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

***Российская Федерация***

**167 Теория информации, данные и знания. Практическая 4. Инструкция по разработке собственного приложения 2020-10-27**

## Обсуждение тем студенческих проектов по анализу данных и требования к регистрации

**Резюме**

Видеозапись представляет собой фрагмент онлайн-занятия (практического занятия №4 по дисциплине "Теория информации, данные, знания" для группы Т-2003, проводимого 27 октября 2020 года профессором Луценко Е.В.). Основная цель занятия – обсуждение и утверждение тем студенческих проектов (работ), связанных с анализом данных.

**1. Вводная часть и организационные моменты:**  
Профессор приветствует студентов, уточняет детали занятия (дата, время, номер пары, дисциплина, группа). Происходит настройка связи и отметка присутствующих, подчеркивается важность этой процедуры.

**2. Цель занятия и требования к студентам:**  
Основная задача – помочь студентам выбрать, разобраться, утвердить и описать темы для их работ, включая реализацию моделей анализа данных. Профессор напоминает о необходимости регистрации студентов в системах ResearchGate и РИНЦ (Science Index) и заключения договора на размещение непериодических изданий для последующей публикации работ.

**3. Обсуждение тем студенческих работ:**  
Профессор призывает студентов предлагать свои темы.

* **Пример 1 (Студент): Анализ игры в дартс.** Студент предлагает проанализировать дартс, но сталкивается с проблемой поиска систематизированных данных. Профессор допускает возможность самостоятельного сбора и структурирования данных студентом.
* **Пример 2 (Профессор): Анализ астрономических данных.** Профессор приводит в пример успешную работу студента первого курса по анализу связи физических свойств звезд с результатами их наблюдений с Земли. Студент для этой работы самостоятельно собрал и структурировал данные. Профессор делится ссылкой на эту статью.
* **Пример 3 (Профессор): Спектральный анализ изображений.** Профессор напоминает о ранее показанном методе спектрального анализа изображений (на примере картин художников) и предлагает студентам применить его, например, к картинам русских художников определенного периода (времена года), скачав изображения из сети. Указывает на потенциальную сложность скачивания большого количества изображений по одному и предлагает поискать архивы.
* **Пример 4 (Профессор): Анализ данных индийской кухни.** Профессор находит набор данных по индийской кухне и предлагает его для анализа. Демонстрирует процесс первичной обработки данных (преобразование в формат DBS, работа с текстовыми и числовыми шкалами, обработка пропусков и специфических форматов записи) и постановку задачи: определить характерные для разных регионов Индии ингредиенты.
* **Пример 5 (Профессор): Анализ данных по недвижимости.** Профессор находит набор данных по оценке недвижимости и предлагает его в качестве темы. Отмечает, что данные выглядят подходящими для анализа зависимости стоимости от различных характеристик.

**4. Призыв к активности и заключение:**  
Профессор выражает неудовлетворенность пассивностью студентов и настоятельно призывает их активнее предлагать темы и использовать время занятия для консультаций, пока есть такая возможность. Демонстрирует готовность помочь с выбором и формулировкой тем. Занятие подходит к концу, обсуждается одна из предложенных тем (индийская кухня) и передается студенту Шамилю. Другая тема (недвижимость) остается доступной для выбора.

**Детальная расшифровка текста по словам**

**1. Вводная часть и организационные моменты**

Здравствуйте, ребята.

Микрофоны выключайте.  
Так.  
Ребята, ну я тоже очень рад встретиться с вами, хоть уже и вечер. 27 октября 2020 года. Шестая пара 17:20-18:50.  
Это занятие проходит.  
Вот. По дисциплине. Лабораторная работа номер семь по дисциплине. Или нет, это практическое занятие, наверное, уже. Номер четыре будет.  
Сейчас я напишу. Практическое занятие.  
Сейчас мы определимся.

По дисциплине теория информации, данные, знания.

И у нас, в общем-то, сегодня была пара с этой группой Т-2003. Сейчас, наверное, будет продолжение занятия, которое было у нас на третьей паре.

Сейчас я, ребята, отмечусь везде, и вас отмечу, сам отмечусь. Потому что это самое главное в нашем деле - отметиться. А потом можно уже вести занятия.  
Так, у нас сейчас 2003 группа.  
Практическое занятие, шестая пара.  
У вас аж 28 человек вас.  
Занятие ведёт профессор Луценко Евгений Вениаминович.

Так.  
И ориентируемся. Я вам доступ дал. Теперь, ну, ёлки-палки, я не не обратил внимание.  
Так, 4. Практическое занятие номер четыре. То есть даже правильно отметил.

**2. Цель занятия и требования к студентам**

Так, ребят, ну, на этих занятиях я вам уже говорил. У нас задача смотреть, какие вы выбрали темы. И я вам помогаю с этими темами разобраться, утверждаю их и помогаю вам их описать, реализовать модель и описать.  
Вот такая у нас тематика этого занятия. Поэтому давайте этим и займёмся.  
И вам надо регистрироваться, я вам уже каждый раз повторяю это, в ResearchGate и РИНЦе. И не просто в РИНЦе, а в системе Science Index и потом заключать договор на размещение непериодических изданий.

**3. Обсуждение тем студенческих работ**

Ну а сейчас, ребята, давайте говорите, какие у вас темы, какие вы темы выбрали. Или вы ещё не успели с прошлого занятия очухаться? С третьей пары?

* **Пример 1 (Студент): Анализ игры в дартс.**
  + *(Студент, прерывая)* Ну, я уже...
  + *(Профессор)* Не, ну тот, кто выбрал, тот с тем вопросов нет. Я ж утвердил темы некоторым ребятам, да? Вот вы можете дальше уже всё описывать, там брать прямо статью за основу и описывать.
  + *(Студент)* А можно взять, допустим, анализ дартса?
  + *(Профессор)* Анализ чего?
  + *(Студент)* Дартс.
  + *(Профессор)* А-а, не понял немножко. Чего именно?
  + *(Студент)* Игра такая.
  + *(Профессор)* Вы знаете, игра? Нет, не слышал даже.
  + *(Студент)* Ну там, где дротики надо кидать в мишень.
  + *(Профессор)* Ну, давай, данные присылай, ссылочку, посмотрим.
  + *(Студент)* Да я пока ещё не могу найти такого сайта, чтобы там были систематизированные данные.
  + *(Профессор)* А нам и не надо... Даже если что, придётся самому это всё составлять. Ага, вот как. Ну, такое тоже может быть.
* **Пример 2 (Профессор): Анализ астрономических данных.**
  + *(Профессор)* Вы ж видели, ребята, там вот есть работы среди уже сделанных работ. Есть очень интересные работы. Мм, такие, которые ребята самостоятельно данные создали для них.
  + Сейчас я попробую вам найти даже, может быть.
  + Вот эта работа мне очень понравилась, ребята.
  + Тут парень сам собрал данные. Так их упорядочил, в таком виде представил к юта там обозначил. Вот.
  + Видите, как интересно? Там, знаете, в чём речь идёт?
  + Вот смотрите. У звёзд есть определённые физические свойства. Эти звёзд свойства известны сейчас науке. При этом они как-то наблюдаются определённым образом. Вот возникает вопрос: связаны ли вот результаты наблюдений этих звёзд с Земли с их физическими свойствами? То есть как влияют физические свойства звёзд на вот эти вот результаты их наблюдения? Ну, ясное дело, что они должны быть связаны, но вот фактически, как они связаны или нет? Вот этот анализ проведён, те красные. Интересный анализ, дающий интересные, в общем, результаты. Вот.
  + Я сейчас вам ссылочку на эту статью пошлю. Это тот самый случай, когда студент не зарегистрировался в ResearchGate и в РИНЦ. Поэтому я эти статьи его эту статью разместил. Такая, знаете, полновесная статья. Имейте в виду, что это студент первого курса написал эту статью. Максим.
* **Пример 3 (Профессор): Спектральный анализ изображений.**
  + *(Профессор)* Значит, ребят, значит, давайте тогда я буду вам высказывать мысли, а вы будете как бы буду вам подсказывать, какие темы можно взять. Попытаюсь навыдумывать разных тем побольше.
  + Значит, уже какой-то опыт есть. Значит, вот есть я вам показывал показывал спектральный анализ изображений в системе Aidos, правильно? На примерах картин с художниками.
  + Отвечайте: показывали, не показывал. Как вы помните, не помните это, что я показывал вам спектральный анализ? Ребят, отвечайте, давайте.
  + *(Студент)* Показывали.
  + *(Профессор)* Показывал. А теперь берите, значит, ищите. Вот.
  + Вот, картины времён года русских известных русских художников. Кто вам мешает вот это вот изображение скачать и обработать? М? Прекрасная работа получится.
  + Ну правда, их не хочется скачивать по одному.
  + Да, к сожалению, тут скачивать неохота по одному. Ну, наверное, можно найти. Где-нибудь там даже в торренте, в торренте, может быть.
  + Поищите, ребята, архив картин.
  + Так сразу я не найду что-то. Ну вот такая идея вам.
* **Пример 4 (Профессор): Анализ данных индийской кухни.**
  + *(Профессор)* Выбираем данные, которые не очень большого объёма. Здесь, допустим, 492 МБ, не надо их выбирать. Вот 397 КБ можно и выбрать.
  + Вот индийская кухня. Смотрим. Индийская кухня. Данные. Indian Foods.
  + Смотрите, что я делаю. Я нашёл тему, сейчас кому-нибудь её отдам эту тему. Кто первый скажет: мне эту тему, тому и отдам.
  + Что что мы делаем, ребята, с самого начала? Сначала берём и преобразуем просто в DBS файл это. Этот файл появляется в папочке, где был файл исходных данных. Вот.
  + И вот мы смотрим, ребята, что здесь у нас есть? Название 256 продуктов питания. Вот. Значит, здесь у нас есть что? Штат, регион. И какие ингредиенты, какое название.
  + И давайте вот решим так. Смотрите, ребята. Мы будем по ингредиентам определять, по ингредиентам будем определять, какой продукт. Вот. Вернее так, по ингредиентам будем определять, какой регион, откуда этот продукт.
  + Можно вообще-то продукт определять, но они, дело в том, что они их очень много. 256.
  + Почему-то здесь надо название, надо числа. Не очень понятно. Смотрим. Регион.
  + Да. Попробуем Excel. Данные из текста. С... *(Профессор работает с файлом, слышны звуки интерфейса и лай собаки)* Да. Пробуем Excel. Ну ладно. Странно, что он не захотел преобразовывать его.
  + Ну ладно, значит, давайте смотреть. Значит, вот у нас наименование продукта, ингредиенты, какая диета, время приготовления, наверное, там что-то там, для чего, штат, регион. Вот мы должны определить, допустим, для какого региона какие наиболее характерные ингредиенты, например, которые используются для приготовления пищи.
  + Значит, для этого мы что сделаем? Возьмём, значит, сейчас... если какие-то зависимости национальные, скажем так, связанные с регионом. Вот. И посмотрим.
  + Здесь минус один, и, наверное, это отсутствие данных, скорее всего. Минус один заменим на пробел. А здесь пробел заменим на ноль. В числовых шкалах пробел, пробел на ноль заменим. Не, давай вот так вернёмся. И в числовых шкалах заменим минус один на ноль. Потому что в числовых шкалах отсутствие данных - это ноль в системе. А в текстовых шкалах отсутствие данных - это пробел. Поэтому мы... ну, можно не пробел, а вообще отсутствие символа. Вот. Вот тут у нас получилось.
  + Что мы видим, ребята? Смотрите. Значит, у нас, э, вот первая колонка, вторая, третья, четвёртая, пятая, шестая, седьмая, восьмая - это у нас описательные шкалы. А девятая - это у нас регион. Это у нас классификационные шкалы. Вот.
  + И мы в системе теперь вводим эти данные. Указываем тип Excel. Классификационная шкала девятая, описательная со второй по восьмую.
  + И вот здесь у нас, ребята, нюанс. Вот смотрите, значит, этот файл у нас, э, видите, ингредиенты у него э выражены через запятую, словами через запятую. И эту запятую заменю на пробел. Вот. Почему? Потому что в системе Aidos, э, ну, можно и запятую было оставить, но компоненты вот эти вот, они через э запятую, через пробел могут быть компоненты. Вот.
  + И я указываю, что у нас, э, описательные шкалы, мы применяем специальную интерпретацию текстовых полей и признаков. И смотрим, что у нас, э, как признаки рассматриваются слова. Слова рассматриваем. Но длиной больше трёх символов. Вот такой вариант. Задаёшь параметров, и у тебя всё это вводится. Получается модель.
  + Смотрите, ребята, значит, у нас классы получились вот эти вот э регионы: центральный, там, северный, южный, западный, восточный. А как признаки у нас получились, смотрите, ингредиенты. Это все эти слова, которые там были, они у нас стали признаками. 913 ингредиентов. Какая диета? Почему-то кроме вегана ничего нету. Нон-веган. Ага, понятно.
  + Тогда, значит, мы должны э пробел вот здесь вот заменить на нижнюю чёрточку. Вот. Тогда у нас это будет одно слово, но не будет интерпретироваться вот эти отдельные слова как признаки. То же самое сделать и здесь. Ну, можно и так вот сделать. Вот так то же самое. То есть мы заменили пробел на нижнюю чёрточку.
  + И ещё раз создаём модель. Теперь, э, выходим отсюда. Удаляем все приложения и ещё раз создаём модель при тех же параметрах. Можно адаптивные интервалы взять. Будет более достоверно, наверное. Ну ладно, давай так пока что. Неадаптивные. Вот. И вперёд.
  + Вот, что у нас получилось? Что у нас получилось, что э м-м... Опять что-то не то получилось. Почему ингредиенты у нас не воспринялись? А ну, странно. Регионы. Ингредиенты, они стали подряд у нас. Это неправильно. В адаптивных интервалах не работают вот эти вот слова. Смотрим. Всё правильно. Ну сейчас всё правильно.
  + Смотрим модели. Есть какая-то специфика по ингредиентам у регионов?
  + Вот. Модель S5, семантический резонанс знаний, наиболее достоверная модель.
* **Пример 5 (Профессор): Анализ данных по недвижимости.**
  + *(Профессор)* Ну вот, ребята, неплохая задача. По оценке недвижимости. Вот, так нормально. Понравилось. Вот. *(Профессор продолжает работу с файлом)* И преобразуем сразу же Dbase... *(Ошибка, повтор действия)* Ну да. *(Попытка с Excel, неудачно)* Сидней годится для таких целей. *(Работа с файлом продолжается)* Странно, что он не захотел преобразовывать его.
  + Ну ладно, значит, смотрим, что у нас тут такое. Значит, у нас, э, ребята, здесь описаны... Отлично преобразовалось. Вот. И нам это больше пока не нужно. Значит, я по недвижимости. Значит, смотрите. Э, мин, средняя стоимость частных домов в 1.000 долларов. То есть у нас, э, получается, классификационная шкала вот эта, средняя стоимость. Вот. А это у нас будет, ставим мы сюда колоночку. И у нас будет просто номер наблюдения. Вот, и мы эти наблюдения потянем сейчас. Тут не так уж много. Вот.
  + Теперь смотрим, ребята, значит, у нас описательная шкала со второй по четырнадцатую. А пятнадцатая - это классификационная. Вот. Что мы здесь делаем? Удаляем все приложения и решаем задачу по недвижимости. Простой Excel. Э пятнадцатая, пятнадцатая, со второй по четырнадцатую. Ну, адаптивные интервалы и не будем применять специальную интерпретацию текстовых полей. И вводим данные. А где он там? Что его там нету, разве? Ну вообще. Я же его сделал вроде как. *(Повторяет действия)* Странно, странно. Ладно. Есть каким-то образом удалил его нечаянно.
  + Так, значит, туда мы опять повторяем. Значит, смотрите, здесь, ребята, есть расшифровка этих вот показателей, которые здесь есть. Эту расшифровку можно использовать. Это хорошая задача по недвижимости. Мне она нравится. *(Работа с файлом продолжается)*

**4. Призыв к активности и заключение**

Что-то я смотрю, как-то вы не собираетесь особо заниматься ничем. А надо бы.  
Уж не забывайте, что сейчас у вас возможность такая есть. Вот занятие, я сижу, помогаю вам. А потом этого не будет. Будете сами что-то выдумывать.

Ребят, ну так не годится. Давайте работайте как-то. Или надо чтобы давать задания вам? Должны сами искать тему, которая вам больше понравится. Я же вам объяснил, где их искать, как искать. Тут какие-то проблемы возникают с этим. Давайте, занимайтесь этим. Вот сейчас время для этого выделенное. Давайте, быстренько.

*(Возвращение к теме индийской кухни)* Значит, э, ребята, вот это приложение я вам сейчас посылаю в чат его. А, ребята, эти данные всё равно на каком языке, ребята. Ну, обычно на английском. Э хорошо, Шамиль, тебе это, тебе это данные. Вот. Лишь бы, лишь бы там э, если, скажем, иероглифы, там не будет видно, не поймёшь ничего. Если арабские, то тоже не очень-то. А если э другие языки, у которых вот э алфавит используется латинский, английский, немецкий, французский, испанский, итальянский, ну тогда проблем нет особых. Ну и другие там языки, которые вот используют обычный алфавит.  
*(Возвращение к теме индийской кухни и ее результатам)* Вот, индийская кухня. Смотрим теперь достоверность. Ну там, судя по этой форме, достоверность получилась очень хорошая. 0,901. И здесь мы видим тоже по графику тоже подтверждается это. Вот этот график он означает что? Что чётко разделяются положительные решения истины от ложных. Ложные решения при низких уровнях сходства, положительные истинные решения при более высоких уровнях сходства. А при решении отрицательном непринадлежности всегда истинных решений гораздо больше, чем ложных. То есть прекрасный результат, ребят, хорошая модель. Ну это не так чётко видно. Ну да, вот это вот получше. Всё прекрасно видно. Видно, что начиная с 10% уже значительно больше истинных решений, чем ложных. То есть это что значит? Что мы прекрасно можем идентифицировать регионы, ребята. 3221 по по ингредиентам пищи. Какие какие ингредиенты используются. И можем посмотреть. Вот у нас эти регионы. Они все друг на друга не похожи. Ну, такой результат нам не очень интересен. Нам интереснее тогда дендрограмму посмотреть. Какие регионы сходны друг с другом по ингредиентам пищи, особенностям пищи национальным разных регионов. Вот мы видим, смотрите, центральные и восточные, да, наверное? Западные и южные. Короче, вот они объединяются таким образом, дендрограмму. То есть всё можно изучать, всё это прекрасно получается.  
*(Передача темы Шамилю)* Значит, можем посмотреть и нейронную сеть. Как эти регионы выглядят? Наиболее чётко выражен центральный регион, остальные слабо выражены по своим признакам. А если взять N3, тогда что? Ну тогда более картина такая естественная. Короче, ребята, что у нас получилось? Что для одних регионов одно характерно, другое не характерно, других - другое. То есть у них есть ярко выраженная специфика у регионов. Вот, допустим, для западного региона очень характерно э что? Сахар там, да? Ну что там ещё? Ну я не знаю, что это значит, ну короче там. Ну, прекрасная получилась модель. Тогда, ребята, значит, э вот э Шамиль, ты ж э эту хочешь тему взять? Давай свою эту э почту в чат. Я сейчас тебе пришлю этот файл исходных данных. Просто введёшь его и будет... А, ты передумал, да? Ну тогда кто-нибудь другой, может быть, возьмёт.

*(Обращение к Ивану по поводу темы недвижимости)*  
*(Студент)* Я скинул.  
*(Профессор)* Что? ...не расслышал.  
*(Студент)* В чат, в чат я скинул.  
*(Профессор)* Мм. А, всё-таки ты согласился, да? Ну хорошо. Тогда сейчас тебе пошлю этот файлик исходных данных. *(Профессор ищет почту)* Как нет, не видно письма твоего? *(Находит)* Вот. Сейчас напишем тебе письмо. *(Готовит письмо)* Вот, во-первых, сама ссылочка на dataset. Вот, во-вторых, файл исходных данных. CSV и экселевский. И уже посчиталась достоверность. Сейчас глянем, интересно же, что там. Вот. 0947, да? Ну, что можно сказать, ребят, это не означает, что не получилось. Вот так где-то на три балла получилось. Ну получилось. Так, смотрим инсуммы. Ну, по-видимому, от наилучшим является инф1 критерий частный. Вот, значит, что мы видим? Ценовые категории: дешёвая, очень дешёвая и дешёвые - исходные. Вот. И 5 7 6 7 э 7 7 - это дорогие. Ну то есть всё вполне разумно. То есть у нас получилась недвижимость разумно классифицирована на дешёвую, дорогую по её признакам. То есть зависимости выявлены. Существует зависимость между вот этими показателями и э стоимостью недвижимости. Э очень дорогая и самая дорогая и очень дорогая - сходные, средние по цене сходные и дешёвые сходные, вторая и седьмая, и третья и седьмая. То есть всё очень логично, ребята, то есть это признак хорошей модели. Так что тебе э шлю вот эту экранную форму. Вот точно так же надо будет задавать параметры. Ну здесь понятно, что тип, диапазон этих вот шкал, а потом вот это вот самое главное. То есть как признаки рассматривают э слова. То есть это что означает? Что у нас вот эти ингредиенты будут рассматриваться как признаки. Послал. Так что, э Шамиль, ой, Шамиль, Иван, задача есть, Иван. Нормальная задача, не что ты как-то ты немножко это. Если будут вопросы, пишите вот на почту. Или на занятии спрашивайте на таком же. Я подскажу, помогу. Ну, идеальным является на занятии, конечно. Получил, да, уже это файлик? М-м. Хорошо.

Вот. Смотрим следующую задачу теперь. Кто-нибудь ещё, может, возьмётся.  
Всё, ёлки-палки. Так.

Ну ладно, ребята, неплохая задача. По оценке недвижимости... *(Повторяется)*

*(Заключение)* Так, ребят, сейчас у нас конец занятия. Но у меня вот почти закончена эта проверка файлика. Кому этот файлик нужен? Давайте почту на мою почту присылаете письмо. Вот, я вам пошлю этот файлик. Файлы исходных данных. Вот у нас получилось модель. Ну ещё я не знаю, насколько там достоверность высокая. Вот. И смотрим, что здесь у нас. У нас, значит, платформа, год. Год тоже надо было числовым сделать. Ну ладно. Ну, короче, если сейчас запустить на расчёт только синтез моделей и потом только верификация на небольшом количестве, там 100 случайных объектов. Ну сейчас она мгновенно найдёт этих случайных объектов эти. Сейчас, может быть, и успеем мы ещё быстро посчитать. Но это будет не точная достоверность модели на основе всех данных, но как бы оценка будет её. Это оценка, э то есть как бы, скажем так, причины, причины популярности игр. На каких платформах, в каких странах, где что, как что влияет на популярность. 19:52 планирует закончить. Ну, в смысле, э 18:52, сейчас закончится расчёт, мы увидим достоверность. Поэтому я советую вам э не уходить, а тем, кто заинтересовался, остаться и посмотреть. Вот, Иван, да? Иван, ну я ж не могу тебе отсюда почту взять, она у меня видишь, выскакивает блоком, улетает сразу. Поэтому ты мне напиши письмо на мою почту, которую я привёл. Я пришлю. А, написал, да? Хорошо. Да, я сейчас тебе пришлю. Что-то я не вижу. А, не туда, не тому пишешь. Да не, не понятно, честно сказать. Я скину. Что? В чат, в чат я скинул. Мм, а, всё-таки это согласился, да? Ну хорошо. Сейчас напишем тебе письмо. Во-первых, сама ссылочка на датасет. Во-вторых, файл исходных данных CSV и экселевский. И параметры интерфейса, при которых получился вот этот результат, который что можно ещё сделать? Можно попробовать не адаптивные интервалы, а одинаковые можно попробовать. А потом там появится калькулятор, можно разное число интервалов попробовать. Понимаете? То есть там можно пытаться получить модель более достоверную. Есть разные способы это сделать. Всё. Можете делать всё это, ребята. Теперь, э теперь конец занятия уже. Сейчас я посмотрю, что там у меня получил сообщение, секундочку. А, спасибо, да? Ну, послал я тебе, Иван. Всё, ребята, конец занятия на этом. На следующем занятии продолжим в том же духе, но я вам советую всё-таки самим попробовать поискать. Вы уже поняли, как это делается. Всё, самого хорошего вам, до свидания, ребята. До свидания. До свидания. До свидания. До свидания.