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

***Российская Федерация***

**197 Лабораторная работа №9. По дисциплине "Интеллектуальные информационные системы и технологии". 2020-11-11**

**Заголовок:** Лабораторная работа №9: Разработка собственных приложений в системе Эйдос и анализ данных

**Резюме:**

1. **Введение и Организационные Моменты:**
   * Занятие посвящено лабораторной работе №9 по дисциплине "Интеллектуальные информационные системы и технологии" для группы ЭТ 1823 (11 ноября 2020 г.).
   * Ведущие: профессора Луценко Е.В. и Аршинов Г.А.
   * Цель занятия: разработка собственных приложений в системе Эйдос.
   * Обсуждение текущего статуса студентов по выбору темы и началу работы.
2. **Проблемы Дистанционного Обучения:**
   * Сложность контроля и помощи студентам в освоении материала при дистанционном формате (отсутствие визуального контакта, непонимание реакции).
   * Дистанционные лекции частично заменяют очное обучение, но не позволяют видеть реакцию аудитории.
   * Лабораторные работы особенно страдают, так как требуют практических навыков, которые сложно формировать удаленно (аналогия с обучением плаванию).
   * Важность получения практических умений и навыков, а не просто прослушивания лекций или просмотра экрана.
3. **Цели Лабораторной Работы и Компетенции:**
   * Научиться создавать приложения в системе Эйдос.
   * Научиться описывать результаты исследований в принятой научной форме (литературной) и публиковать их.
   * Формирование компетенций: способность профессионально излагать результаты исследований и представлять их в виде научных публикаций (важно для аттестации и аккредитации).
   * Задание разработано для формирования этих навыков в условиях дистанционного обучения (выполнение задания под руководством, описание по шаблонам, размещение на площадках для получения статуса публикации).
4. **Выбор Темы и Работа с Данными:**
   * Обсуждение выбора тем студентами (многие еще не выбрали).
   * Преподаватель предлагает помощь в выборе темы.
   * Ограничения по выбору темы: не должна повторять уже выполненные, не противоречить морально-этическим нормам и законодательству РФ.
   * Ограничения по объему исходных данных: не более 10 МБ (лучше КБ) для разумного времени расчетов.
   * Источники данных: поиск в интернете (datasets), порталы UCI и Kaggle.
   * Формат данных: предпочтительно CSV (текстовый, поля разделены запятой или точкой с запятой), так как он экономный и поддерживается многими системами.
5. **Конвертация Данных для Системы Эйдос:**
   * Необходимость преобразования CSV в формат DBF (внутренний формат Эйдос).
   * Способы конвертации:
     + Встроенный CSV-DBF конвертер в Эйдос (режим 2.3.2.10).
     + Онлайн-конвертеры (рекомендуется Convertio за корректность и бесплатность).
     + Excel (не рекомендуется из-за возможных ошибок, например, замены чисел на даты).
   * Важность корректной обработки разделителей (запятая vs точка с запятой) и отсутствующих данных (NA заменять на 0 в числовых полях, но учитывать значимость нуля в контексте).
6. **Практический Пример Анализа Данных в Эйдос:**
   * Демонстрация на примере данных "Доноры крови и гепатит C" с портала UCI.
   * Подготовка данных: копирование в папку indata, переименование в input.data, замена NA на 0 в числовых полях Excel.
   * Ввод данных в Эйдос (режим 2.3.2.2): указание типа файла, номеров классификационных и описательных колонок, настройка интервалов.
   * Создание и анализ моделей (режим 3.5): оценка достоверности, просмотр результатов (кривые распределения, дендрограммы).
   * Выбор оптимальной модели (рекомендуется INF3 с критерием "сумма знаний").
   * Демонстрация на примере данных "Амфибии": сложности с обработкой текстовых данных и разделителями (точка с запятой), необходимость использования онлайн-конвертера.
   * Важность правильной интерпретации данных (например, 0 может быть значимым значением, а не отсутствием данных).
   * Предложение студентам взять разобранные темы для своей работы.
7. **Тестирование и Завершение:**
   * Обсуждение предстоящего тестирования студентов.
   * Преподаватель выражает готовность помочь с подготовкой, но признает нехватку времени.
   * Напоминание о пройденных лабораторных работах, включая анализ спектров изображений.
   * Подведение итогов: задание на разработку приложения – это способ проверить и научить студентов проводить исследования, описывать их и публиковать в условиях дистанционного обучения.

**Детальная расшифровка текста:**

**1. Введение и Организационные Моменты**

* **Приветствие и тема занятия:**
  + Здравствуйте, Георгий Александрович. Здравствуйте, ребята.
  + Здравствуйте.
  + Здравствуйте, здравствуйте.
  + Вот. Так что сегодня у нас 11 ноября 2020 года.
  + Первая пара 8:00-9:30. Лабораторная работа номер девять с группой ЭТ 1823 по дисциплине интеллектуальные информационные системы и технологии.
  + Занятия ведут профессор Луценко Евгений Вениаминович и профессор Аршинов Георгий Александрович.
  + Вот таким образом.
  + Занятие у нас посвящено разработке собственных приложений.
* **Проверка статуса студентов:**
  + По этому поводу, значит, у меня возникает вопрос, конечно, естественно. Как у вас с этим дело обстоит?
  + Пожалуйста, мне каждый ответьте. Вас не так много здесь человек.
  + Значит, Денис, давай, отвечай. Как у тебя стоят дела по разработке собственного приложения?
  + (Пауза, обращение к чату) В чате, пожалуйста, напишите: делаю там, допустим, тема утверждена, работаю. Или тема, нету темы там, допустим. То есть можно написать там плюс-минус. Нету темы. Вот. Молодец. Выбираю тему.
  + Так, нет темы, выбираю тему, не выбрал тему, выбираю, выбираю, нет темы. То есть у вас никого нет темы. Ни у кого нет темы.

**2. Проблемы Дистанционного Обучения**

* **Сложности контроля и взаимодействия:**
  + Значит, ребята, сейчас у нас дистанционная система обучения.
  + Вам это, конечно, ну, трудно, наверное, будет объяснить. Ну, в общем, смысл такой, что раньше, когда мы учились в аудиториях, то мы видели, как вы работаете на компьютерах, понимаете? Как вы работаете в системах, что вы научились, чему не научились, подходили к вам, подсказывали. Я не только о себе говорю, вообще о преподавателях.
  + А сейчас, например, я вас даже не знаю, как вы выглядите, понимаете? То есть хотя бы видео включили, что ли? То есть я вижу просто некие там надписи на экране.
  + Вот, и конечно, я совершенно не представляю, что вы там, чему научились, чему не научились.
* **Ограничения дистанционных лекций и лабораторных:**
  + Вот, когда мы читаем лекции дистанционно, то это напоминает, конечно, то есть в какой-то степени можно сказать, ну более-менее заменяет вот это личное обучение. Хотя, когда я читал лекции в аудитории, я видел вашу реакцию. А сейчас я читаю монитору, можно сказать, лекцию, понимаете? Никакой реакции с вашей стороны я не вижу, абсолютно. Это очень усложняет. То есть я не понимаю, как вы слушаете меня, не слушаете, там вы, не там вы.
  + Значит, сейчас, пожалуйста, камеры включите все.
  + (Реплика студента: Вы точно уверены, что хотите наши видеть лица в 8 утра?)
  + Да. Да. Давно никого не видел уже, студентов.
  + Вот, а что касается лабораторных работ, практических, то здесь вот это дистанционное обучение очень ущербно, ущербно. Почему? В чём это заключается ущербность? Вы же знаете, что для того, чтобы научиться плавать, нужно прыгать в воду и плавать, да?
  + Так вот на лабораторных работах вы должны получить навыки применения тех знаний, которые вы получили на лекциях. А для этого вы должны не слушать, что я рассказываю, не смотреть на монитор на меня там, допустим. Или вообще, может быть, вы и не смотрите, может там где-нибудь сидите там, спите, может быть, я не знаю даже там, что вы тут делаете.
  + Вот. То есть человек учится тому, что он делает. Вот вы смотрите на монитор, значит, вы учитесь смотреть на монитор, понимаете? Вот в чём дело.

**3. Цели Лабораторной Работы и Компетенции**

* **Формирование практических навыков:**
  + Вот, то есть, а вы должны получить соответствующие умения и навыки. То есть вы должны научиться делать приложение и научиться их описывать в литературной форме, принятой в науке, и публиковать.
* **Важность публикаций и компетенций для аттестации:**
  + Вот сейчас комиссия работает у нас, знаете, в аттестации, аккредитации университета. Летом они будут работать очень долго, я так сейчас вот слышал краем уха.
  + Так вот, они обращают на это внимание. Я очень много веду разных дисциплин и в бакалавриате, в магистратуре, в аспирантуре, огромное количество различных дисциплин.
  + И ко мне обращаются люди, чтобы там то сделать, то сделать. В частности, вот тесты. Обращаются руководители управлений, направлений учебных, программ образовательных, и говорят: надо сделать тест. И все говорят, знаете что? Вот они ж не сговариваются, у них есть какой-то общий источник информации. То есть от них этого требуют, и они тоже, соответственно.
  + (Демонстрация списка дисциплин на экране) Вот смотрите. Видите дисциплины, нет? Вот. Видите, да? Да. Вот, и сейчас ещё вам покажу.
  + (Переключение на другой список дисциплин) Вот, это вот в другом университете. Я просто нечаянно не с той, ну, в общем, короче, вышел из этого Вебекса. Ну, по-моему, продолжается всё, нормальная запись идёт и всё. У вас должно там, то есть те дисциплины сохранились, которые были в этом, в аграрном. А это, которая КГУ.
  + Вот, и ко мне обращаются с этим вопросом, я так вижу, что за вопросы, к чему они клонят. Вот, и все спрашивают о том, что именно вот студенты должны приобрести. То есть нужны тесты, которые позволят определить, приобрели ли студенты компетенции определённые. И компетенция - это именно в области проведения научного исследования. Сейчас я попробую даже вам зачитаю.
  + Способность профессионально излагать результаты своих исследований и представлять их в виде научных публикаций. Как вам это нравится? Это то, чему я вас пытаюсь научить, понимаете? То есть я пытаюсь именно это дать, именно эти компетенции вам и дать. Плюс там ещё проведение исследования само.
* **Обоснование текущего задания:**
  + Недавно я изменил, вчера буквально, изменил инструкцию задания. Почему вообще вот этот вариант, я пришёл к этому варианту, дать вам вот такое задание, чтобы вы выполняли его? Потому что это единственное, что сейчас можно придумать. Я вот даже другого варианта даже и не знаю. Как вот можно сейчас дать вам какие-то навыки работы в системе, проведение исследования, публикации, то есть описание его результатов, хода и результатов, публикации его результатов? Ну как вам такие навыки дать, когда я вас вообще не вижу?
  + Вот я пытаюсь это сделать каким-то путём. Что вот вы выполняете некое задание под моим руководством и с моей помощью, описываете его по шаблонам, которые я вам даю, тоже с моей помощью, и размещаете это на площадках, которые означают, что ваше исследование, описание этого исследования приобретает статус публикации.
  + Вот, собственно говоря, что мы, чем мы занимаемся. Я с вашей стороны вижу такое, скажем так, равнодушие к этому полное. Это мягко сказать.

**4. Выбор Темы и Работа с Данными**

* **Помощь в выборе темы:**
  + Значит, давайте мне сейчас, пожалуйста, информацию о том, кто из вас уже делает, а кто нет.
  + Вот, значит, сейчас мы этим и займёмся, ребята. Может быть, я кому-то написал, ребята, кто-то из вас мне писал письмо, и я написал, что рассмотрим это на занятии.
  + (Реплика студента: А нам темы самим себе придумывать или вы можете помочь с ним с выбором?)
  + Да, я могу помочь. Сейчас вот и буду помогать. Потому что раз вы сами там как-то не особо. Уже ж время-то какое? 11 ноября. Через там буквально недели две-три уже у вас не будет совершенно для этого времени.
* **Ограничения по теме и данным:**
  + Значит, смотрите, ребята. Значит, тема может быть любой, но есть несколько ограничений.
  + Первое ограничение - её не должно быть уже среди уже выполненных работ или какой-то очень похожей. Здесь вот список выполненных работ. Откуда он взялся этот список? Вот в этой инструкции задание, тут есть, видите, пункт четвёртый. И там вот есть ссылочка на уже существующие облачные Эйдос-приложения.
  + Второе. Значит, ваша тема не должна противоречить общепринятым в России морально-этическим нормам и действующему законодательству. Ну то есть она должна быть в пределах разумного в этом в этом плане.
  + И третье. Значит, исходные данные должны ни в коем случае не быть больше 10 МБ, потому что файлы больше 10 МБ на FTP-сервере сразу же стираются моментально. Просто такой хостинг там, так он работает. А вообще лучше вообще, чтобы они измерялись килобайтами, потому что тогда расчёты могут быть в разумное время выполнены. Если же там будет, скажем, 5 МБ исходные данные, то могут они занять, скажем, сутки эти расчёты. Вот. Смотря, сколько там исходных данных, там, какие модели, там, что там конкретно.
  + Поэтому вот такие вот есть ограничения. Чтобы не было среди существующих, чтобы объём был исходных данных не очень большим. И чтобы это было всё в пределах, так сказать, общепринятых норм морально-этических и в пределах законодательства, не противоречило законодательству.
* **Источники данных:**
  + Теперь, где брать данные? Значит, данные есть вот на сайте UCI. Ну, во-первых, есть вот наборы данных для машинного обучения - это запрос просто, где очень большое число различных задач. Кстати, сами эти данные называются dataset. Dataset. Вот.
  + То есть можно поиском найти. А можно найти на порталах данных, на которых эти специально на этом... то есть порталы, которые на этом специализируются - это портал UCI и портал Kaggle. Kaggle сейчас является уже довольно длительное время лидером в этой области. А UCI - наиболее древний портал, который когда-то, ещё в девяносто пятом году, видите, уже разместил данные для машинного обучения.
  + И вот здесь мы можем увидеть, ребята, что есть характеристика этих данных. Сколько там атрибутов, то есть сколько у нас признаков у объектов, и сколько объём обучающей выборки. И типы данных, и какая задача ставится. Вот категориальные данные - это текстовые. Целочисленные и действительные - понятно. Это числа, но без дробной части, без знаков после запятой. А это со знаками после запятой.
  + Ну вот обратите внимание, есть, допустим, задача, где 37.000 примеров наблюдений и 294 признака у объектов. Это довольно большая задача. Вот такие я советую вообще не смотреть даже, что там такое, понимаете? А есть вообще просто что-то кошмарное. Вот, допустим, 581.000 примеров. Вот, и 54 признака.
  + Значит, кроме того, я вам не советую брать старые данные, а лучше брать новые. Почему? Потому что новые данные, они в формате CSV.

**5. Конвертация Данных для Системы Эйдос**

* **Формат CSV:**
  + Что такое формат CSV? Знаете, ребята, нет? CSV формат.
  + Это, ребята, сейчас у нас лабораторная работа, поэтому вы должны со мной общаться. То есть я вас спрашиваю, знаете вы или нет? Вы должны говорить, знаете вы или нет. Что такое CSV формат?
  + (Реплика студента: Это для представления табличных данных, нет?)
  + Это текстовый формат. То есть файл досовского формата, то есть стандарта досовский текст. То есть это вообще текст без каких-то там стилей там, ничего, без разметки. Вот. Но он чем отличается от других текстов досовских? Тем, что там есть поля. Первая строка - это обычно наименование полей. И потом значения полей идут в последующие строки. Ну и размер CSV файла может быть значительно больше и по просто общий размер файла, и по числу записей, чем в Экселе, например. Вот Excel не позволяет больше миллиона записей, да, сделать. А CSV файл, он может быть там и 10 млн, и 20 млн, и там и 50 млн записей. Может быть, даже и 100 млн. То есть, ну сколько места на диске есть. Притом этот формат довольно экономный, потому что там нет никаких никакой разметки, никаких пробелов, ничего. То есть там подряд всё идёт. А разделены между собой значения вот этих полей в строках, разделены обычно запятой. Но иногда бывает точка с запятой. Понятно, да?
  + То есть это почти что как база данных. И большинство языков программирования с таким файлом могут работать, просто могут обращаться к этим полям по номерам и к строкам по номерам. Ну как с базой данных почти что, практически. Только единственное, нельзя её там отсортировать там, и по индексному массиву искать. Ну вот так вот просматривать можно. Вот. Но однако это происходит довольно-таки медленно, если большие CSV файлы по размеру.
* **Способы конвертации в DBF:**
  + Вот. И есть специальные средства, позволяющие это ускорить эти процессы. И, в общем, возникает вопрос о том, как преобразовать такую форму эти CSV файлы, чтобы можно было ими пользоваться в системе Эйдос.
  + Значит, есть три способа это сделать. Первый способ. В самой системе Эйдос есть CSV конвертер. Сейчас мы попробуем его использовать для... найдём данные и попробуем использовать. CSV-DBF конвертер. DBF - это внутренний стандарт баз данных в системе Эйдос. Значит, DBF файл - это тот же самый экселевский файл, только Excel 2. Ну, в общем, тогда, когда ещё Windows не было операционной системы, в ней уже был и Word 2, и Excel 2. И вот формат этого старинного совершенно Экселя, вот он, в общем-то, и это, по сути дела, и есть DBF. И Excel до 2003 года версии, они поддерживают DBF формат. То есть они могут его считывать, записывать. Вот, работают с ним практически.
  + Значит, второе, значит, способ. Значит, я сказал, первый способ - это в системе Эйдос есть CSV-DBF конвертер.
  + Второй способ - это сам Word или Excel. Ну, допустим, возьмём, допустим, Excel. Значит, там есть в Экселе вкладочка Данные, где появляется меню соответствующее, универсальное такое кнопочное со списками. И там есть в этом меню слева, там есть место, где можно выбрать ввод данных из текста. Excel вводит данные из CSV файла, но при этом у него есть погрешности при этом вводе. Погрешность в чём заключается? В том, что иногда, и довольно часто, числа заменяются на даты почему-то. Ну вот, допустим, такое число 11 11 20 20, да? Вот цифры такие там подряд идут. Он возьмёт, он напишет 11 там ноября, например, понимаете? Вот. Это очень неудобно.
  + И третий вариант конвертирования - есть онлайн-конвертеры. И есть довольно хорошие онлайн-конвертеры. Значит, я их разные видел много этих конвертеров. Ну, когда возникали вопросы, конечно, я искал их. И вот есть моё впечатление такое, ребята, что практически, можно сказать так, однозначно, что наилучшим является конвертио. Вот этот конвертер. Я это и в самой системе Эйдос советую, этот конвертер использовать. Во-первых, он бесплатный, без всяких там дай там или там ещё чего-нибудь там, или купи там, вот без этого, понимаете? Нет, там это есть, но очень так тактично и в меру, и, в общем, не раздражает. И, в общем, она бесплатно работает и позволяет конвертировать файлы до 100 МБ размером, что довольно-таки много. И очень редко встречается, когда это вообще может потребоваться. И он хорошо работает, у него результат работы корректный.
  + И он работает со всеми стандартами CSV файлов, а у них стандарты всё-таки разные есть, довольно разнообразные. А вот этот конвертер системы Эйдос, он работает с наиболее распространённым таким стандартом, вот, который я сказал, через запятую разделены.
  + (Реплика студента: То есть, в принципе, можно выбрать любой набор данных, который текст из себя представляет, да?)
  + Этот текст, да. И потом мы его должны конвертировать для ввода в систему Эйдос либо в самой системе, либо с помощью онлайн-конвертера. Excel я даже... не советую вообще использовать, потому что он плохо это делает, понимаете?

**6. Практический Пример Анализа Данных в Эйдос (Доноры крови)**

* **Выбор и подготовка данных:**
  + (Поиск данных на UCI) Значит, это аномалия рефракции, наблюдения. Вот, я сейчас смотрю, какие темы. Это вот у нас сложные, тут много данных. Данные о походке. 48 наблюдений. Значит, здесь вот выглядит таким образом. Вот. Dataset - это и есть исходные данные. Значит, здесь вот описывается задача. Вот. Набор данных был создан путём расчёта параметров ходьбы 16 различных добровольцев... Ну, давай посмотрим. 48 этих примеров. Характеристика классификации, нет данных. Посмотрим. То есть это вот подобные вещи, они в начале пишутся статьи. Где тут загружать, я вообще, честно сказать, не вижу. Кнопка данных. Dataset. И что там у нас? Что-то оно не очень-то это самое. Не CSV формат. Опыт был неудачный. Что-то оно уже и не хочет закрываться. Мутация эсовский файл. Нам это не надо.
  + Вот. Теперь назад возвращаемся и смотрим другие данные. Значит, выбираю из них те, где не очень много, не очень большой объём данных.
  + Что такое ВГС, кто-нибудь знает? Да. Когда будете описывать, ребят, задачу, то надо обязательно ссылаться на источник информации, на ссылочку вот, откуда взята, взяты исходные данные. Вот. Категория донор крови против гепатита C, включая... Вот. У нас сами эти публикации, этих авторов, нужно обязательно ссылаться на тех, кто это сделал.
  + Ну здесь вот у нас классический CSV файл, наверное. Сейчас мы посмотрим, насколько он классический. Вот. И смотрим мы вот сюда. Заходим в систему Эйдос, папочку indata. Копируем туда этот файл CSV. И преобразуем его, переименовываем его имя input data. Смотрим его просто в обычном текстовом редакторе. Значит, здесь, видите, в чём проблема? Что поля в кавычках. Да. Но конвертер, который в системе Эйдос, он эти кавычки убирает. Вот. Но вот кавычки, которые в колоночке самой, уже они останутся.
* **Конвертация и ввод в Эйдос:**
  + Ну давайте попробуем что-нибудь сделать с этим файлом. Значит, я сначала попробуем в самой системе Эйдос его преобразовать. Значит, система Эйдос должна быть у вас версии 7 ноября, ребята. 7 ноября. Что мы делаем? Мы сразу же удаляем все приложения в системе. А потом заходим в режим 2 3 2 10. 2 3 2 10. А это CSV конвертер. Видите, CSV-DBF конвертер. И смотрим. Не получилось.
  + Тогда я вам показываю 2 3 2 2 режим. Да, ну теперь надо выйти из системы. Теперь мы используем вот этот конвертер, который, раз не получилось, то мы используем конвертер, который есть, я уже сказал, онлайн-конвертер. Вот он по этой ссылочке находится сразу же. И вот мы им сейчас используем этот конвертер. Выбираем файл на диске C, система Эйдос, у неё есть папочка indata, и там находятся у нас исходный файл. Вот. Он подготовлен для конвертирования. И теперь мы его конвертируем. Вот. И скачиваем.
  + Смотрим, что у нас получилось. У нас получился вот этот экселевский файл. Расширение у него старого Экселя. Ну и вот смотрим. Вот прекрасно получился результат конвертации. 615 примеров.
* **Анализ данных и настройка параметров:**
  + Вот. Ну и что мы должны определить? Мы должны определить вот эту категорию доноров. Сейчас мы разберёмся, сколько. Ребята, я хочу подчеркнуть, у нас уже не CSV файл, а экселевский.
  + И вот мы смотрим. Значит, ну здесь возраст, пол. А дальше какие-то специальные параметры, похожие на медицинские какие-то показатели. Во всех колоночках должны быть имена. Значит, в первой колоночке может быть имя номер, например. Вот такое вот.
  + И вот мы должны посчитать, какие у нас колоночки соответствуют, ну являются классификационными шкалами, которые содержат то, что мы хотим определить на основе других значений других колоночек, которые являются у нас признаками этих объектов. А вот эта вот первая колоночка - просто номер наблюдения. Это анонимные, так сказать. Здесь не написано, кто это был за человек, этот донор. А просто номер написан.
  + И вот здесь вот у нас есть пол. М - это мужчина, наверное. Вот. А F - это, наверное, значит, женщина.
  + Вот. И вот мы должны посчитать эти колоночки. Смотрите, я ставлю курсор в клеточку А1. Нажимаю кнопочку Shift и двигаюсь вправо. Вверху здесь вот я вижу в формулу, которая показывает мне формулу этого диапазона, который я выделяю. Вот сейчас это вторая колоночка классификационная. Ну, третья, естественно, она описательная. И потом четырнадцатая идёт, видите? Я, соответственно, записываю себе, что у меня классификационные колоночки со второй по вторую, а описательные с третьей по четырнадцатую.
  + И выхожу отсюда в систему Эйдос. Да, файл должен оказаться по вот этому пути, для того, чтобы система Эйдос могла с ним работать. Сейчас я вам этот путь в чат отправлю. Там эта папочка специально для исходных данных.
  + Запускаем систему. Удаляем все приложения, какие там были Артёма созданы. Закрываем вот эти вот всякие окошки, которые открывались мой сайт.
  + И дальше мы используем API, автоматизированный программный интерфейс 2 3 2 2, ребята. 2 3 2 2. Это универсальный интерфейс ввода данных из экселевских файлов. Данные могут быть и числовые, и текстовые. Вот, значит, я указываю здесь тип файла сам. А вот здесь в CSV конвертере я написал, что вот если не получается, тогда вот использовать этот конвертер. Видите, вот конвертер. Вот. Потом указываю номер классификационной колоночки или диапазон классификационных колоночек. А потом указываю диапазон описательных колоночек. Вот. И вот здесь вот ставлю разные интервалы. Вот. Видите, как? Такие вот параметры. К сожалению, в чат я не могу это послать, потому что чат только досовский текст воспринимает. И нажимаю О'кей. Начинается конвертация этого файла.
  + Вот. Значит, у нас классификационная шкала одна обнаружена, в ней пять значений различных вот этой вот гепатита там, видимо, этого. А здесь шесть шкал у нас числовых и шесть текстовых. Сейчас мы с ними будем разбираться. И в описательных шкалах по пять градаций. При этом у нас получается матрица модели пять колоночек на 1.180 строк, что довольно-таки много.
  + Поэтому мы можем посмотреть здесь, какие шкалы у нас числовые, какие текстовые. Вот шкала классификационная, видите, у нас Blood Donor, Suspect Blood Donor там, гепатит, фибро... ну, в общем, короче, медицинские термины. И вот здесь вот мы видим, что у нас есть числовая шкала возраст. Данные в этой шкале распределены очень неравномерно. В основном это молодые люди и очень пожилого возраста. А вот в промежуточных возрастах сравнительно мало. Нет, наоборот. Очень молодых и очень пожилых мало, а промежуточных довольно много, поэтому получаются интервалы маленькие по знаку, по величине.
  + И вот тут мы смотрим дальше пол. Вот. А теперь смотрите, параметры, видите, 190 - 14,9, да? Это текстовые параметры. То есть 37,4 - текстовый. А всего 190 таких значений. То есть нам нужно... Почему они являются текстовыми? Потому что там в этой колоночке нет запятых, а там точки стоят. И здесь то же самое. Вот это текстовая колоночка.
* **Обработка ошибок и корректировка данных:**
  + Значит, давайте обратимся сейчас к режиму соответствующему. Ну, войдём до создания модели, пусть он конвертирует этот файл система Эйдос. И посмотрим сейчас на то, что нам надо сделать с этим экселевским файлом, чтобы система Эйдос воспринимала данные как числовые.
  + Вот. Значит, смотрите, ребята, видите, вот у нас здесь 546 колоночки, 547 NA написано. И вот 542, видите, да? NA. Ну, видимо, это отсутствие данных, я так думаю. Значит, нам нужно там, где отсутствие данных... Да, сейчас я вам покажу общий способ, как это определить. Заходим в самый конец данных, нажимаем равно, сумму. И она нам показывает, что до следующего NA, вот оно, видите, показывает блок. То есть Excel сам определяет, где там нечисловые данные. Вот. Ну, то есть мы находим, что у нас в этой колоночке есть нечисловые данные. Эти данные надо заменить на ноль. Потому что отсутствие данных - это ноль в числовых шкалах.
  + Теперь выделяю все колоночки боком и заменяю NA, заменяю на ноль. Нет, не все колоночки. Я выделяю числовые колоночки. И в числовых колоночках NA заменяю на ноль. Вот, и выполнена 31 замена. И вот здесь вот, где возраст, тоже это же самое делаю. Нет, нету никакого.
  + Ну, в общем, вот сейчас вот есть надежда, что всё это будет как числа восприниматься. Записываем обязательно. И смотрим равно. Находят ли система Excel, находит ли Excel нечисловые данные? Вот мы видим здесь колоночка C2, C616. Значит, не находит. Смотрим здесь вот равно E2 и E616. И вот так по всем колоночкам можно было бы проверить, но я так думаю, что особого смысла нет это делать, потому что у нас уже понятно, что мы сделали вот это. Убрали тексты.
  + Зачем же мы убирали тексты, ребята, из колоночек? Там было вот это было NA, мы его убрали. Зачем? Потому что система Эйдос, она, когда находит в числовой колоночке текст, то она считает, что эта колоночка уже текстовая.
* **Очистка и структурирование данных в Excel:**
  + Теперь смотрите, ребята, я ставлю курсор в клеточку А1, Shift и Ctrl держу, нажимаю End. У меня выскакивает, видите, 617 строка. Мне это не нужно. Я беру область данных, копирую, создаю новый лист и вставляю. То есть я вставил данные в новый лист. А вот тот лист старый, я его просто возьму да и удалю, сам лист. Для чего это я сделал? Для того, чтобы у нас область данных совпадала с фактическим расположением данных. Вот.
* **Повторный ввод и создание моделей:**
  + И повторяем попытку. Опять удаляем все данные, все модели. И вводим данные. Здесь все параметры запоминаются, которые мы вводили.
  + Вот. Теперь смотрите, ребята, что у нас получилось? У нас получилось колонок пять, как и было. А строк стало вместо 1.000, помните, там было 1.000 с хвостиком, стало 57.
  + Посмотрим. У нас получается на описательные шкалы по 123 наблюдения. Это более чем достаточно. Значит, адаптивные интервалы, они приводят к чему? К тому, что сами эти вот диапазоны числовые, они разного размера, но у них примерно одинаковое число наблюдений. Это очень разумный вариант. Получается достоверность модели лучше в этом варианте. Почему примерно одинаковое число наблюдений? Потому что, если у нас три землекопа копают две ямы, то на одну яму приходится не полтора землекопа, ребята. А на одну приходится два, а на другую один. Вот по этой же причине наблюдение будет там примерно одинаковое количество. То есть идёт речь о чём? О том, что не может быть полтора наблюдения или 0,7 наблюдения. Только целое число наблюдений. Поэтому задача в целых числах решается. Получается, то есть получается количество наблюдений очень близкое, практически равное в разных диапазонах числовых. Но может отличаться на единицу.
* **Анализ результатов моделирования:**
  + Теперь что произошло? Были созданы классификационные, описательные шкалы и градации. И с помощью них закодированы исходные данные. Что у нас получилось? У нас получились пять классов, которые были найдены во второй колоночке, текст уникальный. И найдено было колоночки описательные, все являются числовыми, кроме пола. Что и требовалось.
  + Смотрим на саму обучающую выборку. Вот у нас эти наблюдения. Вот их признаки. А вот можно в такой форме посмотреть ещё. Это наблюдение, это категория, к какому классу относится. Здесь уже коды мы видим. И коды признаков. Коды соответствуют этим вот справочникам, которые сейчас мы смотрели.
  + Теперь создаём модели в режиме 3.5. Здесь можно просто по умолчанию нажимать О'кей, да и всё. Значит, вот сейчас это не очень долго будет продолжаться, потому что я взял задачу, где не очень большой объём данных.
  + Вот. И давайте посмотрим на саму эту задачу. Я из этого источника данных копирую вам информацию об этом в чат, об этом источнике данных. Причём там, видите, чуть-чуть чат ещё и короче этого. Вот. Это описание полностью желательно копировать прямо вот отсюда начинать, и вот так вот раз, и всё в статью, в описание, там, где описывается, откуда взята таблица для расчёта модели и всего прочего.
  + Вот. Значит, тут у нас около минуты формировались модели и проверялись на достоверность. Теперь мы смотрим название задачи. Где это название задачи? Вот это название задачи надо в системе Эйдос в режиме 1.3 здесь вот название поменять. Вот.
  + И смотрим на достоверность модели. Значит, ну я своим опытным взглядом, когда всё это вижу, то могу сказать, что получилось, конечно, не самый идеальный вариант. Значит, ну посмотрим сейчас. Значит, нас больше всего меня интересует вот эта кривулечка. Значит, смотрите, ребята, у нас получилось классическое совершенно распределение. Что это такое? Значит, здесь вот мы видим уровень сходства положительный, здесь отрицательный. А здесь количество таких вот случаев истинных и ложных, положительных и отрицательных решений. Здесь положительные решения - это решения, когда уровень сходства больше нуля. Вот здесь отрицательные, когда уровень сходства меньше нуля. И мы видим здесь, что когда уровень сходства меньше нуля, всегда истинных решений больше, чем ложных. Синенькие - это ложные, а коричневые, красные - истинные. А вот при положительных решениях, начиная с определённого момента, где-то процентов 18 уровень сходства, доля истинных решений больше, чем доля ложных решений. Начиная с 30%, ложных решений практически нет.
  + Вот. Ну, в разных моделях можно это смотреть, как это получится. Здесь вот это менее удачно выглядит. Вот. А в каких-то других моделях по-другому. Здесь тоже неудачно. Тоже не очень удачно. Ну вот эта модель INF3, модель, она обычно самая такая лучшая в этом плане.
  + В общем, я рекомендую в данном случае модель INF3 с интегральным критерием сумма знаний. Получается 670 критерий достоверности при максимальном значении единица. То есть это, как скажем, не очень хорошо, но, в принципе, задача, видно, что задача решена. То есть получается разделить случаи истинных и ложных решений.
  + Ну дальше можно решать задачи. То есть вот на этом моменте я бы остановился и дальше сказал бы уже, что просто можно решать задачи дальше. Вот. Вот в этой модели INF3 мы можем посмотреть, какие доноры исходные друг с другом по показателям, какие отличаются. Вот. Можем это в виде дендрограммы посмотреть. И так далее. Дальше всё описано в этом шаблоне описания в статье, ребята. То есть можно вполне... Ну вот здесь мы видим, что вот эти вот похожие доноры, и вот эти похожие. Вместе они друг с другом похожи, но отличаются вот от этого все. Ну, в таком плане. То есть там дальше я не буду описывать, потому что всё это описано в шаблоне.
* **Предложение темы студенту:**
  + И теперь, ребята, значит, я могу вам предложить, собственно, я и предлагаю. Кто-то из вас эту задачу может взять. Задача вполне нормальная. Не очень большой размерности. Модель получилась. То есть можете взять эту тему. У кого есть, кому она приглянулась. Вот. Если кто-то её берёт эту тему, тогда мне он пишет письмо сейчас на почту.
  + Ну, говорит прямо сейчас вот в чате об этом. И присылает письмо на почту. Для чего? Для того, чтобы я послал уже экселевский файл преобразованный. И даже могу скриншоты послать с параметрами, какими начинать создание модели и дальнейшее уже продвижение.
  + Так что давайте, ребята, кто-то берите эту тему.
  + (Реплика студента: А повторите название темы, чтобы я вам в письме написал?)
  + А, название темы? Ну тут... Ну, в общем, АСК-анализ доноров крови, наверное. Ну, значит, давай так, я тебе напишу тему сейчас. (Пауза, проблема с компьютером) Что-то оно не очень-то это самое. Ну, вообще не работает, застряло. Типа пауза. Доноров... Человек сохранён. Человек сохранён. Видишь, что происходит, нет? Виснет. Фактически не работает. Ай-яй-яй. Может быть, я попал в спам. Вот. Ну сейчас я повторюсь немножко, чтобы параметры у тебя были. Кстати, параметры надо, чтобы у тебя были заданы. Вот. (Пишет в чат) АСК-анализ доноров крови на основе данных UCI. Вот. Ну сейчас я тебе напишу тему сейчас. (Снова проблема) Что-то оно не очень-то это самое. Ну, вообще не работает, застряло. Типа пауза. Доноров... Человек сохранён.