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

***Российская Федерация***

**45 Лабораторная работа № 2 по дисциплине "Интеллектуальные технологии и представление знаний". Вопрос - Лабораторная работа 3.03. 2020-11-19**

## Заголовок

Обзор системы Aidos: Установка, Лабораторная работа 303, Шкалы и Моделирование данных.

## Резюме

Сессия посвящена работе с интеллектуальной системой Aidos для студентов группы 412. Занятие началось 19 ноября 2020 года.

**1. Установка и Запуск Системы:**
Студентам было показано, как обновить систему Aidos до последней версии, извлекая файлы из архива в свою рабочую папку (рекомендовано X:\Program Files\Aidos). Запуск системы осуществляется через файл StartAidos.exe. При первом запуске в новой папке система регистрирует свой путь. Вход в систему не требует логина/пароля по умолчанию, но их можно установить позже.

**2. Загрузка Лабораторной Работы 303:**
В режиме 1-3 системы Aidos студенты научились скачивать необходимое приложение (лабораторную работу 303) из облака, так как локальные приложения не были установлены для экономии места. После скачивания и установки приложения работа продолжилась в интерфейсе ввода исходных данных.

**3. Работа с Данными и Шкалами:**

* **Ввод данных:** Исходные данные для ЛР 303 загружаются из Excel-файла (режим 2-3-2-2). Структура файла: строки – наблюдения (объекты), столбцы – признаки (свойства/факторы).
* **Типы шкал:**
	+ **Описательные шкалы** (столбцы без фона): Содержат значения факторов или свойств объекта (например, числовые значения фин. показателей, текстовые значения цвета). Могут быть числовыми или текстовыми (номинальными). Для числовых шкал система автоматически определяет диапазон и делит его на интервалы (по умолчанию 10 равных).
	+ **Классификационные шкалы** (столбцы с желтым фоном): Определяют принадлежность объекта к определенному классу или категории (например, уровень успешности предприятия, наличие/отсутствие экрана).
	+ **Порядковые шкалы:** Особое внимание уделено корректному представлению порядковых шкал (например, размер). Рекомендовано использовать префиксы (1-маленький, 2-средний, 3-большой) для обеспечения правильной сортировки и интерпретации системой, так как простая алфавитная сортировка нарушает смысловой порядок.
* **Пример:** Данные могут представлять предприятия, где описательные шкалы – фин. показатели, а классификационные – уровень успешности (отлично, хорошо, ..., банкрот).

**4. Моделирование и Анализ:**

* **Формирование модели:** Система Aidos на основе обучающей выборки (исходных данных) автоматически формирует классификационные и описательные шкалы, кодирует данные и строит модели (статистические и системно-когнитивные), выявляя причинно-следственные связи.
* **Просмотр данных и шкал:** Режимы 2-1 (классификационные шкалы), 2-2 (описательные шкалы) и 2-4 (просмотр закодированных данных) позволяют изучить результаты формализации.
* **Нормализация:** Процесс кодирования и создания справочников для повторяющихся значений (как в базах данных) называется нормализацией, что обеспечивает целостность данных.
* **Синтез и Верификация:** Режим 3-5 запускает синтез моделей и их верификацию (проверку достоверности). Система оценивает достоверность моделей и позволяет отсеивать наименее достоверные результаты. Для больших данных возможен выборочный анализ (Bootstrap).

**5. Производительность:**
Отмечена возможность использования графического процессора (GPU) для ускорения расчетов (синтеза и распознавания) по сравнению с центральным процессором (CPU). Ускорение может достигать тысяч раз, что критично для больших объемов данных. Однако совместимость GPU (особенно не Nvidia) с используемыми технологиями (OpenCL/CUDA) может быть проблемой.

**6. Итог:**
Студенты прошли этапы установки, загрузки ЛР, ознакомления со структурой данных, типами шкал и базовыми режимами работы системы Aidos для последующего анализа и моделирования.

## Детальная расшифровка текста

### Введение

Сегодня 19 ноября 2020 года, вторая пара, 9:40–11:10, с группой 412. Лабораторная работа номер два по дисциплине "Интеллектуальные технологии представления знаний".

Вопрос, который мы будем изучать сейчас, — это лабораторная работа 303.

Мы изучаем сейчас, начинаем рассматривать эту работу с того, что такое классификационные и описательные шкалы, градации, обучающие выборки и так далее.

### 1. Установка и Запуск Системы

#### 1.1. Обновление и Запуск

И сейчас я... вы сделали демонстрацию экрана у себя. И вы должны запустить сейчас файл StartAidos.

Каждый запускает систему в своей папочке, ребята. На диске X, Program Files.

Так, к сожалению, я не вижу ничего. Ну сейчас попробуем.

Так, ещё, пожалуйста... ещё раз, пожалуйста, поделитесь экраном. Потому что я там некоторые вещи сделал, которые... Ну, в общем, экран покажите свой. Чтобы я его видел.

Вот. Напишите "Извлечь", нажмите кнопочку "Извлечь". Извлечь. И "Для всех". Вот сейчас вы обновите систему до текущей версии. И все точно так же должны сделать.

Потом запустится сама система. Версия должна быть 17.11.2020.

Это нормально.

#### 1.2. Вход в систему

О'кей нажать? Да. О'кей нажимаем.

Что мы вводим в логин и пароль? Ничего. Ничего, ничего не вводим. Просто нажимаем О'кей. Если мы там что-нибудь введём, но если это будет неправильно, тогда она нам выдаст соответствующую информацию. Если это будет правильно, мы войдём в систему, и потом в системе можем поменять имя и пароль, и потом придётся их вводить именно те, которые вы введёте в системе. То есть вы можете войти первый раз в систему, как хотите, а то есть просто О'кей нажав, а потом можете сделать так, чтобы чтобы нужно будет для входа вводить имя и пароль. Но, значит, можно вообще ничего этого не делать. Просто таким образом вы закроете систему для других пользователей.

Значит, ребята, значит, каждый запускает в своей папочке, соответствующей номеру в списке.

### 2. Загрузка Лабораторной Работы 303

#### 2.1. Начальная Настройка и Загрузка из Облака

Значит, можете закрыть спокойно вот этот вот мой сайт. Нажать О'кей. Это просто информация о том, что система впервые обнаружила себя в этой папке и прописалась по пути на эту папку. Вот. Ну, мой сайт, который открывается, можно закрыть. Он открывается при запуске.

Вот. Теперь дальше, что мы должны сделать? Мы должны перейти в режим 1-3. У всех обновилась, ребят, система?

Ну, значит, и вот здесь вы должны выбрать внизу "Скачать приложение из облака". Потому что у нас локальных приложений нет. Дело в том, что они там вообще-то в инсталляции есть эти локальные приложения, но Геннадий, системный администратор, установил минимальную версию, потому что эти приложения все занимают место, и было бы не 40 МБ, а 120 МБ на копию. А это уже перебор, когда их когда их 40 штук там на сервере. Для каждой станции вот этой вот сетевой своё есть своя папочка, в которой система именно для этой станции. Ну или, вернее, для студента, как вы поняли, да, потому что мы по списку берём номер из списка.

Сейчас нажимаете кнопочку "Скачать приложение из облака". Это в режиме 1-3. Внизу есть кнопочка, где жирным шрифтом написано, правее "Скачать приложение из облака". Вот.

#### 2.2. Выбор и Установка ЛР 303

И в самый верх списка поднимаемся. Сейчас есть 211 облачных приложений в системе. А нам нужно 303, лабораторную работу 303. Ставим курсор на эту работу. И справа внизу там есть "Установка Aidos приложения".

Это Unix-овский каталог того, что будет сейчас скачиваться. Закрываем просто крестиком его. Скачивается это приложение, устанавливается.

Потом сразу же, значит, что там написано? "Вероятно содержит..." Ну да. Значит, О'кей нажимаем.

### 3. Работа с Данными и Шкалами

#### 3.1. Интерфейс Ввода Данных (Excel)

И сейчас запустится универсальный программный интерфейс 2-3-2-2 ввода исходных данных из внешнего экселевского файла. Это наша базовая лабораторная работа, которую вот мы изучаем подробно. Ну, по-моему, это всё я уже рассказывал вам и показывал на своём экране, своего компьютера. Вспоминаете это, нет, ребята? Или вы первый раз видите вообще всё это?

Вроде было.

Ну, это я всё объяснял, поэтому, то есть у меня-то записано, что я это объяснял. Вот. Ну, ну я это объяснял в другом режиме. Вы не сами это нажимали, всё и делали, а смотрели на экран моего компьютера.

Вот. Да, совершенно правильно. Значит, у вас там запустился Excel, как я понимаю.

Теперь надо... смотрите, значит, в этом окошке мы видим, что у нас обнаружена одна числовая шкала описательная. Она спрашивает, сколько числовых диапазонов там в этой шкале делать. Ну, сейчас мы просто выходим на создание модели, кнопочку нажимаем. Если мы поменяем число диапазонов, тогда надо нажать левую кнопочку "Пересчитать".

Вот теперь смотрите, ребята. Значит, у нас сформированы классификационные описательные шкалы и градации на основе базы исходных данных. И исходные данные закодированы с помощью этих вот классификационных описательных шкал и градаций. И нажимаем О'кей.

И вот на этом пункте мы и остановились. Вот здесь вот О'кей не нажимаете, а читайте внимательно, что здесь написано. "Для дальнейшего выполнения необходимо..." Вот. Значит, нужно открыть файл исходных данных по пути... Она даже пишет, какой путь, видите? Вот по этому пути откройте этот файл исходных данных.

#### 3.2. Структура Данных в Excel

Значит, что мы здесь видим? Значит, мы видим, что у нас есть файл экселевский, у которого строчки — это наблюдения. Ну, в данном случае, это вот работа, она чему посвящена? Это простейшая лабораторная работа, в которых мы видим, как система может определять... сначала она формирует на основе примеров описаний обобщённые образы, классы, а потом сравнивает конкретные объекты с этими классами и решает ряд других задач.

Вот у нас здесь каждая строчка — это наблюдение какого-то объекта. Ну, вы экономисты, поэтому я вам могу сказать, что вот если, допустим, вы будете анализировать предприятие с помощью этой технологии, а есть ряд статей и книг о том, как это хорошо получается, и с примерами, то, значит, у вас строчки — это будут наблюдения предприятий. Предприятия могут быть самые различные. Вот, какие угодно.

#### 3.3. Описательные Шкалы (Факторы/Свойства)

И у них вот эти колоночки, которые без фона, они могут иметь смысл такой, что это какие-то их, этих предприятий, финансово-экономические показатели. Ну, допустим, значения на счетах бухучёта, например, или субсчетах.

#### 3.4. Классификационные Шкалы (Категории)

А вот эти с жёлтеньким фоном колоночки — это так называемые классификационные шкалы, в которых описано, к каким классам относятся эти предприятия, к каким категориям. Ну, например, если мы хотим прогнозировать банкротство или идентифицировать банкротство, то мы можем сделать там такие градации: отлично работающее предприятие, ну, скажем, пять градаций. Отлично работающее, хорошо работающее, средне работающее, плохо работающее, отвратительно работающее. И шестая градация — вообще не работающая, обанкротившееся предприятие. И вот мы приводим, допустим, там 20-30 примеров или 100, или 300 примеров этих предприятий разных таких вот, разной успешности деятельности.

И система выявит взаимосвязи, если они существуют, между прочим. А если их нет, то она их и не выявит. Между показателями финансово-экономическими этого предприятия и успешностью его деятельности. И мы можем в результате по этим показателям определять успешность деятельности предприятия и вырабатывать рекомендации, что ему делать, чтобы оно нормализовало свою работу.

Вот таких вот колоночек, их может быть очень много. В новых версиях Excel до 16.000. Вот. Вот эти шкалы, которые на белом фоне, называются описательные. Это факторы и их значения, либо свойства и их значения. Вот, скажем, свойство "цвет", а значение "чёрный", например.

А вот эти на жёлтеньком фоне — это классификационные шкалы и градации. Они описывают принадлежность предприятия к каким-то категориям. Можно также точно не банкротство прогнозировать, скажем, или успешность деятельности, а можно определить направленность деятельности или масштабность предприятия и другие характеристики предприятия тоже можно точно так же изучать. Причём можно всё это одновременно изучать. То есть одна шкала вот у вас классификационная, а может быть успешность деятельности, а вторая может быть там, к примеру, объём деятельности, а третья — направленность деятельности. Понятно, да? А четвёртая может быть город, где это предприятие расположено. Вот. А направленность — это отрасль или подотрасль, тип предприятия: торговое там, или производственное, там, в какой области оно работает и так далее.

#### 3.5. Идентификатор Наблюдения

Первая колонка не является шкалой, и там просто содержится информация о том, откуда взята информация от данного наблюдения. Ну это простенькая работа, она требует некоторого воображения. Здесь вот то, что я сейчас вам рассказал, надо чуть-чуть напрячь воображение, что это могут быть либо объекты, описанные их свойствами и принадлежностью к каким-то обобщающим категориям, либо это могут быть не какие-то предметы, а это могут быть системы довольно сложные, такие как предприятие или поле, например. Вот здесь вот в аграрном университете, там могут быть поля, например, строчки. Вот эти вот строчки... колоночки, которые без фона — это агротехнологии и природно-климатические факторы описывают они. А вот эти жёлтенькие описывают количество, качество продукции, допустим, полученной на этом поле. Вот. И можно указать также финансово-экономические результаты деятельности, скажем, там, ну, прибыль, рентабельность с этого поля полученную. Понимаете? В результате мы можем создать модель, которая будет нам показывать, как влияют технологии и природно-климатические факторы на количественные, качественные и финансово-экономические результаты деятельности. Вот. Ну, поле можно считать таким предприятием, мини-предприятием тоже в этом смысле.

#### 3.6. Обзор Шкал в Aidos (Режимы 2-1, 2-2)

Вот. Ну, примерно вот так. То есть это вот я вам сейчас рассказал про эти шкалы и градации. Теперь, не закрывая этой формы, на панели задач, на вот этом иконочке системы Aidos, такая спиралька цветная, на ней нажимаем и выбираем, значит... Значит, здесь окна надо закрывать в порядке убывания. То есть последнее окно надо первым закрывать, как как в этом, в стеке. Вот открыли окно, теперь закрыли его, да. Вот все окна позакрывали, кроме главного. Главное окно.

Ну можете сейчас вот прямо сразу на главное окно выйти. На самую верхнюю строчечку. Вот. И теперь вы можете посмотреть на то, как в системе Aidos это выглядят эти классификационные описательные шкалы и градации. Для этого мы нажимаем, выбираем режим 2-1.

И видим здесь вот слева, видите, названия классификационных шкал. Они те же самые, которые в экселевском файле. А справа мы видим уникальные текстовые значения, которые были в этих шкалах. Шкалы могут быть текстовые и числовые. Если вы поставите курсор на "Обобщающий класс", вы тоже справа увидите уникальные текстовые значения, которые в этом, в этой шкале встретились. Видите? Они здесь просто по алфавиту рассортированы. Сбоку мы видим их коды. Они, всё это, всё это присваивается автоматически и выясняется, выявляется. Справа там ничего особо интересного нету, ну, при когда модель уже создана, то там кое-что появляется.

Закрываем эту форму. Да, на крестик просто. И открываем 2-2. Сейчас мы делаем фактически то, что у вас там было написано на этой экранной форме, где было написано "Открыть, открыть файл вот этот, который вы открыли", а потом было написано 2-1, 2-2 посмотреть. Помните, да, там было? Вот мы сейчас это и делаем.

Значит, теперь возьмите сейчас, отсюда выйдите из этого режима. Зайдите в режим 1-3 и нажмите там "Помощь", самая левая кнопочка внизу. И вот видите, вот у вас здесь перечислено всё, что я сейчас вам рассказываю. Что мы ввели данные в режиме 2-3-2-2, потом посмотрели классификационные шкалы 2-1, потом описательные шкалы 2-2. Мы ещё не успели их толком посмотреть. И будем дальше всё вот это делать.

Закрываем это. И здесь мы видим, ребята, что у нас вот здесь уже есть одно приложение. Название приложения автоматически присвоено, потому что это лабораторная работа. Если уже будете сами создавать приложение, тогда надо будет самим название это задавать. Вы можете там клацнуть два раза и поменять это название. Вот. Ну, если нажмёте Insert, тогда вот этот курсор синенький исчезнет, это будет вставка работать. А так работает замена символов. Ну, понятно. Закрываем эту форму, выходим в режим 2-2.

#### 3.7. Анализ Описательных Шкал (Цвет, Размер, Наличие Экрана)

Значит, здесь мы видим текстовую шкалу "Цвет". Потом слева двигаем дальше. Тоже названия шкал полностью взято из экселевского файла. Здесь всё по алфавиту расположено. На любом языке это может быть информация. То есть если в исходном файле, допустим, английский язык, значит, здесь тоже будет всё по-английски. Потому что берётся всё оттуда полностью.

Вот. Если какие-то вопросы есть, спрашивайте. Лабораторная работа, вы имеете полное право. И на лекциях тоже имеете полное право спрашивать.

Значит, теперь смотрите, ребята. Смотрим следующие шкалы: Материал, Размер 1. Вот Размер 1 нажмите. Значит, что мы здесь видим? Размер 1. Мы видим, что у нас тут градации "Большая", "Большой". Как вы думаете, это разные или одно и то же? Ну, в общем, русский язык богат. В русском языке мы говорим не только о размере, но и о размере какого-то объекта мужского или женского рода одновременно получается. То есть "Большой" — это как бы мужского рода, "Большая" — женского рода, да? А я вам советую такие вещи не делать. То есть если вы, допустим, в этом файле экселевском пишете "Размер", то пишите там всё, допустим, в единственном числе мужского рода, в именительном падеже там все названия, понимаете? Единственного числа мужского рода. Почему? Ну, если объекты, то они могут быть и женского, конечно. А вот свойства лучше всего писать так, чтобы они... Ну, ясно, что это означает одно и то же, "Большой" и "Большая", ну, в смысле размера. Вот. Но здесь это лишнее, что оно повторяется, да ещё, ну, с разными, по-разному, разного написания.

И мы здесь видим, что при сортировке по алфавиту здесь нет никакого смыслового порядка. У нас порядок должен быть такой: минимальный код должен соответствовать минимальному значению размера, потом увеличение, увеличение и максимальный код — максимальному значению размера. То есть чтобы это произошло, нужно в экселевском файле, в колоночке "Размер 1", эту колоночку выделить блоком и в ней поменять: "Маленький" поменять на "1/", 4 - "маленький". Она во всей колоночке поменяется это, да, сразу же. Потом "Средний". Берём "Среднее", меняем на "2/", 4 - "среднее". Она поменяет во всей колоночке "Среднее" на это вот дело. "Большой" поменяем на... "Большая" на "Большой", а "Большой" поменяем на... А, подождите, что-то я не то сказал. "Большой" поменяем на "4/4-большой". А "Средний" поменяем на "2/4-средний". Да? "Маленький" поменяем на "1/4-маленький". А "Под руку", ну, наверное, "3/4-под руку". И у нас получатся при сортировке по алфавиту правильный смысловой порядок. В этом случае это будет уже порядковая шкала.

Значит, какая разница для модели? Порядковая она или номинальная? Никакой разницы нет. Модель будет одинаково достоверна, всё будет одинаково нормально, хорошо работать, но при этом у вас получится, что при выходных формах легче будет интерпретировать результаты, если будет шкала порядковая, более осмысленные будут результаты.

Смотрим на "Размер 2" теперь. Там у нас числовая шкала. Найдено минимальное и максимальное значение в этой колонке "Размер 2". И весь этот диапазон изменения величины разделён на заданное число числовых диапазонов. Здесь эти диапазоны равного размера. А в системе есть возможность сделать их разного размера, но так, чтобы было одинаковое число наблюдений. Второй вариант лучше обычно, но где-то такой, наиболее простой вариант. Число этих диапазонов, я вам показывал, можно изменять. Там оно по умолчанию 10 стоит, можно сделать пять там, или три там, как хотите.

Теперь смотрим "Наличие экрана" шкалу. Есть, нет. Значит, может быть, когда-нибудь вы видели такие анкеты, где, если вы согласны с каким-то пунктом анкеты, вы там птичку ставите, да? Правильно? Да. А если не согласны, то тогда вы смотрите, читаете следующий пункт и так далее. То есть вы ставите птички только там, где вы согласны. Вот. И при этом вы, может быть, не обращали внимание особого, что там нет варианта, если вы не согласны. Вы с такими встречались анкетами? Что вот вопрос, допустим, собираетесь ли вы там покупать какой-то товар? И если собираетесь, там, ну написано там такое-то вот как прямо ананас, например. Вот вы пришли в магазин, собираетесь ли вы покупать ананас? И вы ставите птичку: "Собираюсь", значит. Понимаете, да, о чём я говорю? Угу. А если не собираетесь, то вы не ставите птичку. И сдаёте эту анкету на обработку. Анкета интересов покупателей, да? Вот. А там менеджер возьмёт и поставит птичку, что вы собираетесь покупать ананас. За вас. Вы поняли, да, о чём я говорю? Угу. То есть смысл какой? Что в анкете должно быть явно указано вариант ответа. То есть, если вы ответили "Да", то значит, надо там птичку ставить, где "Да". Если "Нет", то птичку, где "Нет". А вот тот вариант, который я сейчас рассказал, он соответствует такому, какой ситуации? Что если вы согласны, то вы ставите птичку. А если вы или не согласны, или просто не хотите отвечать, или просто у вас не было времени отвечать, то вы не поставили птичку. Это, значит, означает... если птичка отсутствует, то это не означает, что вы не собираетесь его покупать. Вот в чём всё дело. Это может означать, что вы просто этот вопрос пропустили. То есть это разные вещи вообще-то. Поэтому нужно обязательно в явной форме указывать и положительный ответ, и отрицательный. Вот в системе Aidos так и сделано здесь. Но это зависит от того, как сделаны исходные данные. Там может быть, если "Да", то написано "Да", а если "Нет", то там может быть ничего не написано. Должно быть написано "Нет" явно, понимаете? Вот на это я ваше внимание обращаю.

Закрываем эту форму.

### 4. Моделирование и Анализ

#### 4.1. Просмотр Данных в Режиме 2-4

И смотрим на форму 2-4-1 режим. Я их так пронумеровал, чтобы легче было вам говорить, что там открывать.

Ребят, а скажите, а на других компьютерах вы тоже сейчас запустили систему? Там тоже обновили вы, да, её? Да. Вот так вот скачался архивчик маленький там, развернули его, да? Да. Поэтому у вас версия 17.11, да?

Значит, теперь смотрите, как устроена вот эта вот форма. Верхняя часть — там у нас информация из первой колонки, где написано, откуда взяты данные данного наблюдения. Там довольно длинное такое текст может быть. Там может написано быть: Предприятие, название предприятия, в каком году там, ну, в общем, в каком городе. То есть там может быть вообще хоть адрес можно написать. То есть там есть место. Вот. Ну, при этом надо, конечно, в разумных пределах всё это стараться делать. Ну вы сами увидите в выходных формах. Это присутствует информация.

И вот в левом нижнем окошке — там коды классов, в соответствии со справочником, который мы смотрели в 2-1 режиме. А в правом нижнем окошке — там коды признаков, которые мы смотрели в режиме 2-2. И вот таких объектов может быть... Ну я испытывал систему на исходных данных, где было 4 млн записей. Вот вы сейчас можете взять, поставить курсор в верхнее окошко и Ctrl+Page Down нажать. Ну на любое поле. Ну, там, где вот единичка или название. И нажать Ctrl+Page Down. Вот, и у вас скакнуло, получилось 20 объектов обучающей выборки. А у меня получалось 4 млн объектов. Вот. То есть такие данные, их из Excel не введёшь, там больше миллиона не получится. Вот, а из CSV файла можно или DBF файла. И там можно и больше. Значит, я оценивал возможности системы Aidos, вот этой реализации, которая сейчас, примерно 12 млн объектов обучающей выборки она позволяет обрабатывать. Но это будет долго, ребята. Просто такие объёмы информации, они будут уже долго обрабатываться. Там уже речь о гигабайтах идёт.

Закрываем эту форму.

#### 4.2. Обзор Процесса Моделирования (Режим 6-4 и Помощь)

И смотрим режим 6-4. Не вижу, чтобы вы это сделали. Ага. Да, 6-4. Вот. Открывается... Ну это можно на весь экран развернуть, иначе там нечитабельно будет.

Ну, разверните на весь экран. И, значит, вы здесь вот можете увидеть... Ну здесь нужно Ctrl+ нажать, наверное, чтобы как-то по ширине выровнять. Вот там вверху кнопочка есть такая... Самое правое — это лист, смотрим лист. Нет, правее, правее. Ну, левее теперь. Вот, значит, там, где у вас проценты написано, а правее там по ширине кнопочка есть. Вот это, да. Вот. И, значит, прокручиваете вниз.

Значит, вот сейчас вы видите что? Что здесь... Что делает система Aidos, ребята? Вот это... эллипс покажите, эллипсоидную форму эту. Значит, чуть-чуть прокрутите, чтобы её видно было. Ой, Владимир Борисович, у меня занятие идёт.
*[Телефонный разговор об отправке статьи в журнал ВАК по определённым специальностям и условиях публикации. Уточняется, что номер будет декабрьский, но выйдет в следующем году, справку можно получить сразу.]*

Вот. Так, смотрим. Что делает система Aidos, ребята? Она берёт исходные данные, вот типа тех, которые в этом экселевском файле, находит там причинно-следственные связи. А чтобы это сделать, сначала создаются справочники, а потом кодируются исходные данные с помощью справочников, а потом обнаруживаются причинно-следственные связи между событиями, которые в этих вот, в этой обучающей выборке описаны.

Когда это всё сделано, то тогда исходные данные преобразуются в информацию, то есть появляются базы моделей. То есть когда вот исходные данные, там эти справочники, плюс база событий, плюс потом модели, отражающие взаимосвязи, то это уже всё информация, всё вместе. И вот эта информация может быть использована для решения задач различных, которых довольно много, но их можно разделить на три группы основных: это идентификация, прогнозирование — это одна группа; принятие решений — вторая, и третья — исследование моделируемой предметной области. Если мы решаем задачу принятия решений для достижения цели, то есть для управления, то тогда информация уже считается, называется знаниями.

Теперь прокрутите вниз, пожалуйста. Это в общем, то, что это делает система Aidos. И потом на основе знаний позволяет решать задачи. Следующую страницу, пожалуйста, откройте. Здесь тоже можно по ширине... Или и так вы видите, я не знаю. Наверное, вполне видно, в общем-то.

Вот первый пункт, серенький, — это неавтоматизированный пункт. На этом этапе мы решаем, что у нас в качестве будет причин или факторов, а что в качестве результатов их действий рассматриваться. То есть мы как бы ставим задачу на смысловом уровне. А вот потом следующий этап — формализация предметной области. Мы сейчас только что его выполнили, ребята. И в системе Aidos это полностью автоматизировано. Это одно из важных достоинств системы, которое делает её практически применимой в широком круге областей. И это нетрудоёмко. Я могу вам сказать, что разработка диссертации, скажем, кандидатской, исследование, оно может год занимать вот этот этап формализации предметной может занимать год: сбор данных, кодирование, внесение в систему, потом, чтобы обработать это всё, понимаете? Это очень трудоёмкий, сложный этап. В системе Aidos он происходит за минуты там, секунды, минуты, ну, если очень много данных, то за час, например, там, за полтора, понимаете?

#### 4.3. Синтез и Верификация Моделей (Режим 3-5)

Потом выполняется синтез и верификация моделей. Что под этим понимается? Синтез моделей — это создание моделей статистических. Потом я буду вам их показывать, объяснять, как они, что они собой представляют. А потом на их основе рассчитываются системно-когнитивные модели или модели знаний. Потом выбирается наиболее достоверная модель из них по критерию F-критерию Ван Рисбергена и его обобщениям. И потом в этой наиболее достоверной модели решаются задачи распознавания, идентификации, прогнозирования, принятия решений и исследования моделируемой предметной области. Теперь прокрутите вниз чуть-чуть.

Эти задачи, они друг от друга не отделены совершенно непроходимой какой-то стенкой, они между собой взаимосвязаны. Когда мы будем их рассматривать, я тогда про эти взаимосвязи расскажу. Это вот общая логика, так сказать, работы в системе, ребята. Что она вообще делает.

Закрываем эту форму. И теперь мы это и делаем, то есть переходим к следующему этапу. Следующий этап и в соответствии с хелпом, который я показывал в режиме 1-3, и в соответствии с этим вот хелпом, который графически я показал сейчас в режиме 6-4. Это следующий этап — это синтез моделей и их проверка на достоверность, то есть их верификация.

Значит, для этого мы запускаем режим 3-5. В режиме 3-5, не спешите. Значит, вверху написано, что мы создаём три модели статистические и семь моделей знаний. Они здесь прямо с названиями приведены своими.

#### 4.4. Использование CPU/GPU

Теперь мы можем... Обратите внимание, справа внизу там есть ГПУ, ЦПУ. Что это такое, кто-нибудь знает, нет? Ребята, кто-нибудь из вас знает? Вот у вас ребята, Сергей? Сергей, ты знаешь, что такое ГПУ, ЦПУ?
*Графический процессор, центральный процессор.*
Да. Правильно. Значит, ЦПУ — это центральный процессор, а ГПУ — это графический процессор, процессор графической карты. Значит, я могу вам сказать, ребята, что вот на центральном процессоре, на самых-самых современных компьютерах, там, допустим, 32 вычислительных ядра. Это на самых-самых крутых таких. Вот. И из них там, скажем, 16 физических и 16 виртуальных. А вот у меня компьютер, где i7 процессор, на нём всего лишь восемь ядер от нуля до семи нумерация, с нуля идёт. И четыре физических вычислительных ядра, а остальные виртуальные четыре. А на графической видеокарте, даже самой-самой плохонькой, вот, допустим, GeForce GT 240 Nvidia, на ней 96 шейдерских процессоров стоит. Каждый шейдерский процессор примерно эквивалентен ядру вычислительному на центральном процессоре, но он на порядок мощнее, чем ядро на центральном процессоре. Я не знаю, почему так получается. Но, в общем, графическая карта — это суперкомпьютер, который там стоит на плате материнской, понимаете? Или даже в виде чипсета просто, просто чип на материнке стоит Nvidia — это суперкомпьютер, понимаете? Вот. То есть там десятки, там, сотни вот этих вот ядер вычислительных, которые по своим возможностям на порядок превосходят центральный процессор. Вот у меня на компьютере слабенькая видеокарта GT 240 GeForce. Вот, на которой 96 шейдерских процессоров. Я на ней производил расчёты, получается от 200 раз быстрее до 4.000 раз быстрее, чем на центральном процессоре. Ребята, 4.000 раз быстрее! Что это такое? Ну это в зависимости от того, сколько там объектов выборки, там от размерности модели, там от разных параметров. Что это значит? Это значит, что вы считаете либо сутки, либо 12 лет. Есть разница, нет? Как вы скажете? Да принципиальная разница! То есть есть вещи, которые вообще невозможно посчитать на центральном процессоре. На графическом, ну, легко их можно посчитать. Ну, не легко, но можно.

Вы, может быть, слышали про майнинг, про биткоины, про такие криптовалюты? Слышали, нет, ребята? Вы же экономисты, должны были слышать. Все, все про это слышали. Так вот, для майнинга, для того, чтобы огромные объёмы вычислений проводить, используются графические процессоры. Сейчас фирма Nvidia осознала, уже несколько лет, как это произошло, наверное, так лет пять, наверное, так, что вот это применение графических процессоров для неграфических вычислений является чрезвычайно перспективными. Они стали выпускать не графические карты с этими процессорами, а просто вот процессоры одни. То есть просто мультипроцессорные системы стали выпускать. Втыкаешь её в материнку, у тебя получается суперкомпьютер, грубо говоря, понимаете? То есть там, допустим, 4.000 шейдерских процессоров. Каждый из которых, как 10 вот этих центральных. И всё. И начинается вот это вот... Ну их обычно для майнинга используют. Но можно также использовать для научных расчётов, как сделано в системе Aidos. Но у неё только два модуля на графическом процессоре реализовано, потому что это было сделано на языке OpenCL, который низкоуровневый и похож на ассемблер. И он, там требуется явное указание распараллеливаний там, сборки данных по разным этим линиям обработки и так далее. В общем, короче, получается, что это довольно трудоёмко, и сделано было только для модуля расчёта моделей и модуля распознавания. Это было сделано Димой Бандыки из Беларуси.

Вот. Значит, дальше мы что можем сказать? Значит, что... Вот вы посмотрите внизу строчечка там. Там написано, что данная модель, которая сейчас у нас будет обрабатываться, составляет 60% от того, что может обрабатывать система Aidos. Что это значит? Это значит, что она может обрабатывать значительно большие модели по размерности, несопоставимо большие. Ну, десятки тысяч раз большие, понимаете, там, так примерно. Поняли, да? Вот. А если, значит, она не может обработать. Вот если мы там задали 4 млн там и 500, 1.500 классов и 30.000 признаков, то тогда что? Тогда она напишет, что эта вот модель там у вас составляет 150% от того, что может делать система Aidos. И тогда, пожалуйста, там вот задайте, там где вот 100 написано, задайте, сколько процентов наиболее значимых результатов оставить, а остальные выбросить. Тогда мы сможем это обработать. То есть она вам прямо рекомендует это сделать, иначе она отказывается дальше работать.

#### 4.5. Верификация Моделей и Bootstrap

Вот здесь вы можете указать пять, например, или 10 там. И у вас будет оставлено только 10 наиболее достоверных результатов идентификации. Что это значит? Система сама не только принимает решение о том, что объект относится к какому-то классу или не относится, но она и знает достоверность этого решения и способна наименее достоверные решения выбросить вообще, не показывать их нигде. Тогда размерность модели сохраняется, сокращается существенно выходных форм, и она тогда может обрабатывать и большего объёма модели. Но сейчас не надо там этого делать. Сейчас вы напишите там 100. Ну, или 50 хотя бы, не знаю. Ну 100 пусть будет там, как было.

Значит, и теперь ещё такой момент. Значит, если модель содержит очень большое число объектов обучающей выборки, ну, скажем, 4 млн, то синтез модели может осуществляться, скажем, час, например, на центральном процессоре, а проверка достоверности путём решения задачи идентификации может производиться, скажем, полгода будет происходить на 4 млн, допустим, объектов и 500 классах там и, допустим, 30.000 признаков. Это чудовищный объём вычислений, понимаете? Каждый объект обучающей выборки копируется в распознаваемую. Ну не вся целиком копируется. Потом организуется цикл по объектам распознаваемым, потом цикл для каждого объекта по всем классам. Потом для того, чтобы сравнить каждый объект со всеми классами, а для того, чтобы сравнить его хотя бы с одним классом текущим, для этого нужно произвести 30.000 сложений и произведений, то есть оно вычисляет суммарное количество информации в признаках объекта о принадлежности к этому классу. Если там 30.000 признаков. И вот представьте себе, что там этих циклов, сколько она будет там коркаться, пока это всё посчитает. Это очень трудоёмко. Вот это реализовано на графическом процессоре, происходит в очень короткое время.

Значит, мы можем осуществить только синтез модели, смотрите. Только верификацию или синтез и верификацию. Если модель небольшая, то можно спокойно запускать одновременно синтез и верификацию модели, как сейчас вот. И при гораздо больших размерах тоже так же точно можно запускать первый пункт. Но если модель, допустим, считается там, скажем, 15 минут на графическом процессоре, а потом, допустим, полгода происходит верификация, тогда это уже безумие это запускать. Тогда делается следующим образом: осуществляется только синтез модели сравнительно быстро, а потом осуществляется верификация модели на каком-то подмножестве обучающей выборки. Вот, допустим, выбираем пункт такой: "Копировать n случайных объектов". Вот поставьте там курсор. На птичку там эту, отметьте этот пункт. Ниже. Вот. И у вас вот здесь вот вы можете задать 100, например, или 1000. Вот. Тогда будет случайным образом выбрано 1000 объектов. Ну курсор надо поставить там полевее, наверное, как-то. Что-то как-то у вас получилось странно. Вот. Поняли вы, да, короче? Можно в самом начале его поставить, у вас как-то неудачно немножко получилось. Вот. И, значит, представьте себе, вы верификацию модели проводите на 1000 примеров, на 1000 объектов из 4 млн. И оцениваете приблизительно достоверность модели. И это система измеряет в сотых долях секунды время исполнения и пишет вам это время на экране после окончания расчёта. Вы берёте и пишете пропорцию такую, что 8 минут {тире} идентифицировалось 1000 объектов. Да? Вот. Сбоку пишете: 480 минут {тире} идентифицировалось X объектов. Или можно написать там 1000 {тире} 8 минут, X {тире} 480 минут. 480 минут — это сколько там? 8 часов умножить на 60 минут, да? 480, 6\*8=48, 480 минут. То есть вы можете посчитать, сколько объектов вы можете вот здесь вот задать, если у вас есть 8 часов на решение задачи верификации модели, проверки их достоверности. У вас получится там, допустим, 35.000 там или 48.000. Вы задаёте эти 48.000, идёте спать, утром приходите, у вас модели проверены на достоверность на 48.000 случайных объектах обучающей выборки. Это уже, ребята, вполне хорошая будет оценка достоверности. Вот всё это вместе, то, что я сейчас вам рассказал, называется Bootstrap подход, Bootstrap. Bootstrap. Это слово означает складной нож на скандинавском каком-то там языке. Вот, который этот метод так назван, почему? Потому что часть выборки используется для одной цели, скажем, для синтеза модели, а другая часть — для другой цели, для проверки, например, модели. Но мы можем использовать одни и те же объекты и для выборки для и для формирования модели, для проверки их на достоверность. Это будет зависеть от того, поставите птичку его ниже здесь вот: "Не удалять", "Удалять". Можно удалять, можно не удалять, как хотите.

#### 4.6. Запуск Расчета

Значит, сейчас давайте все параметры по умолчанию. То есть ставим птичку вверх, вверху, копировать всю обучающую выборку на графическом процессоре. И нажимаем О'кей.
Вот, сначала происходит синтез модели, потом их проверка на достоверность.
Вот это плохо. Вот это означает, ребята, что на этих компьютерах видеокарта, скорее всего, там чипсет несовместим с чипсетом Nvidia. Поэтому не получится у нас использовать графический процессор. О'кей нажимаете. Вот. И придётся на выход нажать из системы. Вверху там вот. На выход, либо вообще придётся диспетчер задач, можно запустить здесь, не знаете? В общем, короче, застряли мы. Прерывайте главную программу. Да, хорошо, что можно задать его. Снять задачу, потому что просто не выполняется OpenCL на графическом процессоре.

Так, и заново запускаем систему Aidos. И, как вы догадываетесь, задаём этот же самый режим 3-5, но уже используем для этого... Ну да, это тоже модель, конечно, тоже это... Вот это как раз и есть этот модуль, который синтез модели на графическом процессоре. Запускаем систему Aidos. Start Aidos, да. Сейчас, поскольку у вас обновление уже проведено, второй раз оно не будет происходить. Этот модуль проверяет, насколько ваш исполнимый модуль устарел. Если его дата не отличается от даты обновления, тогда, в общем-то, то есть он не новее, чем... то есть он не старее, чем обновление, то тогда обновление не скачивается. Всё, заходим сюда, сюда. И режим 3-5 сразу быстренько заходим. Это она сейчас отмечается на сайте, на FTP-сервере системы.

Это можно закрыть. Да, и заходите в режим 3-5, когда он откроется. И задаёте центральный процессор внизу. И О'кей.

Вот, созданы модели и проверяются на достоверность. И прогноз времени исполнения.

Вот, поскольку там разные операции происходят, то происходит корректировка постоянная этого времени исполнения, прогнозируемого. Вот, но оно потом так реалистично, в общем, показывает.

### Заключение и Итоги

Теперь смотрите, что произошло? Были созданы все модели и все проверены на достоверность путём решения задачи идентификации. Закрываем это. О'кей, нажимаем.

Вот. И переходим в режим 5-5 и смотрим на эти модели глазами. Что ж там за модели? Модели: три модели статистических (ABS, PRC1, PRC2) и семь моделей системно-когнитивных. Вот здесь мы можем нажать на кнопочку "Помощь" и посмотреть, как эти модели рассчитываются. Я вам этого не объяснял, ребят, нет ещё? Скажите, объяснял, не объяснял вот это? Нет. Закрываете эту форму. Сейчас мы подойдём к её объяснению. Смотрим сначала модель ABS.

Вот. Можно чуть-чуть это вверх поднять его, потому что мне там заслоняет ваша эта надпись. Что насчёт демонстрации экрана. Вот. Значит, смотрите, что мы здесь видим? Как устроены эти матрицы моделей? Вертикальная, значит, колоночки... горизонтальная шапка, да, содержит наименование колоночек. Там у нас названия классов. Это градации классификационных шкал. Они пронумерованы все. Вот эти номера полностью соответствуют тем, которые в справочнике классов 2-1. А слева мы видим шапку вертикальную, и там названия строчек. Строчки соответствуют значениям свойств или значениям факторов. То есть вот, допустим, шкала "Цвет" — описательная, а градации там: белый, чёрный, жёлтый, там, видите, да? Прокрутите вниз, пожалуйста, до конца. Что вы здесь видите? Значит, вы видите в самом конце итоговые строки, в которых есть количество объектов обучающей выборки, которые были представлены как относящиеся к определённому классу. Здесь это не будет двигаться, потому что там есть зафиксированная колоночка. Вот. То есть там форматирование таблицы уже осуществлено.

И вот вы можете подвинуться вправо пигуночком, что у вас разное число объектов по разным классам. Вот класс "Элемент компьютера" — семь объектов, семь примеров предъявлено. "Средства связи" — три предъявлено. Теперь давайте вверх прокручиваем чуть-чуть незначительно. И там у нас есть строчка, по-моему, двадцать третья строчка. Чуть-чуть выше прокручиваем. Вот. Это "Размер под руку". И вот смотрите, значит, вы видите справа, что признак "Размер под руку" три раза встретился у средств связи и три раза встретился у элементов компьютера. А при этом мы помним, что элементов компьютера семь, а средств связи три. Поэтому мы не можем считать, что если он там и там этот признак встретился три раза, то он одинаково является характерным для объектов этих категорий. Ясно, что для средств связи он гораздо более характерен, чем для элементов компьютера. Почему? Потому что 100% этих объектов, которые являются средствами связи, этим признаком обладают, и только 42% объектов, относящихся к элементам компьютера им обладают. А в среднем обладают 30% им. Сейчас мы это всё увидим.

Значит, давайте всё это закрываем эту форму и переходим в форму ПРЦ-2. Здесь у нас показано, какое число объектов, относящихся к каждому из классов, обладают тем или иным признаком. Давайте вправо прокручиваем и до двадцать третьей строки. Или сначала до двадцать третьей строки, а потом вправо до конца. Как хотите. Вот смотрите, у нас, видите, 100% средств связи с этим признаком "Размер под руку", и 42% и 8 там с копейками, тысяч элементов компьютеров, да, с этим признаком. А всего по всей выборке 30% объектов с этим признаком. Это что значит? Что для средств связи он более характерен, чем для элементов компьютера. А для тех, и для тех, и для тех он более характерен, чем для других категорий объектов.

То есть мы можем вот так вот сравнивать эти вот условные, безусловные процентные распределения, делать выводы. Можем их друг с другом сравнивать, можем сравнивать их с безусловными относительными частотами. То есть есть условные и безусловные. Условное — это частотное распределение у объектов определённой категории, а безусловное — по всей выборке. Понятно, да? Придите, отец, слушайте внимательно. Это похоже на вероятности, но это не вероятности, это относительные частоты, условные и безусловные. Они стремятся к вероятностям, эти значения, при неограниченном увеличении объёма выборки. А сейчас у нас малая выборка, то есть они там довольно далеки от вероятности. Тем более, что у нас разное число объектов наблюдалось по разным категориям. То есть по той категории, где объектов больше, у нас эта погрешность меньше этого числа, относительная частота, она ближе к вероятности. А там, где объектов меньше представлено, там она может больше отличаться от вероятности. Погрешность больше будет этого значения. То есть различие между относительной частотой и пределом, к которому она стремится при неограниченном увеличении объёма выборки, то есть от вероятности.

Ну, значит, я вам могу сказать, что когда вот матрица небольшая сравнительно, ну здесь вот 14 на 50, это ещё такое малюсенькая моделька, 60% от того, что может система Aidos формировать. Вы представьте себе, что там у нас, ну хотя бы 500, например, колонок, а может быть до 1.500-500 или до 2.000 колонок. Но я ограничил 1.500 искусственно, но там сам язык программирования позволяет где-то 2.200 там колонок в базе данных. А строк вообще нет ограничения жёсткого, там может быть и миллионы строк, понимаете? И вот представьте себе, мы имеем матрицу, в которой миллионы строк и 1.500 колонок. И вот так вот глазами будем всё это сравнивать, что ли? Значит, мы эту таблицу, если напечатаем, она весь главный корпус университета заклеит, понимаете? Страничками А4, десятым шрифтом. То есть это полный бред, понимаете, это глазами анализировать. А тем более решать на основе этого какие-то задачи, на основе этого анализа, вообще исключено, если мы это будем делать вручную. Поэтому, значит, в системе Aidos — это очень важное достоинство системы Aidos, что она сравнивает вот эти вот условные, безусловные относительные частоты и по разным алгоритмам происходит сравнение, но это происходит автоматически, ребята.

Закрываем это эту форму. И смотрим помощь. Вот. Ну пошире давайте. Значит, у нас И — индекс И — это у нас индекс строки, Ж — это индекс колоночки. Значит, тут я вам сразу хочу вот что сказать, что если мы хотим сравнить два числа, вот в предыдущей группе я спрашивал студентов, как можно сравнить два числа? Какое из них больше, какое меньше? Мне ответили. Там была одна из студенток, которая сказала, как ответила. То есть как ответ можно сделать? Интересно, у вас вы можете предположить, как это можно сделать? Сравнить два числа? Какие нужно произвести операции и какие сравнения, чтобы сказать, какое число больше, какое меньше? Ну? Давайте, давайте. Вычесть одно из другого и посмотреть, больше нуля или меньше нуля разность. Если больше нуля, первое число больше. Если меньше нуля, первое число меньше. Если равно нулю, значит они одинаковые. Это путём вычитания. А деление — это многократное вычитание одного и того же значения. То есть это тоже, в общем-то, очень похоже на вычитание. Но деление тоже можно использовать для сравнения двух чисел. Если разделить первое на второе и получится у нас результат частное больше единицы, значит, значит, числитель больше, чем знаменатель. Первое число — числитель, второе — знаменатель. А если равно единице, тогда значит они одинаковые. А если меньше единицы, тогда значит первое число было меньше, чем второе. Вот.

Теперь представьте себе, что мы хотим сравнить объекты с обобщёнными категориями, определить, какие признаки наиболее характерны или менее характерны, или вообще не характерны для тех или иных категорий. И хотим сравнить объекты с этими категориями. Вот у нас, допустим, есть объект за дверью, и мы хотим узнать, парень это или девушка. Вот. И мы спрашиваем: "Волосы длинные?" Он говорит: "Да, длинные". О чём это говорит? Это несёт информацию о том, что это девушка, наверное. Хотя бывают и ребята, да, с длинными волосами. Потом мы берём и спрашиваем: "А в брюках?" Нам говорят: "Да, в брюках". Мы тут задумались. Ну вообще-то ребят больше в брюках, чем девушек. Ребята все, а девушки почти все. Но вот всё-таки как-то по тому, насколько вероятно встретить брюки, они мало отличаются парни и девушки. То есть очень высокая вероятность встретить девушку в брюках тоже. Хотя вот у вас тут на фотографиях тут девушки у вас не в брюках. Вот. То есть признак "брюки" он носит некоторое количество информации, содержит о том, что это всё-таки парень, но количество информации небольшое. А вот длинные волосы довольно большое количество информации содержат о том, что это девушка. Почему? Потому что вероятность встретить этот признак "длинные волосы" в группе девушки намного больше, чем в среднем. А признак "брюки", вероятность встретить в группе мальчиков больше, чем в среднем, но совсем чуть-чуть больше, чем в среднем. А у девушек совсем чуть-чуть меньше, чем в среднем. А длинные волосы у ребят намного реже, чем в среднем встречаются. Ну вот примерно такая вот логика. То есть мы можем посчитать, какое количество информации содержится в признаке о принадлежности объекта с этим признаком к тому или иному классу, сравнивая относительные частоты наблюдения этого признака в этой группе и в среднем по всей выборке.

А, значит, есть возможность эти частоты относительные посчитать через абсолютные частоты. И вот сейчас вот мы посмотрим. Вот смотрите, модель INF3 — это слагаемое хи-квадрат. N итое житое — это число встреч итого признака в житой группе. Вот то, что мы сейчас в матрице ABS смотрели. А вот это интересное такое выражение: N итое умножить на N житое разделить на N — это теоретическая частота наблюдения итого признака в житой группе, в объектах житой группы. Это формулу хи-квадрат Пирсона. Он предложил фактическую, то есть фактическое — это понятно, что такое, просто сколько раз встретилось, и теоретическую частоту сравнивать. Теоретическую он предложил выражение для них, для неё. Вот. И он предложил это сравнение делать путём вычитания их друг из друга. А можно это делать путём деления. Тогда у нас получится что? N итое житое — фактическая частота, делённая на вот эту теоретическую частоту. То есть это значит умножить на N и разделить на N итое N житое. N итое — это сумма по строке частот, N житое — сумма по колонкам. Если мы посмотрим, как рассчитать относительную частоту P итое житое, то это, конечно, понятно, что это фактическая частота встреч итого признака в житой группе, разделённая на суммарное количество наблюдений в житой колонке, по житому классу. Либо, значит, признаков, либо объектов. То есть P итое житое — это N итое житое, делённое на N житое. А P итое — это вероятность наблюдения итого признака по всей выборке. Это сумма по строке N итое, делённое на число всех признаков N. Если мы возьмём, подставим сюда вместо абсолютных частот относительные частоты, у нас получится условная частота встречи итого признака в житой группе, делённая на безусловную частоту. Это вариант один из вариантов формулы Харкевича Александра Харкевича. Таким образом, я вам показал, ребята, что частный критерий хи-квадрат очень похож математически на количество информации по Харкевичу, но отличается чем, собственно говоря? Тем, что хи-квадрат сравниваются числа путём вычитания, а количество информации по Харкевичу, они сравниваются путём деления. Но если мы имеем информацию о том, что у нас несколько признаков у объекта, то хотелось бы сравнить суммарное количество информации, которое содержится в признаках объекта о принадлежности к тому или иному классу. Это уже интегральные критерии, про них я попозже буду рассказывать на следующем занятии. А для того, чтобы это можно было делать, у нас хи-квадрат уже для этого подходит, когда нет взаимосвязи, это ноль получается слагаемое. А количество информации не подходит, нам нужно взять логарифм от этого отношения, чтобы единица, когда одинаковое число вероятности условная и безусловная, чтобы получалось у нас ноль. И можно другим способом тоже нормировать к нулю. Если просто взять отношение, не логарифм брать, а просто взять вычесть единицу, то получится выражение для так называемого коэффициента возврата инвестиций (ROI), который широко применяется в экономике, в теории управления инвестициями, портфелем инвестиций.

Здесь я хотел бы вам ещё сказать, что такое факт. Факт — это наблюдение определённого значения свойства у объекта определённой категории. И вот эта таблица ABS, которую мы называем и матрица сопряжённости, и корреляционная матрица, она как раз и содержит суммарную информацию по тем фактам, которые есть в обучающей выборке. В такой форме, на основе которой потом можно рассчитывать модели.

Всё, я вам примерно так рассказал эти вещи. Но каждая, вы видите, что каждая формула два раза написана. Немножко это некорректно с точки зрения математики. Но дело в том, что они совершенно одинаково рассчитываются, но при этом разный смысл переменных. Вот в модели INF2 сумма N житое — это не количество признаков суммарное по житому классу, а количество объектов обучающей выборки по житому классу. А INF1 — это количество признаков. Ну ясно, что чем больше объектов было предъявлено, тем больше будет и признаков. То есть они взаимосвязаны, положительно коррелируют. Но дело в том, что, вообще говоря, это разные вещи. И я вам скажу так, INF1 как бы лучше обосновано теоретически. То есть это прямо совпадает с теорией информации по семантической теорией информации по Харкевичу. А вот INF2 более понятно людям. Поэтому я и тот, и тот вариант использую. И они хотя и похожи, но бывают иногда то один лучше работает, то другой лучше работает. В смысле, достоверность лучше даёт.

Закрываем эту форму. И смотрим дальше. Значит, смотрим модель INF1. Здесь у нас... и смотрим опять же двадцать третью строку и вправо двигаемся. Вот. И мы видим там, что этот признак "Под руку" он содержит 0,779 бита о том, что это средство связи и 0,1031 бит о том, что это элемент компьютера. Потому что вот эта разница между ними как раз и показывает, что отражает, что для средств связи более характерен этот признак. Мы видим здесь некоторые числа с минусами. Это означает, что признак наблюдается у объектов этой категории, но реже, чем в среднем. И его обнаружение несёт информацию о том, что это не относится к данной категории. Связь отрицательная. Хи-квадрат будет меньше нуля. Ну, могу вам привести пример такой ситуации. Допустим, девушка, в классе девушки признак "брюки", он будет содержать количество информации о том, что это не девушка, понимаете? Потому что он наблюдается в этой группе реже, чем в среднем. То есть это будет информация с минусом, но очень незначительная, то есть очень мало отличающаяся от нуля. А признак "длинные волосы" в группе ребята, мальчики, он будет большим числом отрицательным, значительным. Потому что это будет означать, что он встречается в этой группе, но намного реже, чем в среднем. И гораздо больше не будет нести информации об этом.

Закрываем эту форму. И у нас остаётся буквально там 2 минуты до конца занятия. Насколько я понимаю. Пожалуйста, ваши вопросы, ребята.

То есть я тогда пишу, что мы рассмотрели вопрос о моделях, синтез моделей и начинаем на следующем занятии рассматривать вопрос о верификации моделей. Так вот мы будем двигаться, пройдём полностью, а потом ещё пройдём работы реальные, разработка методики риэлторской оценки, обработку текстов посмотрим интеллектуальную и посмотрим обработку графической информации. После этого вы будете готовы создавать свои приложения в любой предметной области практически. Так, на следующем занятии ЛР 303 достоверность моделей. Всё, ребята. На этом у нас сейчас конец занятия будет. Ребят. До свидания.