертер убрали DBF-файлов. Он считается устаревшим форматом. Вот. И тогда можно онлайново конвертировать в Excel. Вот. Спокойно. Без проблем это делается. Поэтому я даже не стал это реализовывать в системе, потому что нет никакого смысла реализовывать то, что легко делается онлайн.
* **Пример обработки данных по вину**:
  + Ну, сейчас мы что делаем? Переименовываем файл indata, во-первых. Размещаем его по пути, там, где он должен находиться в папочке исходных данных. Вот. Что-то совсем не то. Много экранов. Студенты что-то пока не спешат приходить. Так. Вот путь на место, где должен быть файл исходных данных, и как он должен называться. То есть я-то скачал не такой. Видите, да? А хотелось бы, чтобы вот он, наверное, был проанализирован. И сейчас мы тоже используем конвертер сначала. Просто в DBF его преобразуем или в Excel потом. Вот. 2 3 2 10. Вот, всё. Вы обратили внимание, что довольно быстро, да, произошло?
  + В отличие от предыдущей.
  + Ну, догадаться можно с трёх раз, почему так получилось. Так, значит, ну потому что он небольшого размера вообще. Значит, я вам даже скажу, какого размера он – 1600 записей. Ну это вполне такой разумный размер. Я хочу подчеркнуть, что здесь уже всё числа, числовые данные числовые. Вот. Дело в том, что Excel вводит данные, получаются точки, данные получаются нечисловые. Значит, смотрите, ребята, вот это quality – это качество, да? По-видимому. Правда, я не понял, там качество чего? Мы там разобрались, качество о чём идёт речь вообще?
  + Красного вина.
  + А! Красного вина. Вот. Значит, тут есть такой момент. Ну да, точно, Win Red, да. Значит, дело в том, что я писал статью по вину. Вот сейчас давай, её зададим. Вина, видите, классов на вина по свойствам. Потом, вина на основе идентификация классов вина. Анализ качества вина сомелье и по физико-химическим свойствам. Это моя статья вот про это. Ну и дальше там анализ вина по параметрам технологии. То есть с вином много связано задач. Поэтому имейте в виду, надо посмотреть, нет ли этой задачи ещё. Вот. Теперь, что мы делаем? Смотрите, ребята. Значит, вот у нас экселевский файл с исходными данными. Значит, здесь вот нужно сделать колоночку обязательно дополнительную. Почему? Потому что у нас колоночка А была тоже там свойства были. Я пишу здесь "Объект". Вот. То есть это имеется в виду, что это пример, наблюдение. 1 2. Вот. Пример наблюдения. Я беру и протягиваю это до конца вниз, чтобы мы могли провести анализ в системе AIDOS, нам нужен вот такой вид, чтобы это было у них. Вот. Сохраняем я в экселевском виде. Здесь даже старый Excel подходит. Желательно, чтоб было до 65.000 строк, тогда Excel позволяет это старый описать полностью. Значит, смотрите, держу клавишу Shift и двигаю вправо. У меня получается что? До двенадцатой колонки у меня описательные шкалы, а тринадцатая - это классификационная шкала. 4 по 12, а это с 13 по 13. Всё, ребята, у меня есть экселевский файл с исходными данными. Мы сейчас что делаем? Берём, стираем приложение и вносим информацию из старого Экселя. Классификационные шкалы с 13 по 13, описательные со 2 по 12. Разные интервалы я вам советую всегда брать. Это модели получаются лучше. И вводим данные из экселевского файла. Вот. Ну, здесь вот можно, допустим, задать семь, например, там, а здесь так вот пусть и будет. Вот. И мы видим, ребята, сколько у нас наблюдений на каждую градацию шкалы. Вот. Там многовато. Ну ладно, берём тогда вот так 5 5. И сколько наблюдений и размеры интервалов. Вот. И, ну сейчас не будем больше экспериментировать, просто пройдём дальше. Значит, вот оно сейчас создаёт шкалы и градации. Получилась одна классификационная шкала "Качество" от трёх до восьми, пять градаций. Вот. И смотрим описательные шкалы. Вот эти вот, которые там были. Все числовые, видите? Всё нормально получилось. Получилась малюсенькая матрица модели. Мы сейчас эту модель просчитываем на 1600 примерах. И, может быть, сейчас мы увидим, какой результат. Это хорошая модель. Она лучше, чем Excel преобразует. Вот сейчас сначала идёт просмотр файла этого для того, чтобы определить, какие там поля, имена полей определяются. Ну имена там мгновенно определяются, это первая строчка. А потом полный просмотр файла осуществляется и определяется для каждого поля длина по максимальному значению, которое там есть в файле. Определяется также тип данных поля, текстовый или числовой. И если это числовой тип данных, то есть там во всём файле нигде в этом поле текстов нет, только числа. Ну хотя числа тоже являются текстами, но не все тексты являются числами. То есть там именно проверяется, чтобы все были только числа там в тексте, других чисел там, чтоб, то есть других текстов, чтоб там не было, кроме чисел. Если числовой тип данных, то проверяется ещё дополнительно, сколько знаков после запятой. То есть формируется, вся информация вытаскивается по структуре DBS файла или таблицы экселевской, которая будет достаточно, чтобы этот CSV-файл ввести в неё. И вот сейчас уже это сделано, файл создан DBS, и в него вводится информация из CSV-файла по полям. Причём все поля достаточного размера и нужного типа данных. Значит, получается DBS файл немножко меньше даже, чем Excel, э CSV получается по размеру. Потому что там есть числовые форматы данных, которые в CSV как текст идут всё равно. Там все форматы текстовые и числа, числа тоже представляются в текстовом формате. Ну вот так вот примерно это происходит преобразование. Вот я советую сначала это сделать, осуществить такое преобразование. Это именно вот в той версии, которая днём записана, этот режим я добавил. А до этого писал там в инструкции, что возьмите заголовки, вытащите их в Word и проанализируйте. Получился выходной файл меньше 2 ГБ. Иногда он получается больше 2 ГБ. У меня такого случая не было. Для этого нужно там, ну чтоб я так примерно оценивал, где-то больше 12 миллионов строк должно быть в CSV файле, чтобы он был больше 2 ГБ DBF файл получился. Вот. Ну вот получился у нас DBF файл. Мы можем его открыть в Экселе. Делаем форматирование. Но здесь у нас есть где-то длинное наименование. Ну этот, этот CSV-файл с ошибками, ребята. В нём есть места, где поля, поля заданы не запятая разделитель, а точка с запятой. Вот я вам показываю. Вот это вот, видите, это два поля, которые объединены в одно, потому что там не разделитель, разделитель неверный. Вот. И дальше идут текстовая, текстовая информация. Вот. Ну, здесь довольно-таки проблематично будет понять, что вы тут, что у нас в качестве классов, что в качестве признаков. Ну, в принципе, ребят, такие файлы можно обрабатывать в режиме 2-3-2-2, задавать обработку текстов и признаками считать слова, которые есть в полях соответствующих. Вот архив там, JSON там и так далее. И можно их обрабатывать. Так, сейчас я подключу ещё один университет из-за накладки, из-за этой. Вот там пока никого нету. Ну ладно. То есть я сейчас показал, что можно ввести данные с этого, с этого CSV-файла, который вот по этой ссылочке, что вы мне прислали, Данил. Такое впечатление, что здесь, э-э, хмм, насколько я понял, здесь текстовая информация. А вот видите, так, здесь написано, собственно говоря, что вот описание этого, вы ж читайте, что тут написано. 200.000 научных статей, из которых 100.000 с полным текстом по ковиду. То есть это идёт обработка публикаций. То есть анализ публикаций идёт. Вот. Ну, в принципе, можно проанализировать публикацию, но у них должна быть, э, какой-то класс должен быть, к которому они относятся. И мы этот класс должны определять по словам, которые есть в этих публикациях, в их названиях или аннотациях. Вот. Пока я не понял. Вот это надо прочитать и понять, что они хотят извлечь из этого набора данных. То есть если есть желание, можно, в общем-то, и попробовать это сделать. Во всяком случае, я вам могу сказать, что вот на этом ResearchGate там была одна девушка, американка из США, которая спросила меня, можно ли обработать вот такого рода данные, очень похожие на те, которые сейчас вы привели. Даже чуть ли не те же самые. И я ей показал, как это сделать. То есть именно обработать текстовые данные. Она была, значит, там, ну, в общем, сказала спасибо, скажем так. То есть это возможно. Но не знаю, насколько это имеет смысл там, насколько это красиво получится. Но можно попробовать. Пожалуйста, ребята, вопросы ещё. Ну я думаю, что было бы предпочтительнее, ребят, всё-таки числовые данные обрабатывать. Знаете, почему? Модели с текстовыми данными, они получаются очень большой размерности. Там, да, кстати, для системы совершенно всё равно, на каком языке эти данные. На русском, на английском, на немецком, без разницы, она спокойно всё это одинаково обрабатывает. Можно даже и на всех сразу. Вот. Но, значит, если числовые данные, то там вводятся числовые диапазоны. А с числовыми диапазонами получаются размерность модели меньше. И тогда можно полный анализ проводить, там и кластерный анализ, и другие виды анализа получаются в разумное время. А вот с такими моделями, где там сотни тысяч слов, например, там, ну, кластерный анализ слов уже не проведёшь, например. Матрицу сходства 100.000 на 100.000 не посчитаешь, понимаете? Там необозримое время пройдёт на это. Вот, смотрите, ребята, значит, что можно сказать об этой модели, ребята? Значит, смотрите, значит, при увеличении уровня сходства закономерно линейно растёт число истинных решений. И это означает, что уровень сходства является адекватным критерием оценки достоверности решения. То есть это очень хороший признак, разумный. Это означает, что, в общем-то, модель неплохая. Теперь смотрим, какой у неё уровень достоверности по F-критерию - 0,718. Это, ну, как, на три балла, грубо говоря. Но, в принципе, если вот вы бы меня спросили, можно ли это брать, я бы сказал, можно. Но нужно посмотреть, не совпадает ли это ссылочка на другие уже выполненные работы, понимаете? Либо из тех трёх, если не ошибаюсь, тем по вину, там не было именно анализ красного вина по химическим свойствам. Ну, может быть, да, может быть. Я не говорю, что точно. Но если такого не было, то, пожалуйста. Вот. И тут уже можно смотреть, значит, можно смотреть, какие характеристики у вина самого высокого качества. Допустим, смотрим, что для них характерно там. Что у них высокое содержание алкоголя, низкое содержание чего-то там, я не знаю, что это такое. Density, низкая, сульфата, ну, сульфата можно догадаться. Высокая, что это такое? Цитрик тоже высокая. Значит, здесь можно что, пять градаций. Пятая, пятая, пятая - это очень высокая. Пятая - это, это четыре-пять - это высокая. Три-пять - это средняя, два-пять - это низкая, один-пять - это очень низкая. Вот. Это что очень, это что характерно для вина очень высокого качества. Ну, может быть, есть смысл перевести, но не в самой этой таблице экселевской, а просто для себя, чтобы понимать, что это такое вот, что это значит. Названия, конечно, скорректировать. Вот. Ну так неплохо. Я тогда числовые возьму. Да, сейчас вот я посмотрю классификационная шкала. Здесь у нас баллы 3 4 5 6 7 8 получилось, видишь? Вот. Ну надо и на то, и на то посмотреть. 3 4 5 6 7 8 Вот. Ну там достоверность-то выше у модели, ты ж видишь. Вот. Ну так вот. Таковы вот такие приколы бывают. В общем, интересно. Но там дело в том, что с баллами получается больше классов, а это интересно само по себе, когда больше классов. Тогда можно вот такие всякие формы делать, разглядывать их. Вот, допустим, пятая, четвёртая и третья оценка качества, они отличаются от шестой, седьмой и восьмой. То есть это как бы высокое качество, это низкое качество, это между ними сходство есть между четвёртой, третьей и шестой, маленькое сходство. Можно посмотреть эти дендрограммы, как это выглядит. А когда там три, например, у тебя будет их, понимаешь, или там пять, то немножко так слишком упрощённо всё получается. Вот. Ну, вот, смотрите, 2 минуты 26 секунд это заняло создание модели и проверка её на достоверность. И смотрим достоверность модели. Ничего непонятно. А ну-ка.

**VIII. Завершение и следующие шаги**

* **Итоги занятия**:
  + Всё, ребята, на этом мы занятие заканчиваем.
* **Задание на следующее занятие**:
  + Значит, вы что должны делать? Вы должны регистрироваться в РИНЦ, ResearchGate, находить тему и на следующем занятии мы будем рассматривать вашу тему. Для вас это тоже очень полезно, потому что вы видите реальный пример, когда мы прямо берём данные, скачиваем с сайта, от и до, понимаете? Вот это описание, вот, естественно, я вот перевожу на русский, читаю, о чём здесь идёт речь. Это прямо в статью, эту ссылку на данные тоже прямо в статью, в описание, я имею в виду. И всё, в общем, вы делаете по образцу вот этому с помидорами, только вместо помидоров у вас вино. И прекрасно у вас получится.